Journal of Computer Science

Original Research Paper

Speech Signal Analysis in Phase Domain

Husne Ara Chowdhury and Mohammad Shahidur Rahman

Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh

Article history
Received: 21-04-2020
Revised: 22-07-2020
Accepted: 06-08-2020

Corresponding Authors:
Husne Ara Chowdhury
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
Email: husna-cse@sust.edu

Abstract: Speech signal analysis based on the Phase Spectrum (PS) in the recent years becomes more popular to the researchers due to its attractive properties. Though some problem arises during its processing but appropriate modification gives the useful result. In this study, we proposed a new method of speech signal analysis based on its phase spectral representation. After analyzing the PS we modified it vocal tract dominated spectrum by utilizing the minimum-phase stable component of the speech spectrum. The modified signal holding only the phase spectral component is used for the parametric modeling of the speech signal to estimate the characteristics of vocal tract perfectly. This extracted feature was found to provide complementary evidence from the magnitude spectrum of speech but with better resolution. According to perceptual analysis, the PS takes precedence over the magnitude spectrum. This research utilizes the Group Delay (GD) spectrum as a representative of the PS because of its meaningful characteristic. All pole modeling is performed on the signal evaluated from the GD spectrum to find the resonances of the vocal tract system accurately. The effectiveness of the method is tested by synthesizing some vowels over a range of pitch periods from low to high pitched speech. The validity of the proposed method is also verified by plotting the formant contour on the spectrogram of a sentence from the TIMIT database and standard F2-F1 plot of natural speech spoken by male and female speakers. The proposed method performs better than the state-of-the-art methods.

Keywords: Group Delay, Phase Spectrum, Stability, Deconvolution, Spikiness

Introduction

Source-filter separation is an interesting problem in speech signal analysis. Various methods have been exercised to separate filter from source for long years. The magnitude spectrum of speech signal is usually used to decompose it into its fundamental components such as the excitation source and the vocal tract filter. Most of the popular decomposition methods can be either model-based (linear prediction), or non-model based (Cepstrum).

Atal and Hanauer (1971; Makhoul, 1975) used the Linear Predictive Coding (LPC) method to separate the source from the filter. But this technique has some limitations. One of the major limitations of LPC is that the peaks of linear prediction spectrum are highly influenced by the frequency of pitch harmonics. In high pitched speaking, such estimation is very difficult due to the wide spacing of harmonics. Magi *et al*. (2009) tried to compensate for the problem using the stabilized weighted LP. The model-based approach uses a specific order of prediction that coincides with the peaks of the magnitude spectrum envelope rather than the actual number of peaks that may resemble the resonances. Therefore they are not reliable to observe fluctuations that occur in short intervals. So the speech processing methods described in (Noll, 1967; Oppenheim *et al*., 2001; Deller *et al*., 2000; Rahman and Shimamura, 2005; 2007) used the cepstrum method for this purpose in some way. Moreover, the source-filter separation of a segment of speech signal employing the magnitude spectrum is affected by the poor resolution and higher spectral leakage.

All of these approaches use the magnitude spectrum while ignoring the PS completely. This is mainly because the PS is difficult to analyze for discrete-time

signals. Because it is associated with the problem of chaotic nature due to phase wrapping (Loweimi, 2018). Both of these spectrums are a significant part of original signal reconstruction. Various attempts (Murthy and Yegnanarayana, 1991; Duncan *et al*., 1989; Bozkurt *et al*., 2004a; 2004b) have been made to exhibit the significance of the PS in speech analysis for source-filter separation. These methods usually encompass the use of group delay spectrum. The first-order negative derivative of unwrapped PS is called Group Delay (GD) spectrum.

Even though actually the convolution of the time domain signal is same as the addition in the phase domain (Loweimi *et al*., 2015; Vijayan and Murty, 2015) signal, which may be useful for carrying out deconvolution, however, there is no direct phase modeling algorithm available for such decomposition. Thus developing a fundamental phase based model of source-filter in a trustworthy way elevates the PS to be used in speech analysis practically. To compensate the problem of phase wrapping, the GD function of the PS can be used as a representative of phase. GD spectrum conveys the information of PS but its overall contour is more understandable. The GD spectrum has some useful properties described by (Murthy and Yegnanarayana, 2011).

Gowda *et al*. (2013) employed GD function on the LP spectrum for formant estimation. This type of method conveys the error encountered in the magnitude spectrum and there are no extra facilities encountered using the GD spectrum. To extract the features accurately from the GD spectrum, it should be taken from the pure phase spectrum properly.

GD spectrum often shows spurious spikes which creates masks on the formant peaks. Bozkurt (2005) that, the persistence of zeros near the unit circle is responsible for these spikes. He tried to remove these spikes utilizing the analysis window other than the unit circle (Bozkurt *et al*., 2007).

Yegnanarayana *et al.* (1988) have shown a few methods (Murthy *et al*., 1989; Murthy and Yegnanarayana, 1991; Murthy and Gadde, 2003) to overcome the problem of spikiness and derived formant tracking by compensating the spikes from group delay function. They obtained a smoothed magnitude spectrum through cepstral smoothing. Then computed smoothed group delay spectrum of minimum phase signal from this smoothed spectral representation. But minimum-phase GD spectrum contains some roots that coincide with the unit circle. According to the properties of the GD spectrum described in (Murthy and Yegnanarayana, 2011) if a root is coincided with the unit circle, the GD spectrum at these location of roots becomes infinity. So the vocal tract filter extracted from this minimum-phase GD spectrum becomes marginally stable in that case. As a result, spikes retain yet. A stable filter (Deepak and Prasanna, 2015) with roots inside the unit circle overcomes this problem.

