Kadiri, Sudarsana; Alku, Paavo; Yegnanarayana, Bayya

Analysis of Instantaneous Frequency Components of Speech Signals for Epoch Extraction

Published in:
Computer Speech and Language

DOI:
10.1016/j.csl.2022.101443

Published: 01/03/2023

Document Version
Publisher's PDF, also known as Version of record

Computer Speech and Language 78 (2023) 101443

Contents lists available at ScienceDirect

Computer Speech & Language

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

Analysis of Instantaneous Frequency Components of Speech Signals for Epoch Extraction

Sudarsana Reddy Kadiri [a,*], Paavo Alku [a], B. Yegnanarayana [b]

[a] Department of Signal Processing and Acoustics, Aalto University, Finland
[b] Speech Processing Laboratory, IIIT-Hyderabad, India

ARTICLE INFO

Dataset link: http://festvox.org/cmu_arctic/index.html

Keywords:
Speech analysis
Phase processing
Instantaneous frequency
Group delay
Excitation source
Glottal closure instants
Epochs

ABSTRACT

The major impulse-like excitation in the speech signal is due to abrupt closure of the vocal folds, which takes place at the glottal closure instant (GCI) or epoch in each cycle. GCIs are used in many areas of speech science and technology, such as in prosody modification, voice source analysis, formant extraction and speech synthesis. It is difficult to observe these discontinuities (corresponding to GCIs) in the speech signal because of the superimposed time-varying response of the vocal tract system. This paper examines the phase part of different frequency components of the speech signal to extract epochs. Three analysis methods to decompose the speech signal into different frequency components are considered. These methods are the short-time Fourier transform (STFT), narrow bandpass filtering (NBPF), and single frequency filtering (SFF). The locations of the discontinuities in the speech signal are obtained from the instantaneous frequency (IF) (i.e., the time derivative of the phase) of each of the frequency components. A method for automatic detection of epochs using the amplitude weighted IF is proposed. Performance of the proposed epoch detection method is compared with four state-of-the-art methods in clean and telephone quality speech. The performance of the proposed method is comparable with the performance of the existing epoch detection methods for clean speech but better for telephone quality speech.

1. Introduction

Analysis of speech signals using short-time Fourier transform (STFT) or bandpass filtering focus mostly on the magnitude part to represent speech information related to the excitation and the vocal tract system. This paper examines the phase part in these analysis methods, and shows that glottal closure instants (GCIs) or epochs in voiced speech can be extracted from the phase part as well. GCIs are used in many areas of speech science and technology: in prosody modification (Rao and Yegnanarayana, 2006), voice source analysis (Alku et al., 2009; D. Alessandro and Sturmel, 2011), glottal activity detection (Murty et al., 2009), estimation of fundamental frequency (Yegnanarayana and Murty, 2009; Kadiri and Yegnanarayana, 2018), estimation of formant frequencies (Joseph et al., 2006; Gowda et al., 2020), time delay estimation (Yegnanarayana et al., 2005; Murthy et al., 2020), estimation of the number of speakers from multi-speaker data (Swamy et al., 2007), speech enhancement (Deepak and Prasanna, 2016) and parametric speech synthesis (Airaksinen et al., 2018; Drugman and Dutoit, 2012). The importance of features derived around GCIs was studied in analysis and classification of voice qualities (Kadiri et al., 2020b; Kadiri and Alku, 2021), analysis, detection and classification of emotions (Gangamohan et al., 2013; Kadiri et al., 2015, 2020c; Kadiri and Alku, 2020), and pathological speech analysis and detection (Kadiri and Alku, 2020). More details on the GCI detection methods, and the GCI-based

* Corresponding author.
E-mail addresses: sudarsana.kadiri@aalto.fi (S.R. Kadiri), paavo.alku@aalto.fi (P. Alku), yegna@iiit.ac.in (B. Yegnanarayana).

https://doi.org/10.1016/j.csl.2022.101443
Received 9 December 2021; Received in revised form 24 July 2022; Accepted 18 August 2022
Available online 27 August 2022

analysis of speech processing can be found in Yegnanarayana and Gangashetty (2011), Drugman et al. (2014), Kadiri et al. (2020a) and Kadiri et al. (2021).

Analysis of natural signals, such as signals generated by the human speech production mechanism, helps in understanding the underlying phenomenon that is responsible for generating these signals. In general, however, analysis of natural signals is difficult because the underlying signal production system is time varying. Analysis of natural signals is typically conducted by decomposing them into mathematically tractable basis functions. The selection of the basis functions depends on the type of the signals and on the type of information to be extracted from them. The most widely used time–frequency analysis, the STFT, decomposes the signal as a linear combination of complex sinusoids. The coefficients associated with the complex sinusoids at different frequencies represent the frequency spectrum, which consists of magnitude and phase components. Both of these components are needed to reconstruct the signal. The relative importance of these components varies depending on the purpose of analysis of the signal. For example, the Fourier transform phase component carries more useful information in images than the Fourier transform magnitude component (Yegnanarayana et al., 1984; Oppenheim and Lim, 1981). In the case of speech, the Fourier transform magnitude is known to carry useful information (Yegnanarayana et al., 1984; Oppenheim and Lim, 1981). Therefore, most speech processing methods focus on analyzing and modeling the magnitude spectrum to represent the speech production characteristics.

In speech processing, the phase spectrum has received less attention because of the phase wrapping problem (Mowlaee et al., 2016; Vijayan and Murty, 2015). It is sometimes argued that the phase spectrum has a less significant role in auditory perception (Wang and Lim, 1982; Mathes and Miller, 1947; Schroeder, 1975; Mowlaee et al., 2016). Recent studies have examined the significance of phase in human speech perception (Mowlaee et al., 2016; Gerkmann et al., 2015). The importance of phase was analyzed in the perception of intervocalic stop consonants in Liu et al. (1997) by using stimuli that were reconstructed as the magnitude-only and phase-only versions of the consonants (Liu et al., 1997). The experiments in Liu et al. (1997) indicated that if the analysis segment length is longer than 50 ms the phase spectrum takes precedence over the magnitude spectrum. Similar analyses were made in Alsteris and Paliwal (2006), Paliwal and Alsteris (2005) and Paliwal and Wójcicki (2008) to study the effects of the size and shape of the analysis window on the phase of the STFT in speech perception. In Gerkmann et al. (2015), Mowlaee and Saeidi (2013), Krawczyk and Gerkmann (2014) and Paliwal et al. (2011), it was shown that enhancement of speech corrupted by noise can be achieved by suitably modifying the phase spectrum. In Quatieri and Oppenheim (1981), Oppenheim and Lim (1981) and Yegnanarayana et al. (1984), iterative algorithms were studied in phase-only recovery of minimum phase and maximum phase signals. In Alsteris and Paliwal (2004) and Schluter and Ney (2001), it was shown that the phase spectrum -based feature extraction improved the performance of automatic speech recognition. In addition, several studies have reported that features representing the phase spectrum are useful in speaker recognition and detection of synthetic speech (Nakagawa et al., 2012; Wang et al., 2010; Bastys et al., 2010; Rajan et al., 2013; Saratxaga et al., 2016). More details of the application of the phase spectrum in speech processing can be found in recent review articles (Mowlaee et al., 2016; Gerkmann et al., 2015).

Reliable estimation of the phase spectrum is problematic due to wrapping of the phase values into the interval $(-\pi\ \pi]$. The value of $\theta + 2\pi n$ (where n is an integer) is indistinguishable from θ. There may also exist discontinuities of 2π at some frequencies. With addition or subtraction of integer multiples of 2π at these discontinuities, the unwrapped phase can be retrieved in a few cases to preserve the continuity of the phase function (Mowlaee et al., 2016; Gerkmann et al., 2015). An alternative representation of the phase information is to compute the derivative of phase with respect to frequency ($\frac{d\phi}{d\omega}$) and with respect to time ($\frac{d\phi}{dt}$) (Yegnanarayana, 1978; Cohen, 1995).

The frequency derivative of the phase spectrum is called the group delay function (GDF) (Quatieri, 2004; Yegnanarayana, 1978). The GDF can be computed using the Fourier transform relations, and it does not require the explicit computation of the phase spectrum. The computation of the GDF involves division by the squared magnitude spectrum. Therefore, the GDF shows high-amplitude peaks at the spectral nulls which are due to zeros of the transfer function close to the unit circle in the z-plane (Bozkurt et al., 2007). To reduce the effects of these high-amplitude peaks, the GDF can be conditioned using the cepstrally smoothed magnitude spectrum (Murthy and Yegnanarayana, 2011) or the GDF can be multiplied with the squared magnitude spectrum to cancel its denominator (Zhu and Paliwal, 2004). It is worth emphasizing that many of the GD-based methods utilize information from minimum phase equivalent signals (Murthy and Yegnanarayana, 2011).

