Speech Communication 96 (2018) 142–149

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Epoch detection from emotional speech signal using zero time windowing

Jainath Yadav*[,a], Md. Shah Fahad[b], K. Sreenivasa Rao[a]

[a] Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, India
[b] Department of Computer Science, Central University of South Bihar, Patna, India

ARTICLE INFO

Keywords:
Epoch extraction/detection
Emotional speech
Hilbert envelope of the numerator group delay (HNGD)
Zero frequency filter (ZFF)
Zero time windowing (ZTW)

ABSTRACT

The main objective of this work is to enhance the performance of epoch detection in the case of emotional speech. Existing epoch estimation methods require either modeling of the vocal-tract system or a priori information of the average pitch period. The performance of existing epoch estimation methods degrades significantly due to rapid variation of the pitch period in the emotional speech. In the present work, we have utilized the advantage of zero time windowing method, which provides instantaneous spectral information at each sample point due to the contribution of that sample point itself. The amplitudes of spectral peaks are higher at the instants of epochs compared to neighbouring sample points. The proposed method uses the sum of three prominent spectral peaks at each sampling instant of the Hilbert envelope of Numerator Group Delay (HNGD) spectrum, for accurate detection of epochs in the emotional speech. The experimental result shows that the accuracy of the proposed method is better than existing methods in the case of emotional speech. It is also observed that the proposed method works well even for the aperiodic nature of the speech signal and it is robust against emotional speech.

1. Introduction

Epoch is the instant at which excitation to the vocal-tract system is maximum during the production of speech. The epoch corresponds to the Glottal Closure Instant (GCI) in voiced speech, and to some random excitation in non-voiced speech. Extraction of epochs from speech has gained increasing attention in the recent years. Pitch-synchronous speech analysis exploits the knowledge of epoch locations in several speech applications such as prosody modification (Rao and Vuppala, 2013; Rao and Yegnanarayana, 2006), voice conversion (Rao, 2010), speech synthesis (Stylianou, 2001; Moulines and Charpentier, 1990), emotion recognition (Rao and Koolagudi, 2013), enhancement of reverberant speech (Yegnanarayana and Murthy, 2000), and glottal flow estimation (Wong et al., 1979). The time-varying characteristics of both excitation source signal and vocal tract resonances introduce difficulties for accurate detection of epochs (Yegnanarayana and Gangashetty, 2011). Further, epoch detection from the emotional speech is a challenging task due to large and rapid variations in pitch values. Due to these variations, existing methods detect many spurious epochs and miss some of the actual epochs in the case of emotional speech. The epoch parameters such as epoch interval, strength of the epoch and slope of the epoch strength contour vary from one emotion to other emotions (Koolagudi et al., 2010). The pattern of vibration of the vocal folds is also different for different emotions. As epoch and its associated parameters have a significant role in recognition and synthesis of emotions, a reliable and efficient method of epoch extraction is required in the case of emotional speech.

In literature, several methods have been proposed for the detection of epochs from the neutral speech signal. Some methods require electroglotto-graphy (EGG) signal rather than speech signal for detecting epochs (Thomas and Naylor, 2009; Childers and Larar, 1984). The large value within the pitch period of the Linear Prediction (LP) residual signal corresponds to epoch location (Atal and Hanauer, 1971). Therefore, LP residual signal has been used to detect epoch locations in many existing methods. The epoch extraction methods based on LP residual are Hilbert envelope based method (Ananthapadmanabha and Yegnanarayana, 1979), group delay based method (Smits and Yegnanarayana, 1995), Dynamic Programming Phase Slope Algorithm (DYPSA) (Naylor et al., 2007), Yet Another GCI Algorithm (YAGA) (Thomas et al., 2012), Integrated Linear Prediction Residual (ILPR) using plosion index method (Prathosh et al., 2013) and Speech Event Detection using the Residual Excitation And the Mean-Signal (SEDREAMS) based method (Drugman and Dutoit, 2009; Drugman et al., 2012). In Hilbert envelope-based method, epochs are detected from the peaks in the Hilbert envelope of the linear prediction residual signal. The group delay method is based on global phase characteristics of minimum phase signal. The average slope of the phase spectrum, also known as phase slope function, is computed as a function of time. The