In this analysis, utilizing the stable and minimum-phase component of the speech spectrum the GD spectrum is computed that shows better resolution pitch and its harmonics. The possible pitch harmonics are more suppressed with emphasized vocal tract dominated harmonics. This GD spectrum becomes an useful tool to deconvolve the vocal tract filter from the excitation source. Accuracy of vocal tract filter estimation depends on the accurate estimation of AR (auto-regressive) coefficients. The proposed high resolution GD spectrum can mitigate this problem. The Collaborative Voice Analysis Repository (COVAREP) (Degottex *et al*., 2014) package used the pure phase domain based method called the Differential-Phase Peak Tracking (DPPT) method (Bozkurt *et al*., 2004a) for formant tracking. This method is the recent popular method for formant tracking. In a few case the DPPT method suffers from glottal formant (Bozkurt *et al*., 2004c) effect.

The efficacy of the proposed method is tested by estimating the formant frequencies. Then it is compared with the recent DPPT method implemented in COVAREP. The proposed method shows elevated accuracy in formant frequency estimation than DPPT, particularly in the case of high pitched speech.

Problem Identification

Fourier Analysis and Magnitude Spectrum

For speech signal analysis Fourier transformation is an important and popular mathematical tool. The Discrete Fourier Transform (DFT) (Rabiner, 1978) of the discrete signal $x(n)$ can be defined using the Equation (1):

$$X(k)=\frac{1}{L}\sum_{n=1}^{L}x(n)e^{-j\left(\frac{2\pi}{L}\right)kn} \quad (1)$$

where, n, L, k indicates sample number, frame length, frequency index with k = 1, 2, 3….N. Alternatively as follows:

$$X(k)=|X(k|e^{\phi X(k)} \quad (2)$$

where, $|X(k)|$ is the magnitude spectrum and $\phi X(k)$ is the PS, respectively which are also represented as:

$$|X(k)| = \sqrt{X_R(k)^2 + X_I(k)^2}$$
$$\phi X(k) = \arctan \frac{X_I(k)}{X_R(k)} \quad (3)$$

where, $X_R(k)$ and $X_I(k)$ indicate the real and imaginary part.

The Fig. 1-7 is generated from a synthetic vowel /a/ utilizing the Matlab program. A speech frame and corresponding magnitude spectrum, PS and GD spectrum of a synthetic vowel /a/ is shown in Fig. 1a to 1d. The magnitude spectrum has a fine structure of its harmony that better characterizes the speech signal and matches our understanding level. The local minimum and maximum in magnitude spectrum correspond to zeros and poles in *z* domain which is a useful tool to deconvolve the source from the filter. The understandable representation of the magnitude spectrum makes it useful broadly.

The main problem encountered with the magnitude spectrum is a higher spectral leakage and a lower spectral resolution. The short duration of real measurements degrades the overall resolution of the magnitude spectrum. The magnitude spectrum is formed by the multiplication of the magnitude of the individual component spectra. This is another cause of overall resolution reduction as shown in Fig. 2. The lack of integer multiple cycles of each component spectra causes spectral leakage.

Phase Spectrum

Speech signal holding only the phase spectral component is formed by appointing one in place of the magnitude spectral part. If the value of $x(n)$ is real, its PS is an odd function and the PSs values are limited within $\pm\pi$ (wrapped). As a result, it shows a chaotic nature that poses no physical interpretation or mathematical

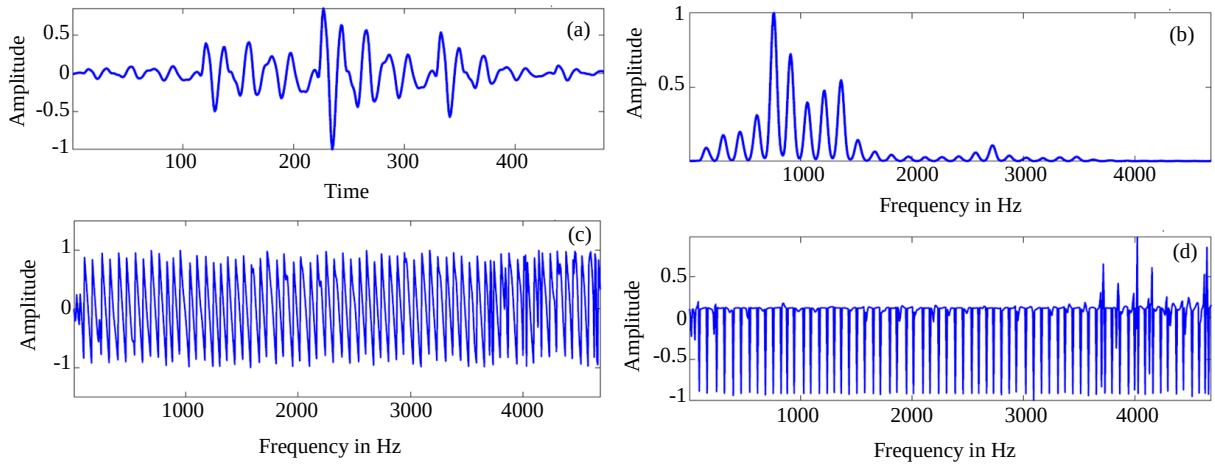

Fig. 1: (a) A speech segment (b) Magnitude Spectrum (c) Phase Spectrum (d) Group Delay Spectrum of a synthetic vowel/a/