Another representation is the time derivative of the phase, called instantaneous frequency (IF). IF carries information about the local frequency behavior as a function of time (Boashash, 1992; Cohen, 1995). Most of the methods for the computation of IF use the derivative of the phase of the analytic signal directly, and are therefore subject to the phase wrapping problem (Murty and Yegnanarayana, 2008; Vijayan et al., 2016). IF has been computed using STFT or a filter bank to extract useful information of speech production such as formant contours and formant bandwidths (Costas, 1981; Ramalingam et al., 1994; Kumaresan et al., 1994; McCowan et al., 2011; Stark and Paliwal, 2008; Vijayan et al., 2019; Tsiakoulis et al., 2013). In most of these studies, the random effects of phase wrapping and the effects of discontinuities due to excitation are smoothed out by averaging IFs over time and frequency.

It is to be noted that the alternative representations of phase, i.e., GD and IF, have been used mainly in basic speech analysis like in pitch estimation (Kawahara et al., 2016), formant extraction (Murthy and Yegnanarayana, 1991; Kumaresan and Rao, 1999) and spectrum estimation (Yegnanarayana and Murthy, 1992; Stark and Paliwal, 2008). However, there is little previous effort in using the phase information to extract the excitation characteristics of voiced speech. The most important characteristics are the discontinuities in the speech signals caused by abrupt closure of the vocal folds. This may be due to computational issues and also due to effects of the size and shape of the analysis window used for processing speech signals (Smits and Yegnanarayana, 1995).

In Vijayan and Murty (2016), the authors exploited the phase information to detect the discontinuities at the epochs. The error associated with the all-pass modeling, referred to as the all-pass residual, was used to characterize the excitation. The all-pass residual

exhibits prominent peaks at the epochs. As the strength of the peaks in the error signal varies with time, a peak-to-neighborhood-energy-ratio measure was used to identify the peaks corresponding to the epochs. The epochs obtained from the error signal were refined using a dynamic programming algorithm. It is to be noted that this method uses modeling in order to estimate the source and system information. In Murty and Yegnanarayana (2008), an attempt was made to obtain the locations of the discontinuities by computing the IF of the filtered output. This approach depends on the choice of the center frequency of the filter.

In the present study, the IFs computed by three different speech analysis methods are explored to highlight the discontinuities in the signal. The discontinuities in the IF are identified as the epochs. The highlights of the present study are as follows:

- A systematic investigation of the IF computed using STFT, narrow bandpass filtering (NBPF) and single frequency filtering (SFF) is carried out to highlight the discontinuities in the signal.
- As information about discontinuities is present at all frequencies, the IFs computed at all frequencies are combined to highlight epochs.
- To further enhance the detection of epochs, a method based on the amplitude-weighted IFs is proposed.
- The proposed epoch extraction method is compared with existing methods using three databases, which include both clean and telephone quality speech. Performance of the proposed epoch extraction method is shown to be comparable with the performance of the existing methods for clean speech, and is better for telephone quality speech.

The organization of the paper is as follows. A general description of the IF is presented first in Section 2. This section discusses the computation of the IF both for analytic signal and for multicomponent signals using STFT, NBPF and SFF. In Section 3, the IF is derived for synthetic signals (aperiodic impulse sequences and synthetic speech signals) and for natural speech signals. Section 4 deals with extraction of impulse-like discontinuities in speech signals. A method for epoch extraction using the IF is described in Section 5. The proposed epoch extraction method is compared with existing methods for clean speech and telephone quality speech in Section 6. Finally, a summary of the studies made in this paper is given in Section 7.

2. Instantaneous frequency (IF)

In this section, the basic definition of IF applicable for monocomponent signal is given. The IF is the derivative of phase of an analytic signal. The concept of IF is extended to multicomponent signals, by defining IF for each component. The multicomponent signal is decomposed into several monocomponents using three different methods, namely short-time Fourier transform (STFT), narrow bandpass filtering (NBPF) and single frequency filtering (SFF). The discontinuities in the IF of each component are caused by the discontinuities in the signal due to impulse-like excitation. By combining the IFs due to all components the impulse characteristics are highlighted at the epoch locations. The three different methods of decomposition of a multicomponent signal yield different phase characteristics. The impulse characteristics in the combined IFs, and their effectiveness in the identification of the epoch locations from these three methods are examined in the subsequent sections.

2.1. IF of analytic signal

The IF of a real signal $x(t)$ is defined as the time derivative of the unwrapped phase of the complex analytic signal $x_a(t)$ of $x(t)$ (Cohen, 1995; Kadiri, 2018). The $x_a(t)$ is given by

$$x_a(t) = x(t) + jx_H(t), \tag{1}$$

where $x_H(t)$ is the Hilbert transform of $x(t)$ (Cohen, 1995). Writing $x_a(t)$ in polar form, we get

$$x_a(t) = |x_a(t)|e^{j\phi(t)}, \tag{2}$$

where

$$|x_a(t)| = \sqrt{x^2(t) + x_H^2(t)}, \tag{3}$$

$$\phi(t) = \tan^{-1}\left(\frac{x_H(t)}{x(t)}\right), \tag{4}$$

are the instantaneous amplitude and phase, respectively. Taking logarithm on both sides of Eq. (2), we get

$$\log x_a(t) = \log|x_a(t)| + j\phi(t), \tag{5}$$

Taking derivative with respect to time t, we get

$$\frac{d}{dt}\log x_a(t) = \frac{x_a'(t)}{x_a(t)} = \frac{d}{dt}\log|x_a(t)| + j\phi'(t), \tag{6}$$

where the superscript $'$ denotes the derivative operator. The $\phi'(t)$ is the IF, and is given by

$$\phi'(t) = \Im\left(\frac{x_a'(t)}{x_a(t)}\right), \tag{7}$$

where $\Im(.)$ denotes the imaginary part.

The IF $\phi'(t)$ can be obtained using the Fourier transform $X_a(\omega)$ of $x_a(t)$ as follows. The analytic signal $x_a(t)$ in terms of $X_a(\omega)$ is given by

$$x_a(t) = \frac{1}{2\pi}\int_0^\infty X_a(\omega)e^{j\omega t}d\omega, \tag{8}$$

Note that $X_a(\omega)$ exists only for the positive frequencies. Taking the derivative of $x_a(t)$ with respect to time, we get

$$\begin{aligned} x_a'(t) &= \frac{1}{2\pi}\int_0^\infty X_a(\omega)e^{j\omega t}(j\omega)d\omega, \\ &= j\left(\frac{1}{2\pi}\int_0^\infty (\omega X_a(\omega))e^{j\omega t}d\omega\right), \\ &= j\text{IFT}\left(\omega X_a(\omega)\right). \end{aligned} \tag{9}$$

where IFT is the inverse Fourier transform.

The IF can be obtained from Eqs. (7) and (9) as

$$\phi'(t) = \Re\left(\frac{\text{IFT}(\omega X_a(\omega))}{\text{IFT}(X_a(\omega))}\right), \tag{10}$$

where $\Re(.)$ denotes the real part.

The IF can be interpreted as the frequency of a sinusoid that fits the signal under analysis. The IF as a function of time shows the deviation of the frequency of the signal from the monotone at every instant of time. For a multicomponent signal, i.e., for signal consisting of multiple sinusoids, the IF is defined for each frequency component. The IFs vary depending on how the multicomponent signal is decomposed into individual frequency components. Three decomposition methods (STFT, NBPF and SFF) of multicomponent signals are considered in this study.

2.2. Decomposition by short-time Fourier transform (STFT)

The discrete-time STFT of the signal $x[n]$ at time n is given as (Rabiner and Schafer, 2010)

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega m}, \tag{11}$$

where $w[n]$ is the analysis window in the interval 0 to $(N_w - 1)$. The value of $X(n,\omega)$ at any frequency ω_k is given by

$$\begin{aligned} X(n,\omega_k) &= e^{-j\omega_k n}(x[n] * w[n]e^{j\omega_k n}), \\ &= |X(n,\omega_k)|e^{j\theta(n,\omega_k)}. \end{aligned} \tag{12}$$

The term $h_k[n] = w[n]e^{j\omega_k n}$ is the frequency-shifted window function. The kth frequency component $y_k[n]$ of $x[n]$ from the STFT decomposition is given by

$$\begin{aligned} y_k[n] &= x[n] * h_k[n], \\ &= e^{j\omega_k n}X(n,\omega_k). \end{aligned} \tag{13}$$

The kth frequency component of the signal can be written (using Eq. (12)) as Quatieri (2004)

$$y_k[n] = |X(n,\omega_k)|e^{j[\omega_k n + \theta(n,\omega_k)]}. \tag{14}$$

2.3. Decomposition by narrow bandpass filtering (NBPF)

A multicomponent signal can be decomposed into monocomponent signals using a resonator at each frequency. The resonator is a second-order IIR filter with a pair of complex conjugate poles ($re^{\pm jw_k}$) in the z-plane. The system function for the kth resonator is given by

$$H_k(z) = \frac{1}{1 + a_1 z^{-1} + + a_2 z^{-2}}, \tag{15}$$

where $a_1 = -2r\cos(\omega_k T)$, $a_2 = r^2$, and T=$1/f_s$ is the sampling interval or inverse of the sampling frequency (f_s). The value of r defines the bandwidth of the resonator, with smaller values of r ($r \ll 1$) corresponding to larger bandwidths. In this study $r = 0.995$ (corresponds to bandwidth of ≈12 Hz) is used. The output $y_k[n]$ of the kth resonator for the input $x[n]$ is given by

$$y_k[n] = -a_1 y_k[n-1] - a_2 y_k[n-2] + x[n]. \tag{16}$$

The $y_k[n]$ corresponds to the kth frequency component of the NBPF output.