* Corresponding author.
E-mail addresses: jainath@cub.ac.in (J. Yadav), shahfahad@cub.ac.in (Md. S. Fahad), ksrao@iitkgp.ac.in (K.S. Rao).

https://doi.org/10.1016/j.specom.2017.12.002
Received 11 October 2016; Received in revised form 2 November 2017; Accepted 2 December 2017
Available online 12 December 2017

positive zero-crossings in the phase-slope function are identified as epochs. The DYPSA method emphasizes the epoch locations using group-delay function of LP-residual and selects actual epoch candidates by using N-best dynamic programming. The YAGA algorithm is an extension of DYPSA method. The YAGA algorithm uses inverse filtering, multi-scale analysis, group delay function, and dynamic programming. In SEDREAMS algorithm, both glottal closure instants (GCIs) and glottal opening instants (GOIs) are obtained using two steps: (i) determination of expected GCI and GOI regions using a mean-based signal, and (ii) refinement of the expected regions where GCIs and GOIs are obtained using the LP residual. In ILPR plosion index method, the speech signal is preprocessed by half-wave rectified integrated LP-residual signal. The preprocessing step reduces ambiguity in finding the epoch locations. Further, epoch locations are detected by applying dynamic plosion index. Although LP-residual based methods work well in most cases, the parameters such as the order of LP analysis and the length of the window need appropriate values for deriving the LP-residual signal. These methods are dependent on the amplitude of LP-residual signal. Therefore, the performance of these methods reduces drastically due to the sensitivity of LP-analysis to noise in the signal. These methods work well when excitation impulses are periodic, but they fail when pitch period varies rapidly in view of emotional speech.

Zero Frequency Filtering (ZFF) (Murty and Yegnanarayana, 2008) is a non-LP-residual based method for epoch detection. In ZFF method, speech signal is passed through the cascade of two ideal zero frequency resonators. The output of the Zero Frequency Resonator (ZFR) is an exponentially increasing or decreasing function of time. The trend in the output of ZFR is removed by performing the mean-subtraction operation with the window length equal to the average pitch period of the speech utterance. The resulting signal is called zero frequency filtered signal. The positive zero crossings of the zero frequency filtered signal are known as epochs. ZFF method is not dependent on the amplitude of LP-residual signal. Therefore, ZFF method works well in signal degradation, and it is more effective than LP-residual based methods. However, ZFF method requires a priori information about the average pitch period. The pitch period of emotional speech varies rapidly in an utterance. Therefore, the performance of ZFF method degrades significantly in emotional speech due to rapid fluctuations of the pitch period. In Vijayan and Murty (2016), the phase spectrum of the speech signal has been exploited to detect epochs.

The characteristic of emotional speech is different from neutral speech due to rapid variation of the pitch period in an utterance (Gobl et al., 2003) (Govind and Prasanna, 2012). In addition, the segments with aperiodicity and sub-harmonics bear other problems as discussed in Mittal and Yegnanarayana (2014a). An effort has been made to enhance the performance of epoch estimation in the case of emotional speech (Govind and Prasanna, 2012). In this method, the average fundamental frequency (F_0) is calculated for every 20–30 ms segment of Zero Frequency Filtered Signal (ZFFS) by computing the highest magnitude frequency value from short-term Fourier transform. The window length for trend removal is computed as the inverse of average F_0 of the segment. In addition, each trend removed ZFF segment is passed through a low-pass filter with the cut-off frequency equal to $1.05 \times F_0$. Even though the accuracy of epoch detection is increased in terms of higher identification rate, lower false alarm and miss rates, but detected epochs deviate from genuine epoch locations (Deepak and Prasanna, 2014). Due to these reasons, this method is not suitable for the application such as epoch based prosody modification in case of emotional speech. The performance of epoch detection for emotions such as happy, angry, and fear is not comparable with neutral and boredom emotions in spite of the improvement in the epoch estimation (Deepak and Prasanna, 2014; Govind and Prasanna, 2012). In Kadiri and Yegnanarayana (2017), a single frequency filtering based approach for epoch extraction has been explored to provide high temporal resolution of excitation information and high resolution of spectral information. In brief, most of the existing methods for epoch extraction are based on the quasi-periodic assumption of the vibration of the vocal folds. Emotional speech contains rapid variations in pitch due to voluntarily controlled vibrations of vocal folds (Mittal and Yegnanarayana, 2014b). Therefore, existing methods are not suitable for detecting epochs from emotional speech.