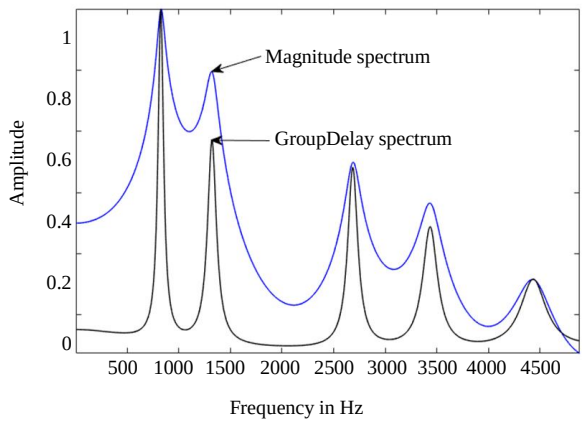

Fig. 2: A speech signal using 12 poles LPC based Magnitude Spectrum and Corresponding GD Spectrum

modeling. A linear phase component of $e^{-j\left(\frac{2\pi}{L}\right)kn}$ is introduced in the PS by shifting the signal $x(n)$ by k samples in the time domain. But the encoded information in the PS become hidden. The main difficulty in unwrapping the PS as shown in Fig. 1c is the summation of integer multiple cycle of 2π to the principle PS. The PS can be unwrapped by taking the derivative of the PS using the following expression:

$$\arg'(X(k)) = \frac{d\{\arg(X(k))\}}{dk}$$
$$= \frac{X_R(k)X_I'(k) - X_R'(k)X_I(k)}{|X(k)|^2}$$

where $'$ denotes derivative w.r.t. k.

The impulse response of a signal $x(n)$ is the cascade of resonators and anti resonators. But the PS becomes the addition of individual component resonators phase spectra. The GD function shows the variations of resonators and anti resonators clearly. To extract useful features from PS researchers (Yegnanarayana *et al*., 1984) like to work with its derivative which is called the GD spectrum.

Group Delay Spectrum

Dealing with the issues in PS, researchers become interested with GD spectrum which conveys all of the PS information with more accountable form. It can be expressed as the negative spectral derivative of the unwrapped PS using the Equation (4):

$$\tau(k) = -\mathrm{Im}\frac{d\{\log(X(k))\}}{dk} = -\frac{d\{\arg(X(k))\}}{dk} \quad (4)$$

where, $Im\{.\}$ and $arg\{.\}$ indicate the imaginary part and unwrapped phase. Since the phase unwrapping is not simple (Baldi *et al*., 2000), the GD spectrum can be computed by forgoing this problem as follows:

$$\tau(k) = \frac{X_R(k)X_I'(k) - X_R'(k)X_I(k)}{|X(k)|^2} \quad (5)$$

The resonance and anti-resonance peaks in the group delay domain have higher resolution comparatively with the spectral domain. As shown in Fig. 2, the GD spectrum has a noticeably better frequency resolution and lower frequency leakage where the resonance information is more clear. If one or more poles or zeros i.e., roots are on the unit circle, the GD becomes $\tau(k) = \infty$ at the location of that roots (Murthy and Yegnanarayana, 2011).

Properties of Group Delay Spectrum

The peaks/valleys of the GD spectrum correspond to poles/zeros of the transfer function. The time domain convolution operation of a speech becomes addition in the group delay.

Suppose a system $h(n)$ in the time domain results from the convolution of two-component resonators $h_1(n)$ and $h_2(n)$. The frequency response of the system is given by:

$$\begin{aligned} F\{h(n)\} &= F\{h_1(n) * h_2(n)\} \\ H(k) &= H_1(k)H_2(k) \\ &= |H_1(k)||H_2(k)|e^{j\left(\arg\{H_1(k)\}+\arg\{H_2(k)\}\right)} \\ \arg\{H(k)\} &= \arg\{H_1(k)\} + \arg\{H_2(k)\} \\ \Rightarrow \tau_h(k) &= \tau_{h_1}(k) + \tau_{h_2}(k) \end{aligned}$$

where, F denotes DFT, X and H denotes DFTs of signal and impulse response and τ indicates GD spectrum.

Issues of Group Delay Spectrum

Practically GD spectrum encounters the following important issues:

i) It suffers from spikiness as shown earlier in Fig. 1d which restricts its applicability. Due to its spiky nature, neither the pitch nor the formants can be distinguished visually. As a result, it becomes confusing to the researcher that, whether it carries any important information or not. The zeros (poles) close to the unit circle are manifested as spikes in the GD spectrum as shown in Fig. 3. The excitation component gives rise to such zeros. Bozkurt (2005) discussed the cause of this issue elaborately. The strength of these spikes rely on the proximity of the roots with the unit circle. These spikes are scattered in a significant part of the spectrum and cannot be eliminated easily.

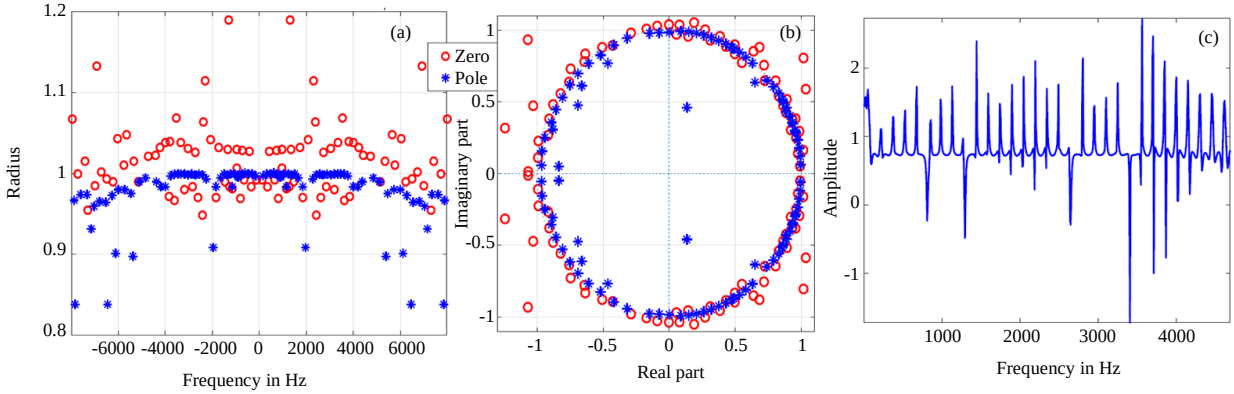

Fig. 3: A non minimum phase speech signal $x(n)$ representation with (a) Pole-Zero plot in Cartesian coordinate. (b) Pole-Zero plot in polar coordinate. (c) Corresponding Group Delay Spectrum.