2.4. Decomposition by single frequency filtering (SFF)

In SFF, the component signals are obtained by passing the frequency-shifted signal $x_k[n] = x[n]e^{-j\bar{\omega}_k n}$ through a single pole resonator, with the pole located on the negative real axis in the z-plane at $z = -r$ (r defines the bandwidth) (Aneeja and Yegnanarayana, 2015; Kadiri and Yegnanarayana, 2017). The $\bar{\omega}_k = \pi - \omega_k$, where ω_k is the desired frequency. The system function of the single pole resonator is given by

$$H(z) = \frac{1}{1 + rz^{-1}}. \tag{17}$$

The kth frequency component is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{18}$$

It should be noted that $y_k[n]$ is a complex signal with real part $y_{kr}[n]$ and imaginary part $y_{ki}[n]$.

By writing $y_k[n]$ in polar form, we get

$$y_k[n] = |v_k[n]|e^{j\phi_k[n]}, \tag{19}$$

with

$$v_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}, \tag{20}$$

$$\phi_k[n] = \tan^{-1}\left(\frac{y_{ki}[n]}{y_{kr}[n]}\right), \tag{21}$$

where $v_k[n]$ is the temporal amplitude and $\phi_k[n]$ is the temporal phase.

Note that SFF differs from NBPF, as the filter is fixed in SFF, whereas in NBPF the filter varies with the frequency of the component. In addition, the filtering at the highest frequency of $f_s/2$ helps to capture the magnitude and phase information at each frequency as in the case of modulation of a carrier frequency. Both in NBPF and SFF, the effects of windowing are avoided due to filtering operation.

2.5. IFs for multicomponent signals

The phase function of the component signals in Eqs. (14), (16) and (18) are different due to the manner in which the signal is decomposed by STFT, NBPF and SFF, respectively. The characteristics of IFs derived from these phase functions will be different too, although all of them are expected to show discontinuities at the impulse locations of the signal.

The IF for the kth frequency component is obtained using the inverse discrete Fourier transform (IDFT) as follows (Murty and Yegnanarayana, 2008):

$$\Omega_k[n] = \frac{2\pi}{N}\Re\left(\frac{\text{IDFT}(lY_{ka}[l])}{\text{IDFT}(Y_{ka}[l])}\right), \tag{22}$$

where $Y_{ka}[l]$ is the discrete Fourier transform of the analytic signal ($y_{ka}[n]$), and N is the total number of samples in the signal. The complex analytic signal $y_{ka}[n]$ of $y_k[n]$ is given by

$$y_{ka}[n] = y_k[n] + jy_{k_H}[n], \tag{23}$$

where $y_{k_H}[n]$ is the Hilbert transform of $y_k[n]$. The IF $\Omega_k[n]$ of each component shows deviation from the corresponding frequency ω_k, except in the case of SFF, where all the IFs show deviation from π, corresponding to the half the sampling frequency. The combined IF $\Omega[n]$ is obtained by summing deviations of all K IFs from the corresponding frequencies. Thus, combined IFs can be written as

$$\Omega[n] = \sum_{k=1}^{K}(\Omega_k[n] - \omega_k). \tag{24}$$

for the STFT and NBPF methods, and

$$\Omega[n] = \sum_{k=1}^{K}(\Omega_k[n] - \pi). \tag{25}$$

for the SFF method.

3. IF for synthetic and natural signals

In this section, the IF is computed for synthetic signals and natural voiced speech signals using the three methods of decomposition described in the previous section. Fig. 1(a) shows a multicomponent synthetic signal, which is an aperiodic sequence of impulses with arbitrary amplitude values. Figs. 1(b), 1(c) and 1(d) show the IFs computed at 500 Hz using STFT, NBPF and SFF, respectively. The parameters used are: f_s=8 kHz and $r = 0.995$. The filtered signals are obtained for every 5 Hz for both NBPF and

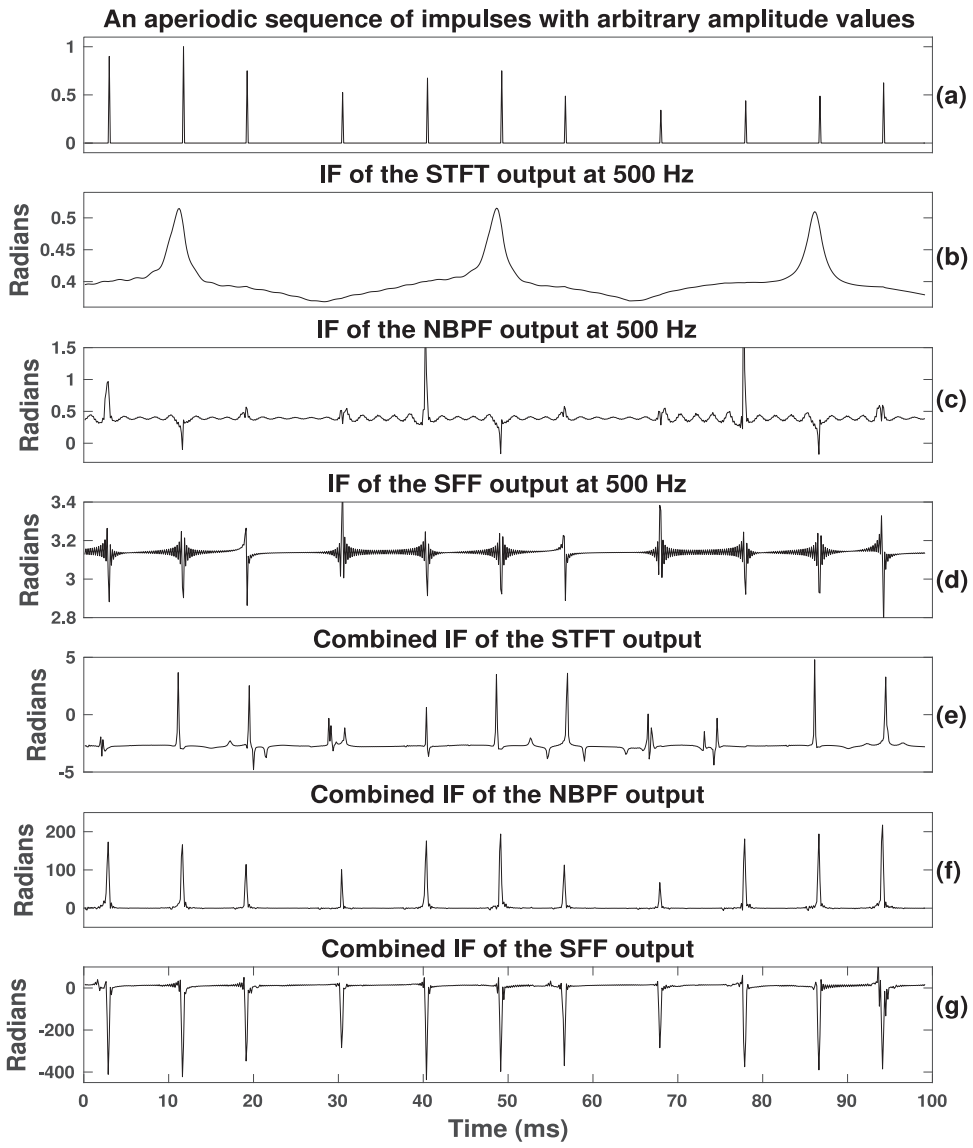

Fig. 1. Illustration of the IF for an aperiodic sequence of impulses. (a) An aperiodic sequence of impulses with arbitrary amplitude values. (b) IF of the STFT output at 500 Hz. (c) IF of the NBPF output at 500 Hz. (d) IF of the SFF output at 500 Hz. (e) Combined IF of the STFT output. (f) Combined IF of the NBPF output. (g) Combined IF of the SFF output.