In this work, we have proposed a new method for epoch estimation that utilizes Zero Time Windowing (ZTW) method (Yegnanarayana and Gowda, 2013) to extract the epoch locations from the emotional speech signal. Zero time window is a high decaying window function (impulse-like window function). Speech segment multiplied with this window preserves the amplitude of present sample and attenuates the remaining sample amplitudes significantly. Therefore, the derived spectral information using ZTW is due to the contribution of the present sample, where the window is superimposed. The window is shifted for every sample, and spectral characteristics of the speech signal are obtained using ZTW at each sampling instant with high resolution. ZTW operation is an approximation to integration operation in the frequency domain. The existing window functions smear the vocal tract information in the time domain and destroy useful information around epochs. The main advantage of ZTW is that it does not smear the discontinuity in the time domain due to epoch. The spectral peaks are stronger at glottal closure instants compared to neighbouring samples. Hence, the ZTW method provides more discriminative spectral characteristics between epochs and non-epochs with high time resolution. As ZTW provides discriminative spectral characteristics with high resolution in the temporal domain, it enables to carry out robust pitch estimation from noisy speech signal (Prasad and Yegnanarayana, 2015). For epoch extraction, epoch evidence signal is obtained by taking the sum of the three prominent spectral peaks from spectral energy profile. It uses Hilbert envelope and group delay function to enhance spectral peaks.

Rest of the paper is organized as follows. The motivation for this work is discussed in Section 2. Section 3 presents the proposed method for epoch extraction. In Section 4, the performance of the proposed method and its comparison with other existing methods are discussed. Section 5 discusses the conclusions derived from the present work.

2. Motivation

Spectral peaks are highly dominant and exhibit dynamic behaviour at epochs compared to neighbouring samples. Therefore, there is a need to compute spectral information accurately at each sample location. Traditional methods use a window size of 20–30 ms for computing the spectrum, and they provide the average spectral characteristics of the window segment. This is the main reason, which motivated us to use ZTW method for epoch detection in emotional speech. ZTW (Yegnanarayana and Gowda, 2013) method has been proposed to extract instantaneous spectral information at each sample. This method is robust against all types of sound units even for semi-vowels, trills, fricatives, voice bar, burst, and aspiration. ZTW method also works well for different types of signal degradation such as babble, vehicle and white noise. It uses impulse like window and numerator group delay function for providing the spectral information at each sampling instant. Fig. 1(a) shows the speech signal and reference epochs marked with circles. Fig. 1(b) and (c) show EGG signal $e(n)$ and one sample shifted EGG $e(n-1)$, respectively. Fig. 1(d) shows Differentiated EGG (DEGG) signal and negative peaks showing the locations of reference epochs. Fig. 1(e) shows the signal corresponds to the amplitude of the prominent spectral peak obtained from ZTW method at each sample point. Here, DEGG signal is computed from the EGG by taking successive sample difference (i.e., $e'(n) = e(n) - e(n-1)$). In digital domain differentiation operation is performed by computing the difference between the signal and its shifted (one sample shift) version (i.e., $x'(n) = x(n) - x(n-1)$). Where, $x(n)$ is the discrete time signal, $x(n-1)$ is one sample shifted version of $x(n)$ and $x'(n)$ is differentiated $x(n)$, which is computed by taking the difference between $x(n)$ and $x(n-1)$. Some existing epoch detection methods require average pitch