ii) Most of the GD spectrum-based analysis methods use a minimum phase signal. But the system characterized by the minimum phase signal shows some roots on the unit circle with radius = 1. GD spectrum extracted from the minimum phase signal shows spurious spikes as shown in Fig. 4.

iii) Murthy and Yegnanarayana (2011) have shown in that for the minimum phase and non-minimum phase signals magnitude spectrum is the same but the GD spectrum is different based on the pole-zero analysis. So the feature extracted from the GD is not applicable for all the cases.

iv) We know that the GD spectrum is additive. In fact, this additive property holds not for vocal tract filter only but also for excitation source. As a result, both of them overlap in the GD spectrum domain. So the problem encountered yet using the filtering method to decompose the filter from the source.

v) The GD spectrum sometimes becomes negative which violates the causality of the speech signal system (Loweimi, 2018).

In this study, we exploited the benefits of the GD spectrum by solving the aforementioned problems.

Proposed Solutions of Group Delay Spectrum

Several techniques have been derived in last few years concerning the GD spikiness, such as Modified GDF (MODGDF) (Murthy and Yegnanarayana, 1991; Murthy and Gadde, 2003), Chirp GD Function (CGDF) (Bozkurt *et al*., 2007), product spectrum (Zhu and Paliwal, 2004) and model based GDF (Yegnanarayana, 1978; Loweimi *et al*., 2013). The MODGDF replaces the denominator in Equation (5) by the cepstraly smoothed power spectrum when $|X(k)|$ approaches to zero. In the CGDF Bozkurt tried to smooth the GD spectrum by moving the analysis window away from the unit circle. Then used the peak picking algorithm to find the formants. In the product spectrum method, Zhu and Paliwal replaced the denominator by unity. The model-based approach uses the Autoregressive (AR) model extracted from the signal and then its group delay is computed. These techniques deal with the problems of the GD spectrum and achieved more success about the spikiness problems. But none of them provides how the source and filter components interact or overlap with each other and how is such information encoded in the PS.

The proper solution of the issues can be acquired by utilizing the signals of the causal output of a stable system. The minimum phase signal ensures the causal output of a system due to its poles-zeros lie within the unit circle. Some of the roots can also stay on the unit circle which causes spikes yet scattered over the GD spectrum as shown in Fig. 4. According to the method described in (Deepak and Prasanna, 2015) placing the roots on the unit circle results in a marginally stable filter. So necessity arises to make the system stable. To do so one of the ways is to place all of the poles and zeros within (less than one) the unit circle (Deepak and Prasanna, 2015). On the way of doing so, the excitation component with its harmonics becomes suppressed and the harmonics that may correspond with the filter component become boosted. If the roots of the speech signal are reflected away from the unit circle the GD spectrum shows signal information more clearly than the magnitude spectrum. For speech signals, the vocal tract associated poles are located in a reliable distance from the unit circle (Loweimi, 2018). So by all-pole modeling the signal from the appropriately processed GD spectrum, resonant peaks can be retrieved with elevated accuracy.

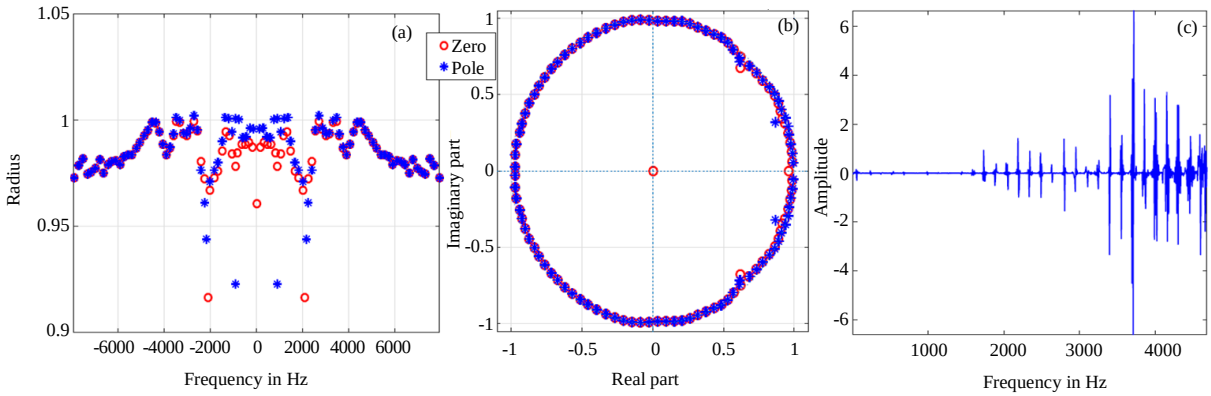

Fig. 4: A minimum phase speech signal $x_m(n)$ representation with (a) Pole-Zero plot in Cartesian coordinate. (b) Pole-Zero plot in polar coordinate. (c) Corresponding Group Delay Spectrum.