SFF. In the case of STFT, a 1024-point FFT with a 30 ms Hann window and 1-sample shift is used. It is to be noted that in the case of STFT and NBPF, the IF deviates from the normalized center frequency $\omega_k = \frac{2\pi}{f_s} f = \frac{2\pi}{8000}(500) = 0.3927$, and in the case of SFF, the IF deviates from π radians. In all the cases, it can be observed that the IF shows discontinuities at the instants of impulse locations, although the discontinuities do not manifest themselves equally at all the time instants, especially for STFT. Since the impulse information is present at all frequencies, the IFs at all frequencies are combined as indicated in Eqs. (24) and (25). Figs. 1(e), 1(f) and 1(g) show the combined IF plots obtained for STFT, NBPF and SFF, respectively. Since the IF is subtracted from the center frequency for all the frequencies, the combined IF deviates from zero. From the figures, it can be seen that the combined IF plots highlight the discontinuities clearly. The combined IF plots obtained from NBPF and SFF show sharp discontinuities at the instants of the impulse locations. It is worth noting that the relative amplitude patterns of the discontinuities in the combined IFs of NBPF and SFF are similar to the amplitude patterns of the aperiodic impulse-sequence shown in Fig. 1(a).

Fig. 2 illustrates IFs obtained for a synthetic voiced speech segment (shown in Fig. 2(a)) generated by exciting a 10th order all-pole filter using the Liljencrants–Fant (LF) (Fant, 1995) model (shown in Fig. 2(b)). Figs. 2(c), 2(d) and 2(e) show the IFs computed at 500 Hz, using STFT, NBPF and SFF, respectively. In all these cases, it is difficult to distinguish discontinuities caused by the LF excitation, although they can be seen to some extent in the IF plots computed using NBPF. Figs. 2(f), 2(g) and 2(h) show the

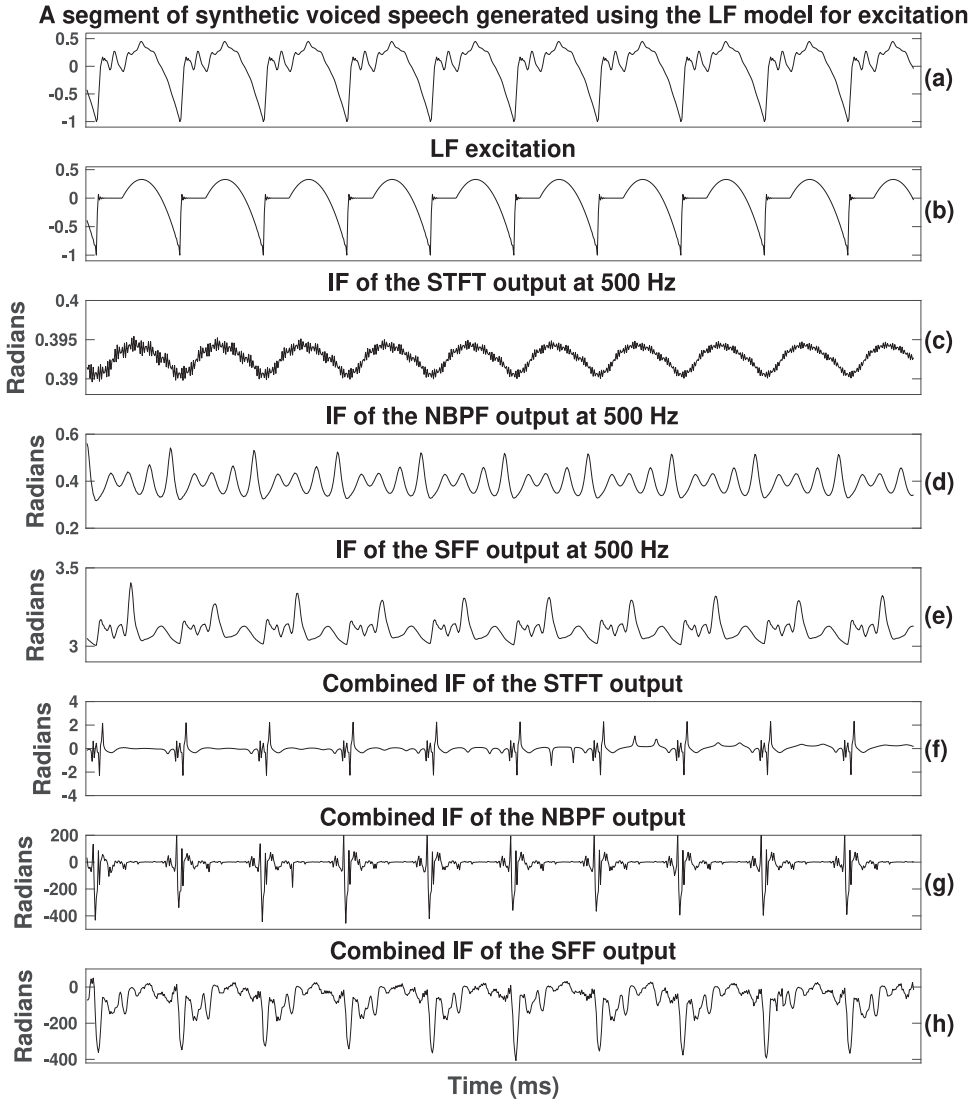

Fig. 2. Illustration of the IF for synthetic voiced speech. (a) A segment of synthetic voiced speech generated using the LF model for excitation. (b) LF excitation. (c) IF of the STFT output at 500 Hz. (d) IF of the NBPF output at 500 Hz. (e) IF of the SFF output at 500 Hz. (f) Combined IF of the STFT output. (g) Combined IF of the NBPF output. (h) Combined IF of the SFF output.

combined IF plots obtained by STFT, NBPF and SFF. From the figures, it can be observed that the combined IFs highlight the discontinuity information clearly.

An illustration of IFs for a natural voiced speech signal is shown in Fig. 3. Figs. 3(a) and 3(b) show the segment of voiced speech and the corresponding differentiated electroglottography (dEGG) signal. The negative peaks in the dEGG signal correspond to the glottal closure instants (GCIs) or epochs. Figs. 3(c), 3(d) and 3(e) show the IFs computed at 500 Hz using NBPF, STFT and SFF, respectively. In all these cases, it is difficult to see discontinuities that are associated with the negative peaks of the dEGG signal. Figs. 3(f), 3(g) and 3(h) show the combined IF plots obtained using STFT, NBPF and SFF, respectively. The combined IF plots in Figs. 3(f) and 3(g) do not show the discontinuities due to impulse-like excitations consistently. The combined IFs obtained using SFF (shown in Fig. 3(h)) show the discontinuities clearly at all the GCIs, and they match well with the GCIs in the dEGG signal (shown in Fig. 3(b)).

4. Extraction of impulse-like discontinuities from speech signals

In Section 3, it was observed that the combined IF obtained using the SFF method highlights the impulse-like discontinuity information. In this section, the SFF method is used for extraction of the impulse-like discontinuities from speech signals. In Fig. 3(h), it can be observed that even though the combined IFs obtained using the SFF method highlights the impulse-like behavior, it contains

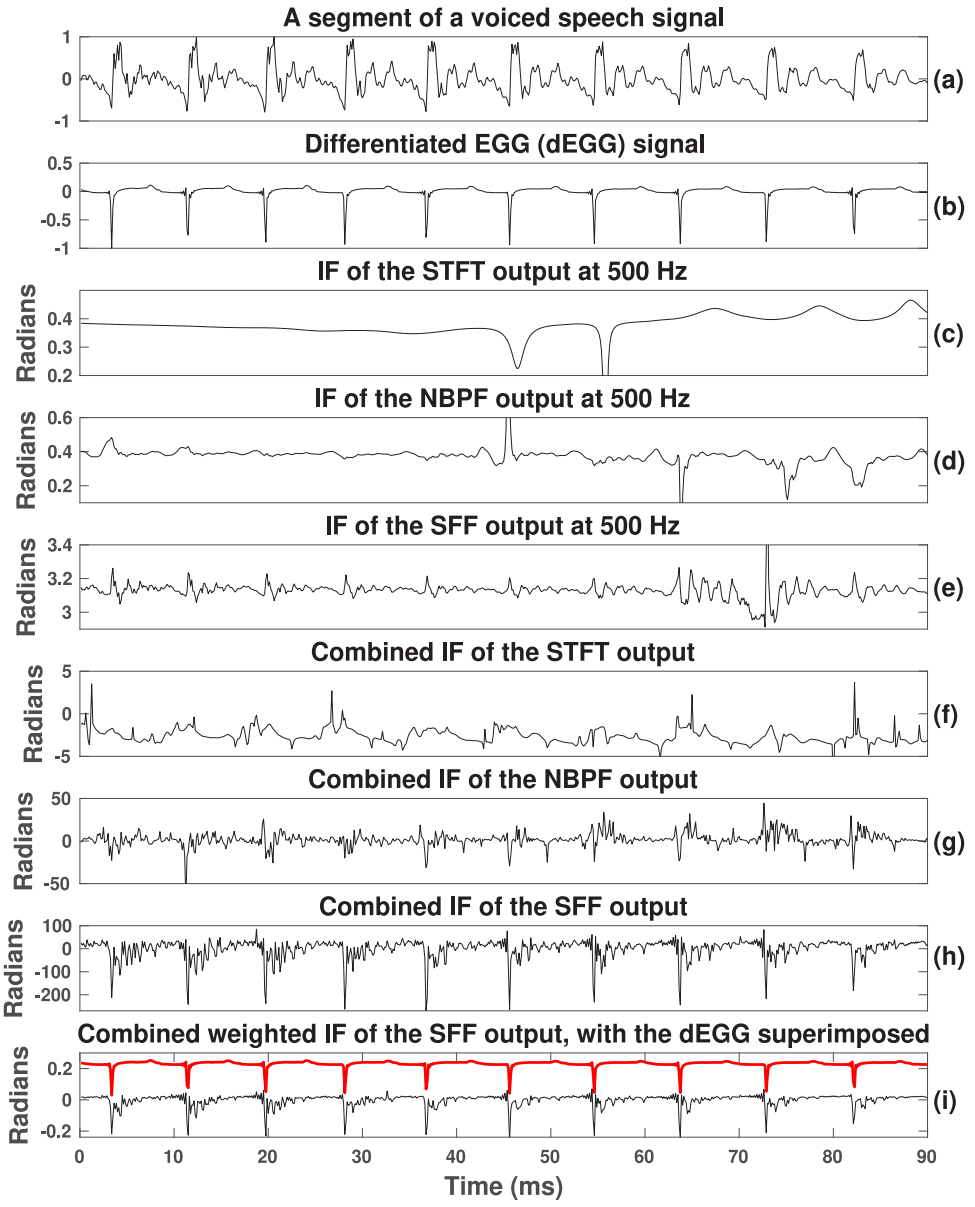

Fig. 3. Illustration of the IF for a segment of natural voiced speech. (a) A segment of a voiced speech signal. (b) Differentiated EGG (dEGG) signal. (c) IF of the STFT output at 500 Hz. (d) IF of the NBPF output at 500 Hz. (e) IF of the SFF output at 500 Hz. (f) Combined IF of the STFT output. (g) Combined IF of the NBPF output. (h) Combined IF of the SFF output. (i) Combined weighted IF of the SFF output, with the dEGG of Fig. 3(b) superimposed.

fluctuations in other parts of the signal. To reduce such fluctuations, amplitude weighting of the IF is proposed. The amplitude weighting of the IF is first defined for the kth frequency as follows

$$\Omega_{AW_k}[n] = \frac{\sum_{n=1}^{N} v_k[n]\Omega_k[n]}{\sum_{n=1}^{N} v_k[n]}, \qquad k = 1, 2, \ldots, K. \tag{26}$$

where $v_k[n]$ is given in Eq. (20).

By combining $\Omega_{AW_k}[n]$ for all frequencies, the combined amplitude weighted IF, called AIF, is defined as

$$\Omega_{AW}[n] = \sum_{k=1}^{K} \Omega_{AW_k}[n] = \sum_{k=1}^{K} \left(\frac{\sum_{n=1}^{N} v_k[n]\Omega_k[n]}{\sum_{n=1}^{N} v_k[n]} \right). \tag{27}$$

Fig. 3(i) shows the AIF computed for the voiced speech segment of Fig. 3(a). From Fig. 3(i), it can be clearly seen that the fluctuations are reduced in the AIF compared to the unweighted IF shown in Fig. 3(h). The discontinuities in the AIFs match well

with the discontinuities in the dEGG signal shown in Fig. 3(i). In the next section, a method is proposed to detect epochs based on AIF.

5. Epoch extraction

Using the AIF defined in Eq. (27), an epoch extraction method is proposed, which is motivated by previous studies in epoch extraction, particularly, the zero frequency filtering (ZFF) (Murty and Yegnanarayana, 2008), and the speech event detection using linear prediction (LP) residual and mean based signal (SEDREAMS) (Drugman and Dutoit, 2009). In both of these epoch extraction methods, the impulse-like excitation characteristics of speech signals are captured by using a filtering operation, which yields an oscillatory signal that varies with the local pitch period. Both the ZFF and SEDREAMS methods have been shown to work well for clean speech (Drugman et al., 2012). However, these methods failed to detect epochs accurately from telephone quality speech because of the attenuated level of low-frequency components in telephone quality speech (Kadiri, 2019; Kadiri and Yegnanarayana, 2020). We examine the effectiveness of the AIF for epoch extraction, especially for telephone quality speech. A oscillatory signal that varies with the local pitch period is derived using the ZFF-based approach (Murty and Yegnanarayana, 2008). The AIF signal is passed through a cascade of two zero frequency resonators as given in Eq. (28). The precise location of the epoch is obtained by searching for the minimum of the AIF in the region between the minimum and maximum of the filtered signal, i.e., around each negative-to-positive zero crossing.

The proposed epoch extraction method based on the AIF, called as the EAIF method, consists of the following steps:

1. The AIF signal in Eq. (27) is passed through a cascade of two ideal zero frequency resonators. The output of this filtering operation is given by (Murty and Yegnanarayana, 2008)

$$z_0[n] = -\sum_{k=1}^{4} a_k z_0[n-k] + \Omega_{AW}[n], \tag{28}$$

   where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$. This filtering is equivalent to cumulative sum in the discrete-time domain. The resulting signal $z_0[n]$ grows or decays approximately as a polynomial function of time.
2. The trend in $z_0[n]$ is removed by subtracting the local mean computed over the average pitch period (Murty and Yegnanarayana, 2008; Drugman and Alwan, 2011) at each sample as follows:

$$z[n] = z_0[n] - \frac{1}{2P+1}\sum_{m=-P}^{P} z_0[n+m]. \tag{29}$$

   Here, $2P+1$ corresponds to the number of samples in the window used for trend removal.
3. The region (interval) between the minimum to maximum of the filtered signal $z[n]$ around each negative-to-positive zero crossing is hypothesized as the region of the epoch.
4. The location of the minimum of the AIF signal in the hypothesized region/interval is marked as epoch or GCI.

The steps involved in the proposed EAIF method to extract epochs are illustrated in Fig. 4. Fig. 4(a) shows a segment of voiced speech and Fig. 4(b) shows the corresponding AIF. Fig. 4(c) shows the filtered signal obtained from the AIF signal in Fig. 4(b). Fig. 4(d) shows the intervals derived from the filtered signal where epochs are present (regions between the minimum to maximum around each negative-to-positive zero crossing of the filtered signal). The precise epoch locations are estimated by finding the minimum of the AIF signal shown in Fig. 4(b) in the intervals of epoch presence shown in Fig. 4(d). Fig. 4(e) shows the estimated epoch locations, which are marked by downward arrows (in red), along with the reference epochs shown by the dEGG signal. It can be clearly seen that the estimated epochs match well with the reference epochs indicated by the negative peaks of the dEGG signal.

6. Performance evaluation

Performance of the proposed epoch extraction method was evaluated using data from three databases, which contain both speech signals and the simultaneously recorded EGG waveforms (Kominek and Black, 2004). The data in these databases was collected in clean laboratory conditions. In order to test the epoch extraction method for telephone quality speech, the clean speech data were encoded using a low bit-rate speech codec (ITU-T, Recommendation, 2005). Epoch extraction was then carried out on both the clean signal and its coded version. The dEGG signals were used to obtain the ground truth. The results of the proposed method were compared with four state-of-the-art epoch extraction methods using the evaluation metrics defined in Naylor et al. (2007).

6.1. Experimental protocol

This section describes the databases, epoch extraction methods for comparison and evaluation metrics.