Proposed Group Delay Spectrum

Fourier transform $X(k)$ of a discrete signal $x(n)$ can be decomposed into two component, namely, Minimum Phase (MP) X_{MP} and All Pass (AP) X_{AP} using the following expression:

$$\begin{aligned} X(k) &= |X(k)| e^{-j.\arg\{X(k)\}} \\ &= X_{MP}.X_{AP} \\ \arg\{X(k)\} &= \arg\{X_{MP}(k)\} + \arg\{X_{AP}(k)\} \end{aligned} \tag{6}$$

From the Equation (6) we can say that the magnitude spectrum is only related to MP part but the PS is related to both of the MP and AP component of the speech spectrum. It is noticeable that minimum phase signal can be retrieved from the magnitude spectrum easily. The source (src) and filter (flt) are manifested in the magnitude spectrum and GD spectrum, which can be expressed by the Equation (7) and (8):

$$|X(k)| = X_{MP}(k) = |X_{flt}(k)|.|X_{src}(k)| \tag{7}$$

$$\begin{aligned} \arg\{X_{MP}(k)\} &= \arg\{X_{flt}(k)\} + \arg\{X_{src}(k)\} \\ \tau_{MP}(k) &= \tau_{flt}(k) + \tau_{src}(k) \end{aligned} \tag{8}$$

The pole-zero plot and GD spectrum of the signal $x(n)$ is shown in Fig. 3. Since the signal $x(n)$ is a non-minimum phase signal, it is observed that all of the poles/zeros are scattered inside and outside the unit circle. As result the filter becomes unstable. So the resulted GD spectrum becomes spiky.

Our goal is to separate the filter from the source to estimate the vocal tract characteristics by modeling the signal retrieved from the phase domain only. To do so, first, we need to compute the minimum-phase component of the $X(k)$, named as $X_{MP}(k)$.

Suppose the z transform of a signal is:

$$X(z) = \frac{b_0 \prod_{i=1}^{m}(1 - b_i z^{-1})}{a_0 \prod_{i=1}^{n}(1 - a_i z^{-1})}$$

where, $\forall_i$, $b_i < 1$ and $a_i < 1$. If all roots of this signal can be reflected within the unit circle then it is called the minimum phase signal. In some cases, a number of roots may arise on the unit circle which are also responsible for GD spikes. Let $x_m(n)$ be the minimum phase signal retrieved using the Equation (9):

$$x_m(n) = \frac{1}{N} \sum_{n=1}^{N} X_{MP}(k) e^{j\left(\frac{2\pi}{N}\right)kn} \tag{9}$$

According to the pole-zero plot of this minimum phase signal shown in Fig. 4 it is evident that a number of roots lie on the unit circle and the GD spectrum remains spiky yet causing the system marginally stable. To make the system converging type it should become stable (Deepak and Prasanna, 2015). Deepak and Prasanna ensured the system stability by reducing the radius of roots (poles/zeros) less than one. So necessity arises to place all of the poles and zeros within the unit circle. For this purpose, we have applied the following formula on the minimum phase signal $x(n)$:

$$x_s(n) = x_m(n) * (\alpha)^n \tag{10}$$

where, $0 < \alpha < 1$.

The pole-zero plot and corresponding GD spectrum is shown in Fig. 5 of the resulted signal $x_s(n)$. It is found that the system characterized by the signal $x_s(n)$ becomes stable. This is because the signal $x_s(n)$ converges with the signal $x_m(n)$ with suppressed excitation source as in Fig. 6a and 6b. The GD spectrum computed from the signal $x_s(n)$ shown in Fig. 6d. In this GD spectrum, the possible harmonics corresponding to resonance peaks are manifested and the others are more suppressed.

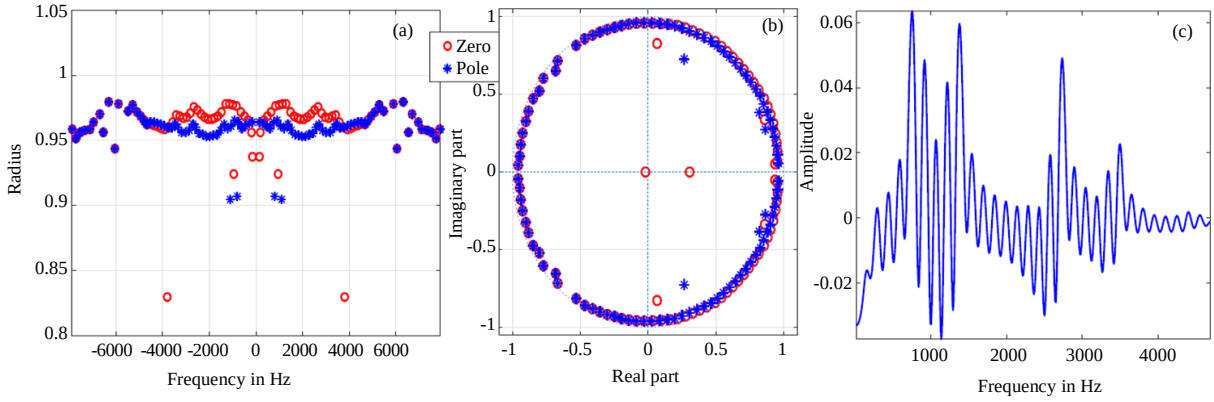

Fig. 5: A stable and minimum phase speech signal $x_s(n)$ representation with (a) Pole-Zero plot in Cartesian coordinate. (b) Pole-Zero plot in polar coordinate. (c) Corresponding Group Delay Spectrum.

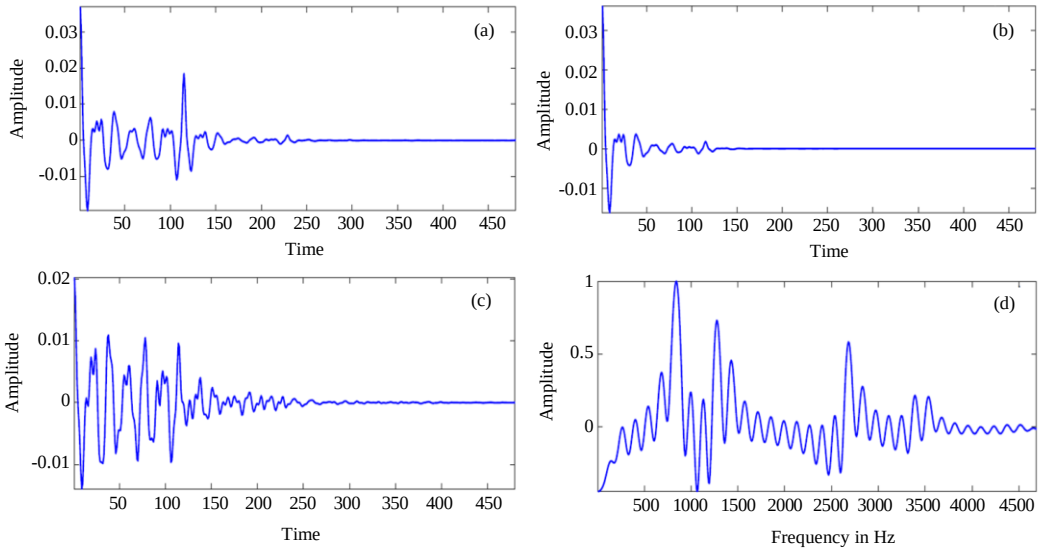

Fig. 6: (a) Minimum Phase Signal $x_m(n)$. (b) Signal $x_s(n)$ by placing poles and zeros inside the unit circle. (c) Signal $x_{ss}(n)$ retrieved from the GD Spectrum. (d) GD Spectrum of the signal $x_s(n)$

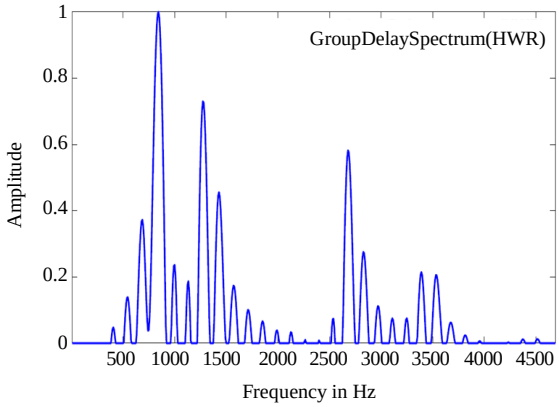

Fig. 7: GD spectrum after half wave rectification

Loweimi *et al.* (2015) used a low pass filter to extract the excitation component and used a simple subtraction method to retrieve the filter component with some post processing. But this method faces the issue mentioned in section 2.3.2 (iv). Now the N point DFT of $x_s(n)$ is $X_s(k)$ obtained by using the Equation (1). GD spectrum of that signal can be obtained by the Equation (11):

$$\tau_s(k) = -\frac{d\left\{\arg\left(X_s(k)\right)\right\}}{dk}$$

$$\tau'_s(k) = sign\left(\tau_s(k)\right).\left(\tau_s(k)\right)^{\beta} \tag{11}$$

where, $3 > \beta > 1$. Since the negative values in GD spectrum discontinues the causality of the characteristic system (Loweimi, 2018) so in order to make the system causal the GD spectrum has to be half-wave rectified by using the following expression:

$$\tau_{ss}(k) = \begin{cases} \tau'_s(k), if\ \tau'_s(k) > 0 \\ 0,\ otherwise \end{cases} \tag{12}$$

The GD spectrum extracted from the Equation (12) can be expressed as follows:

$$\tau_{ss}(k) = \tau_{flt}(k) + \tau_{src}(k) \tag{13}$$

The filter component is more emphasized with the suppressed source component in the GD spectrum obtained from the Equation (13) as shown in Fig. 7. The proposed method is briefly shown by the block diagram in the Fig. 8.

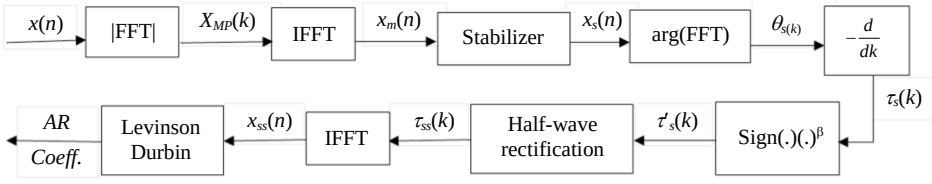

Fig. 8: Block diagram of the proposed method.

Decomposition of Proposed GD Spectrum

By applying the inverse Fourier transform using Equation (14) on that GD spectrum we find the signal denoted by $x_{ss}(n)$ which carries both of source and filter information:

$$x_{ss}(n)=\frac{1}{N}\sum_{n=1}^{N}\tau_{ss}(k)e^{j\left(\frac{2\pi}{N}\right)kn} \quad (14)$$

This signal is stable. So it converges with the input signal $x_m(n)$ by suppressing the excitation component with dominated filter component as shown in Fig. 6c. This signal $x_{ss}(n)$ has all features as $x(n)$ with additional facility named as causality and stability. By all-pole modeling, the signal $x_{ss}(n)$ can be deconvolved into the vocal tract filter and the excitation source successfully.

Synthetic Speech Analysis

The Liljancrant-Fant glottal model (Fant *et al*., 1985) is used to generate synthetic speech by simulating the source. The efficacy of the proposed method is tested using over a range of fundamental frequencies from 100 to 400 Hz by simulating five synthetic Japanese vowels according to (Rahman and Shimamura, 2007). Five formant frequencies for each vowel according to Table 1 is used to simulate the vowels. The fixed values of bandwidths for the five formants are set as 60, 100, 120, 175 and 281 Hz. The sampling frequency used here is 10 kHz. In this study, the formant frequencies at different fundamental frequencies denoted by F0s are estimated. For this purpose, all the parameters of the glottal model are kept constant. The analysis order is set to 12. A Gaussian window of 30 ms is used to segment the speech signal. Each segment is shifted by 10 ms. The DFT size 1024 point is used to analyze the magnitude and PS.