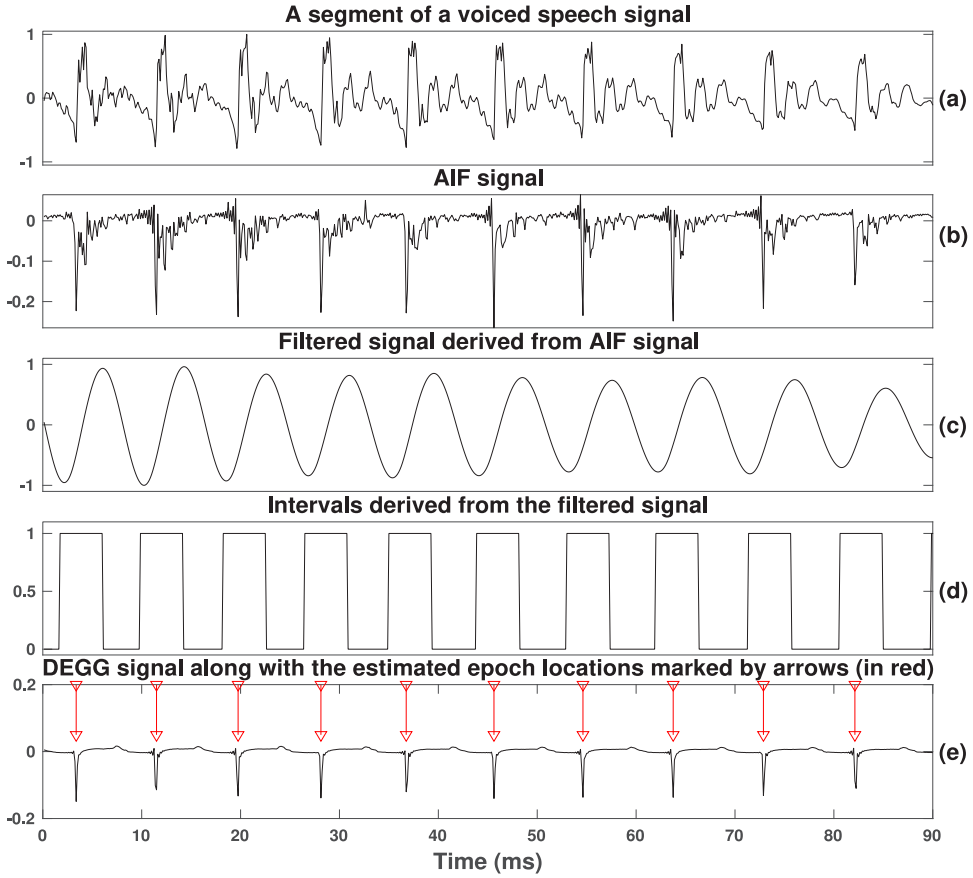

Fig. 4. Illustration of epoch extraction using the proposed EAIF method for a segment of voiced speech. (a) A segment of voiced speech. (b) AIF signal. (c) Filtered signal derived from (b). (d) Intervals derived from the filtered signal. (e) Differentiated EGG (dEGG) signal along with the estimated epoch locations marked by downward arrows (in red).

6.1.1. Databases

The epoch extraction methods were evaluated using speech and simultaneously recorded EGG waveforms of five speakers obtained from three databases. The data of the first three speakers were taken from the CMU ARCTIC database (Kominek and Black, 2004). The speech data of each speaker (BDL (male), JMK (male) and SLT (female)) consists of around 1132 phonetically-balanced English sentences. The data of the fourth speaker (KED (male)) was taken from the KED TIMIT database. This data consists of 453 sentences. The data of the fifth speaker (RAB (male)) was taken from the RAB database. The data consists of a set of 1946 nonsense words containing phone–phone transitions in English. All these databases are available on the Festvox webpage http://festvox.org/cmu_arctic/index.html. The speech and EGG signals were aligned by compensating the larynx-to-microphone delay (around 0.9 ms for BDL, JMK, SLT, 0.6 ms for KED, and 2.3 ms for RAB). Reference epoch locations were identified as the locations of the negative peaks in the dEGG signal. Telephone quality speech was simulated by processing the speech signals taken from the three databases with a narrowband codec, which was implemented using the G.191 software (ITU-T, Recommendation, 2005).

6.1.2. Epoch extraction methods for comparison

The following four epoch extraction methods were used for comparison to the proposed EAIF approach: Zero frequency filtering (ZFF) (Murty and Yegnanarayana, 2008), speech event detection using LP residual and mean based signal (SEDREAMS) (Drugman et al., 2012), most singular manifold (MSM) (Khanagha et al., 2014) and yet another GCI algorithm (YAGA) (Thomas et al., 2012). It is to be noted that in the proposed EAIF method, the AIF signal is passed through a cascade of two zero frequency resonators as opposed to passing the speech signal as in the original ZFF method (Murty and Yegnanarayana, 2008).

6.1.3. Evaluation metrics

Performance of the epoch extraction methods was evaluated using five widely used measures described in Naylor et al. (2007). These five evaluation metrics are the identification rate (IDR1), miss rate (MR), false alarm rate (FAR), identification accuracy (IDA) and identification rate within ±0.25 ms (IDR2). The first three measures are called reliability measures, and the other two are called

Table 1
Performance comparison of the epoch extraction methods averaged across all speakers for clean speech and telephone quality speech. IDR1—Identification rate, MR—Miss rate, FAR—False alarm rate, IDA—Identification accuracy in ms, IDR2—Identification rate within ±0.25 ms.

Condition	Method	IDR1%	MR%	FAR%	IDA(ms)	IDR2%
Clean	ZFF	**99.14**	0.38	0.48	0.43	62.12
	SEDREAMS	98.85	0.74	**0.41**	**0.37**	81.66
	MSM	91.47	3.34	5.19	0.51	75.48
	YAGA	97.63	**0.33**	2.04	**0.37**	**83.45**
	EAIF	97.61	1.09	1.30	0.42	75.85
Telephone quality	ZFF	46.92	1.71	51.37	0.92	32.20
	SEDREAMS	38.68	**1.49**	59.83	0.83	31.20
	MSM	86.32	5.02	8.66	0.68	54.93
	YAGA	37.94	1.94	60.12	0.83	28.87
	EAIF	**93.55**	1.62	**4.83**	**0.64**	**57.69**

Table 2
Performance comparison of the epoch extraction methods for clean speech. IDR1—Identification rate, MR—Miss rate, FAR—False alarm rate, IDA—Identification accuracy in ms, IDR2—Identification rate within ±0.25 ms.

Speaker	Method	IDR1%	MR%	FAR%	IDA(ms)	IDR2%
BDL	ZFF	**99.12**	0.49	**0.39**	0.31	82.02
	SEDREAMS	99.09	0.51	0.40	0.29	83.54
	MSM	94.41	2.78	2.81	0.45	74.39
	YAGA	97.61	**0.32**	2.07	**0.23**	**86.11**
	EAIF	97.11	0.85	2.04	0.36	71.98
JMK	ZFF	**99.19**	**0.30**	0.51	0.49	48.60
	SEDREAMS	97.53	2.46	0.01	0.39	71.28
	MSM	94.14	5.58	0.28	0.41	72.21
	YAGA	99.05	0.83	**0.12**	**0.37**	**75.56**
	EAIF	97.67	1.79	0.54	0.44	64.37
SLT	ZFF	**99.28**	0.40	0.32	0.22	81.64
	SEDREAMS	99.19	0.63	**0.18**	0.26	76.02
	MSM	94.57	5.20	0.23	0.36	65.17
	YAGA	99.02	**0.39**	0.59	**0.21**	**82.78**
	EAIF	98.26	0.73	1.01	0.34	78.97
KED	ZFF	**99.91**	0.04	0.05	0.54	47.91
	SEDREAMS	99.90	0.06	**0.04**	**0.46**	90.60
	MSM	96.78	1.32	1.90	0.53	90.91
	YAGA	98.65	**0.03**	1.32	0.47	**91.53**
	EAIF	98.57	0.38	1.05	**0.46**	86.83
RAB	ZFF	98.23	0.66	**1.11**	0.61	50.42
	SEDREAMS	**98.57**	**0.04**	1.39	**0.47**	**86.87**
	MSM	77.47	1.82	20.71	0.79	74.74
	YAGA	93.82	0.09	6.09	0.58	81.27
	EAIF	96.44	1.72	1.84	0.51	77.11

accuracy measures. The better the performance of an epoch extraction method, the higher are the values of IDR1 and IDR2, and the lower are the values of MR, FAR and IDA.

6.2. Results and discussion

All the epoch extraction methods were evaluated against the ground truth epoch information provided by the EGG. The average performance of epoch extraction across all the five speakers (BDL, JMK, SLT, KED, and RAB) is shown for clean and telephone quality speech in Table 1. Tables 2 and 3 show the results for each speaker for clean and telephone quality speech, respectively.

For clean speech (Tables 1 and 2), the results show that the performance of the proposed EAIF method is comparable or better than the existing methods in both reliability (i.e., in IDR, MR, and FAR) and accuracy (i.e., in IDA and IDR2). From the reference methods, ZFF and SEDREAMS are most reliable, and YAGA is the most accurate for clean speech.

For telephone quality speech (Table 3), the results indicate that the performance of the existing methods is severely affected for all the speakers compared to clean speech (Table 2). It can be observed that even though the performance of the MSM method is lower compared to the other methods (ZFF, SEDREAMS, and YAGA) in clean speech, the performance of MSM is higher for telephone quality speech for all the speakers. On the other hand, the performance of the EAIF method exceeds the performance of all the other methods for all the speakers in both reliability and accuracy. Overall, the results of the proposed EAIF method are comparable with those obtained using the state-of-the-art epoch extraction methods for clean speech, and better for telephone quality speech.

The key findings of this study are as follows:

Table 3
Performance comparison of the epoch extraction methods for telephone quality speech. IDR1—Identification rate, MR—Miss rate, FAR—False alarm rate, IDA—Identification accuracy in ms, IDR2—Identification rate within ±0.25 ms.

Speaker	Method	IDR1%	MR%	FAR%	IDA(ms)	IDR2%
BDL	ZFF	45.91	1.29	52.80	0.89	29.19
	SEDREAMS	45.12	2.91	51.97	0.81	26.73
	MSM	88.79	5.12	6.09	0.67	48.19
	YAGA	35.19	1.97	62.84	0.94	28.19
	EAIF	**95.38**	**0.83**	**3.79**	**0.62**	**54.72**
JMK	ZFF	34.72	2.94	62.34	0.91	27.31
	SEDREAMS	32.81	1.12	66.09	1.17	28.12
	MSM	92.54	2.43	5.03	0.71	53.09
	YAGA	35.15	3.34	61.51	0.95	30.11
	EAIF	**96.45**	**1.01**	**2.54**	**0.63**	**54.88**
SLT	ZFF	68.49	0.77	30.74	0.76	42.22
	SEDREAMS	63.19	**0.86**	35.95	0.70	50.11
	MSM	88.67	6.44	4.89	0.63	52.40
	YAGA	68.19	1.19	30.62	0.74	41.81
	EAIF	**95.11**	2.75	**2.14**	**0.61**	**57.06**
KED	ZFF	40.70	1.38	57.92	1.17	30.19
	SEDREAMS	23.19	0.86	75.95	0.79	28.92
	MSM	93.48	2.38	4.14	0.58	**68.31**
	YAGA	34.18	0.89	64.93	0.92	32.14
	EAIF	**98.04**	**0.38**	**1.58**	**0.54**	65.73
RAB	ZFF	44.76	2.16	53.08	0.89	32.11
	SEDREAMS	29.11	**1.72**	69.17	0.66	22.12
	MSM	68.11	8.73	23.16	0.79	52.67
	YAGA	16.99	2.29	80.72	**0.59**	12.11
	EAIF	**82.75**	3.15	**14.10**	0.79	**56.04**

- The IF derived using SFF analysis for decomposition of speech signal into individual frequency components highlights the discontinuities corresponding to the GCIs/epochs in the speech signal.
- The IF derived using the SFF-based decomposition is shown to highlight the discontinuities better than the STFT and NBPF based decomposition methods.
- The amplitude weighted IF (AIF) is used to develop a method for epoch extraction from clean and telephone quality speech.
- The proposed EAIF epoch extraction method is shown to perform better than four known state-of-the-art methods, in terms of reliability and accuracy, especially for telephone quality speech.

7. Summary and conclusion

In this paper, the excitation information of speech was extracted using the phase component, by highlighting the discontinuities at epochs. The impulse-like characteristics in the speech signal were derived from the IF of the filtered signal at each frequency, obtained by STFT, NBPF and SFF. The sum of the IFs of all the filtered signals shows discontinuities at the locations of the impulses in the excitation. It was observed that the combined IF obtained from the SFF method highlighted the impulse-like discontinuity in the speech signals better compared to the STFT and NBPF methods. The impulse-like discontinuity was highlighted further by amplitude weighting the IF, resulting in the combined amplitude weighted IF, the AIF. An epoch extraction method, called EAIF, was proposed based on the AIF. The performance of the proposed EAIF method was compared with four state-of-the-art epoch extraction methods for clean and telephone quality speech. The performance of EAIF was comparable to the existing methods for clean speech, and better for telephone quality speech, both in terms of reliability and accuracy. The performance improvement is due to the following two features. First, the EAIF method exploits the impulse-like discontinuity in IF, which improves accuracy. Second, the filtered signals derived from the AIF oscillate with the local pitch period yielding improved reliability. The impact of noise on IF computation is a topic for future studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is publicly available: http://festvox.org/cmu_arctic/index.html

Acknowledgments

This work was partly supported by the Academy of Finland (project no. 330139) and Aalto University, Finland (the Ministry of Education and Culture's Global Program Pilots for India). The third author would like to acknowledge the support of Indian National Science Academy (INSA).

References

Airaksinen, M., Juvela, L., Bollepalli, B., Yamagishi, J., Alku, P., 2018. A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (9), 1658–1670.

Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., Story, B., 2009. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. J. Acoust. Soc. Am. 120, 3289–3305.

Alsteris, L.D., Paliwal, K.K., 2004. ASR on speech reconstructed from short-time Fourier phase spectra. In: Proc. Interspeech. Jeju Island, Korea, pp. 565–568.

Alsteris, L.D., Paliwal, K.K., 2006. Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Commun. 48 (6), 727–736.

Aneeja, G., Yegnanarayana, B., 2015. Single frequency filtering approach for discriminating speech and nonspeech. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4), 705–717.

Bastys, A., Kisel, A., Salna, B., 2010. The use of group delay features of linear prediction model for speaker recognition.. Informatica 21 (1), 1–12.

Boashash, B., 1992. Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals. Proc. IEEE 80 (4), 520–538.

Bozkurt, B., Couvreur, L., Dutoit, T., 2007. Chirp group delay analysis of speech signals. Speech Commun. 49 (3), 159–176.

Cohen, L., 1995. Time-Frequency Analysis: Theory and Applications. Prentice-Hall Signal Processing Series, New York.

Costas, J.P., 1981. Residual Signal Analysis – A search and destroy approach to spectral estimation. In: Proc. First ASSP Workshop on Spectral Estimation. IEEE, Hamilton, Canada, pp. 6.5.1–6.5.8.

D. Alessandro, C., Sturmel, N., 2011. Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. Sadhana 36 (5), 601–622.

Deepak, K.T., Prasanna, S.R.M., 2016. Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (7), 1205–1219.

Drugman, T., Alku, P., Alwan, A., Yegnanarayana, B., 2014. Glottal source processing: From analysis to applications. Comput. Speech Lang. 28 (5), 1117–1138.

Drugman, T., Alwan, A., 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In: Proc. Interspeech. Florence, Italy, pp. 1973–1976.

Drugman, T., Dutoit, T., 2009. Glottal closure and opening instant detection from speech signals. In: Proc. Interspeech 2009. pp. 2891–2894.

Drugman, T., Dutoit, T., 2012. The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. Audio Speech Lang. Process. 20 (3), 968–981.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T., 2012. Detection of glottal closure instants from speech signals: A quantitative review. IEEE Trans. Audio Speech Lang. Process. 20 (3), 994–1006.

Fant, G., 1995. The LF-model revisited. transformations and frequency domain analysis. Speech Transm. Lab. Q. Prog. Status Rep. 36, 119–156.

Gangamohan, P., Kadiri, S.R., Yegnanarayana, B., 2013. Analysis of emotional speech at subsegmental level. In: Proc. Interspeech. pp. 1916–1920.

Gerkmann, T., Krawczyk-Becker, M., Le Roux, J., 2015. Phase processing for single-channel speech enhancement: History and recent advances. IEEE Signal Process. Mag. 32 (2), 55–66.

Gowda, D., Kadiri, S.R., Story, B., Alku, P., 2020. Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1901–1914.

ITU-T, Recommendation, 2005. G.191, Software tools for speech and audio coding standardization. Source: http://www.itu.int/rec/T-REC-G.191-200509-I/en.

Joseph, M.A., Guruprasad, S., Yegnanarayana, B., 2006. Extracting formants from short segments using group delay functions. In: Proc. Interspeech. pp. 1009–1012.

Kadiri, S.R., 2018. Analysis of Excitation Information in Expressive Speech (Ph.D. thesis). Speech Processing Laboratory, IIIT Hyderabad.

Kadiri, S.R., 2019. A quantitative comparison of epoch extraction algorithms for telephone speech. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. cICASSP, Brighton, UK, pp. 6500–6504.

Kadiri, S.R., Alku, P., 2020. Analysis and detection of pathological voice using glottal source features. IEEE J. Sel. Top. Signal Process. 14 (2), 367–379.

Kadiri, S.R., Alku, P., 2020. Excitation features of speech for speaker-specific emotion detection. IEEE Access 8, 60382–60391.

Kadiri, S.R., Alku, P., 2021. Glottal features for classification of phonation type from speech and neck surface accelerometer signals. Comput. Speech Lang. 70, 101232.

Kadiri, S.R., Alku, P., Yegnanarayana, B., 2020a. Comparison of glottal closure instants detection algorithms for emotional speech. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. cICASSP, pp. 7379–7383.

Kadiri, S.R., Alku, P., Yegnanarayana, B., 2020b. Analysis and classification of phonation types in speech and singing voice. Speech Commun. 118, 33–47.