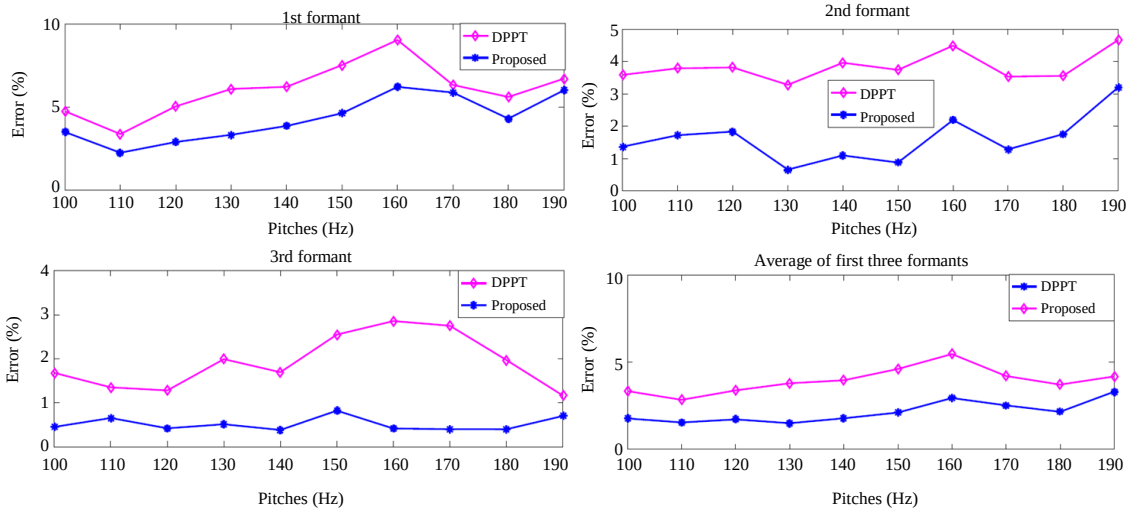

Fig. 9: REE of Formant Frequencies by synthesizing five vowels at different pitch values (100 to 190 Hz)

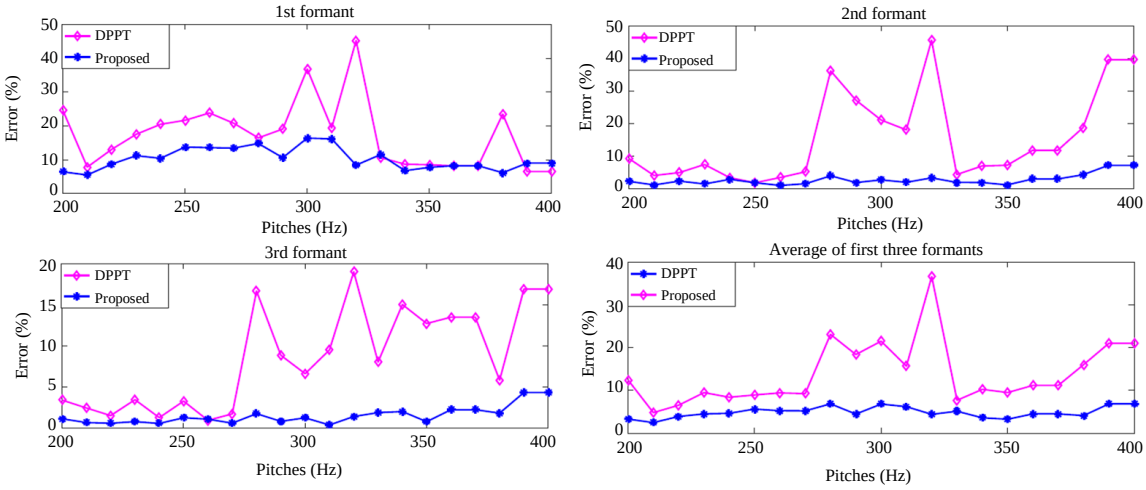

Fig. 10: REE of formant frequencies by synthesizing five vowels at different pitch values (200 to 400 Hz)

Formant Frequency Estimation

The formant frequencies estimated by analyzing the speech signal on different window positions. Each formant frequency values are taken as the arithmetic mean of all of the window positions for each vowel. The Relative Estimation Error (REE), REF_i, of the *ith* formant is calculated by averaging the individual F_i errors of all the five vowels. Now we can express REF_i as:

$$REF_i = \frac{1}{5}\sum_{j=1}^{5} | \hat{F}_{ij} - F_{ij} | / F_{ij} \quad (15)$$

where, F_{ij} denotes the *ith* formant frequency of the *jth* vowel and $\hat{F}_{ij}$ is the corresponding estimated value. The average of the REE, E of the first three formants of all the five vowels are represented using the following expression:

$$E = \frac{1}{15}\sum_{j=1}^{5}\sum_{i=1}^{3}\left|\hat{F}_{ij} - F_{ij}\right| / F_{ij} \quad (16)$$

Table 1: List of formant frequencies for synthesizing vowels

vowel	F1	F2	F3	F4	F5
/a/	813	1313	2688	3438	4438
/i/	375	2188	2938	3438	4438
/u/	375	1063	2188	3438	4438
/e/	438	1863	2688	3438	4438
/o/	438	1063	2688	3438	4438

The REE of the three formants F1, F2 and F3 and average REE of the first three formants are shown in Fig. 9 and 10. For low pitches, the REE shows that the proposed technique is better than the DPPT technique. For high pitches, it is observed that the DPPT method shows higher errors. This is because the DPPT method for high pitches is mostly affected by the glottal formant.