Kadiri, S.R., Alku, P., Yegnanarayana, B., 2021. Extraction and utilization of excitation information of speech: A review. Proc. IEEE 109 (12), 1920–1941.

Kadiri, S.R., Gangamohan, P., Gangashetty, S.V., Alku, P., Yegnanarayana, B., 2020c. Excitation features of speech for emotion recognition using neutral speech as reference. Circuits Systems Signal Process. 39 (9), 4459–4481.

Kadiri, S.R., Gangamohan, P., Gangashetty, S.V., Yegnanarayana, B., 2015. Analysis of excitation source features of speech for emotion recognition. In: Proc. Interspeech. pp. 1324–1328.

Kadiri, S.R., Yegnanarayana, B., 2017. Epoch extraction from emotional speech using single frequency filtering approach. Speech Commun. 86, 52–63.

Kadiri, S.R., Yegnanarayana, B., 2018. Estimation of fundamental frequency from singing voice using harmonics of impulse-like excitation source. In: Proc. Interspeech. pp. 2319–2323.

Kadiri, S.R., Yegnanarayana, B., 2020. Determination of glottal closure instants from clean and telephone quality speech signals using single frequency filtering. Comput. Speech Lang. 101097.

Kawahara, H., Agiomyrgiannakis, Y., Zen, H., 2016. Using instantaneous frequency and aperiodi detection to estimate F0 for high-quality speech synthesis. In: Proc. ISCA Speech Synthesis Workshop. Sunnyvale, California, USA, pp. 221–228.

Khanagha, V., Daoudi, K., Yahia, H.M., 2014. Detection of glottal closure instants based on the microcanonical multiscale formalism. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1941–1950.

Kominek, J., Black, A., 2004. The CMU Arctic speech databases. In: Proc. ISCA Speech Synthesis Workshop. Pittsburgh, PA, USA, pp. 223–224.

Krawczyk, M., Gerkmann, T., 2014. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1931–1940.

Kumaresan, R., Ramalingam, C.S., Rao, A., 1994. RISC: an improved Costas estimator-predictor filter bank for decomposing multicomponent signals. In: Proc. IEEE Seventh SP Workshop on Statistical Signal and Array Processing. Quebec City, QC, Canada, pp. 207–210.

Kumaresan, R., Rao, A., 1999. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. J. Acoust. Soc. Am. 105 (3), 1912–1924.

Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervocalic stop consonants. Speech Commun. 22 (4), 403–417.

Mathes, R.C., Miller, R.L., 1947. Phase effects in monaural perception. J. Acoust. Soc. Am. 19 (5), 780–797.
McCowan, I., Dean, D., McLaren, M., Vogt, R., Sridharan, S., 2011. The delta-phase spectrum with application to voice activity detection and speaker recognition. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2026–2038.
Mowlaee, P., Saeidi, R., 2013. Iterative closed-loop phase-aware single-channel speech enhancement. IEEE Signal Process. Lett. 20 (12), 1235–1239.
Mowlaee, P., Saeidi, R., Stylianou, Y., 2016. Advances in phase-aware signal processing in speech communication. Speech Commun. 81, 1–29.
Murthy, H.A., Yegnanarayana, B., 1991. Formant extraction from group delay function. Speech Commun. 10 (3), 209–221.
Murthy, H.A., Yegnanarayana, B., 2011. Group delay functions and its applications in speech technology. Sadhana 36 (5), 745–782.
Murthy, B.N., Yegnanarayana, B., Kadiri, S.R., 2020. Time delay estimation from mixed multispeaker speech signals using single frequency filtering. Circuits Systems Signal Process. 39 (4), 1988–2005.
Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1602–1613.
Murty, K.S.R., Yegnanarayana, B., Joseph, M.A., 2009. Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. 16 (6), 469–472.
Nakagawa, S., Wang, L., Ohtsuka, S., 2012. Speaker identification and verification by combining MFCC and phase information. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1085–1095.
Naylor, P.A., Kounoudes, A., Gudnason, J., Brookes, M., 2007. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio, Speech Lang. Process. 15 (1), 34–43.
Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69 (5), 529–541.
Paliwal, K.K., Alsteris, L.D., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Commun. 45 (2), 153–170.
Paliwal, K.K., Wójcicki, K.K., 2008. Effect of analysis window duration on speech intelligibility. IEEE Signal Process. Lett. 15, 785–788.
Paliwal, K.K., Wójcicki, K.K., Shannon, B.J., 2011. The importance of phase in speech enhancement. Speech Commun. 53 (4), 465–494.
Quatieri, T.F., 2004. Discrete-Time Speech Signal Processing. Pearson Education, Singapore.
Quatieri, T.F., Oppenheim, A.V., 1981. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. IEEE Trans. Acoust. Speech Signal Process. 29 (6), 1187–1193.
Rabiner, L., Schafer, R., 2010. Theory and Applications of Digital Speech Processing. Prentice Hall, USA.
Rajan, P., Kinnunen, T., Hanilci, C., Pohjalainen, J., Alku, P., 2013. Using group delay functions from all-pole models for speaker recognition. In: Proc. Interspeech. Lyon, France, pp. 2489–2493.
Ramalingam, C.S., Rao, A., Kumaresan, R., 1994. Time-frequency analysis using the residual interference signal canceller filter bank. In: Proc. IEEE-SP Int. Symposium on Time-Frequency and Time-Scale Analysis. Philadelphia, USA, pp. 500–503.
Rao, K.S., Yegnanarayana, B., 2006. Prosody modification using instants of significant excitation. IEEE Signal Process. Lett. 14 (3), 972–980.
Saratxaga, I., Sanchez, J., Wu, Z., Hernaez, I., Navas, E., 2016. Synthetic speech detection using phase information. Speech Commun. 81, 30–41.
Schluter, R., Ney, H., Using phase spectrum information for improved speech recognition performance. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 1. ICASSP, Salt Lake City, UT, USA, pp. 133–136.
Schroeder, M.R., 1975. Models of hearing. Proc. IEEE 63 (9), 1332–1350.
Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. 3 (5), 325–333.
Stark, A.P., Paliwal, K.K., 2008. Speech analysis using instantaneous frequency deviation. In: Proc. Interspeech. Brisbane, Australia, pp. 2602–2605.
Swamy, R.K., Murty, K.S.R., Yegnanarayana, B., 2007. Determining number of speakers from multispeaker speech signals using excitation source information. IEEE Signal Process. Lett. 14 (7), 481–484.
Thomas, M.R.P., Gudnason, J., Naylor, P.A., 2012. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans. Audio Speech Lang. Process. 20 (1), 82–91.
Tsiakoulis, P., Potamianos, A., Dimitriadis, D., 2013. Instantaneous frequency and bandwidth estimation using filterbank arrays. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. cICASSP, Vancouver, BC, Canada, pp. 8032–8036.
Vijayan, K., Murty, K.S.R., 2015. Analysis of phase spectrum of speech signals using allpass modeling. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (12), 2371–2383.
Vijayan, K., Murty, K.S.R., 2016. Epoch extraction by phase modelling of speech signals. Circuits Systems Signal Process. 35 (7), 2584–2609.
Vijayan, K., Murty, K.S.R., Li, H., 2019. Allpass modeling of phase spectrum of speech signals for formant tracking. In: Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC. Lanzhou, China, pp. 1190–1196.
Vijayan, K., Reddy, P.R., Murty, K.S.R., 2016. Significance of analytic phase of speech signals in speaker verification. Speech Commun. 81, 54–71.
Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. IEEE Trans. Acoust. Speech Signal Process. 30 (4), 679–681.
Wang, L., Minami, K., Yamamoto, K., Nakagawa, S., 2010. Speaker identification by combining MFCC and phase information in noisy environments. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. cICASSP, Dallas, Texas, USA, pp. 4502–4505.
Yegnanarayana, B., 1978. Formant extraction from linear-prediction phase spectra. J. Acoust. Soc. Am. 63 (5), 1638–1640.
Yegnanarayana, B., Gangashetty, S.V., 2011. Epoch-based analysis of speech signals. Sadhana 36 (5), 651–697.
Yegnanarayana, B., Murthy, H.A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Process. 40 (9), 2281–2289.
Yegnanarayana, B., Murty, K.S.R., 2009. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Process. 17 (4), 614–624.
Yegnanarayana, B., Prasanna, S.R.M., Duraiswamy, R., Zotkin, D., 2005. Processing of reverberant speech for time-delay estimation. IEEE Trans. Speech Audio Process. 13 (6), 1110–1118.
Yegnanarayana, B., Saikia, D., Krishnan, T., 1984. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans. Acoust. Speech Signal Process. 32 (3), 610–623.
Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 1. ICASSP, Montreal, Quebec, Canada, pp. 125–128.