As a result, the first formant peak shifts towards the nearest glottal formant or harmonics. If the first formant shifts then the other formants such as F2, F3 also shifts. The DPPT method is based on the smoothing of the GD spectrum. This smoothing cannot be perfectly done in the high pitched speech. The proposed method on the other hand, is based on the all-pole modeling of the signal extracted from GD spectrum. So the proposed method results in more accurate formant estimation. It can be summarized from Fig. 9 and 10 that for analyzing the high pitched speech the proposed method can be utilized with elevated accuracy.

Analysis on Real Speech

The formants are estimated from a sentence of the TIMIT database which is sampled at 16 kHz. To segment the speech signal a Gaussian window size of 30 ms and a frame overlapping of 5 ms is used. The DFT size 1024 point is used to analyze the magnitude and PS. The speech signal is pre emphasized by $1\text{-}z^{-1}$ and prediction order 16 is used.

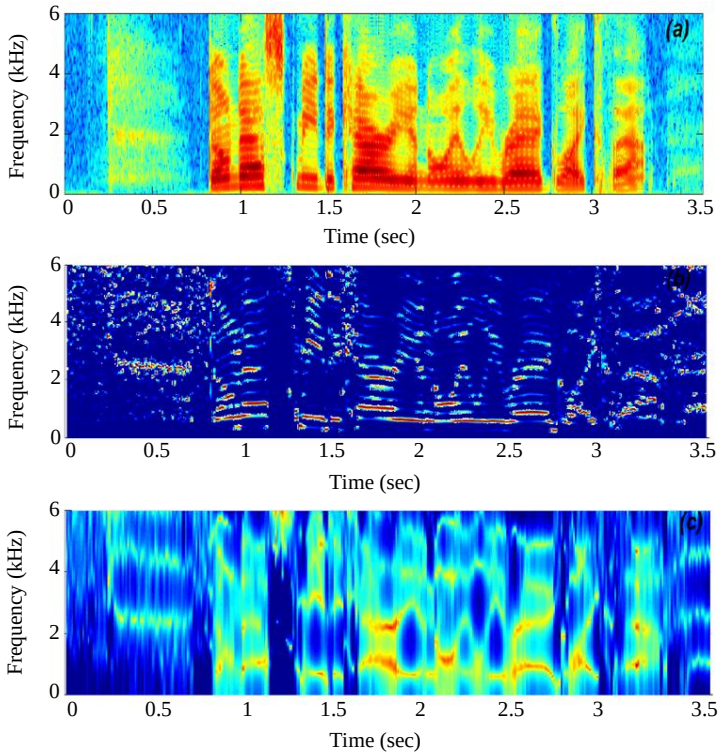

Fig. 11: A sentence from TIMIT database "Don't ask me to carry an oily rag like that". (a) Spectrogram of the sentence. (b) GroupDelayGram of the Proposed method (c) LPGram of the Proposed method.

DOI: 10.3844/jcssp.2020.1115.1127

Figure 11a shows the spectrogram of the input sentence. The GD spectrum is taken from this sentence and processed using the proposed method. The GDgram in Fig. 11b shows the harmonics that may correspond to the formant peaks with higher intensity and others are with blurring intensity. The signal from this GD spectrum is modeled by the all-pole modeling approach. The LPgram in Fig. 11c shows formants clearly but the pitch and its harmonics are absent.

To find the accuracy of the proposed method the formant peaks are plotted on the spectrogram as shown in Fig. 12. It is observed that the all the formant values are reasonably in the confined region than the DPPT method. However, some spurious peaks are also picked up in the unvoiced region.

We also analyzed the real speech signal of vowel sounds /a/ and /o/ uttered by a male and a female speaker. The result of analyzing these vowels are represented by standard F2-F1 plot and formant contour in case of low pitched male and high pitched female speeches which are shown in Figs. 13 to 16. For both of male and female speakers, the proposed method produced an F2-F1 value of almost all frames that exist within a confined region where the DPPT method produced some scattered values. In the DPPT method, some F2-F1 values on the zero line indicate that F2 values are treated as F1 values. This is because the DPPT method is influenced by glottal formant. From the F2-F1 plot of all the cases, it is evident that the proposed method is free from the glottal formant effect.

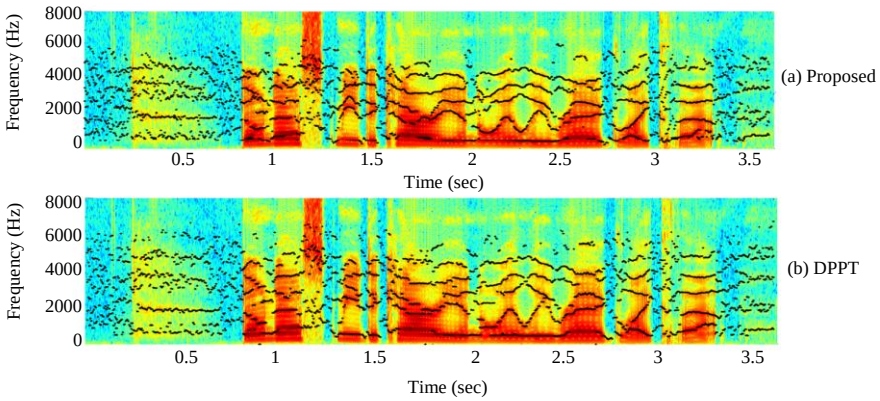

Fig. 12: A sentence from TIMIT database "Don't ask me to carry an oily rag like that". Spectrogram and Formant contour of F1-F5 (a) Proposed method (b) DPPT method.

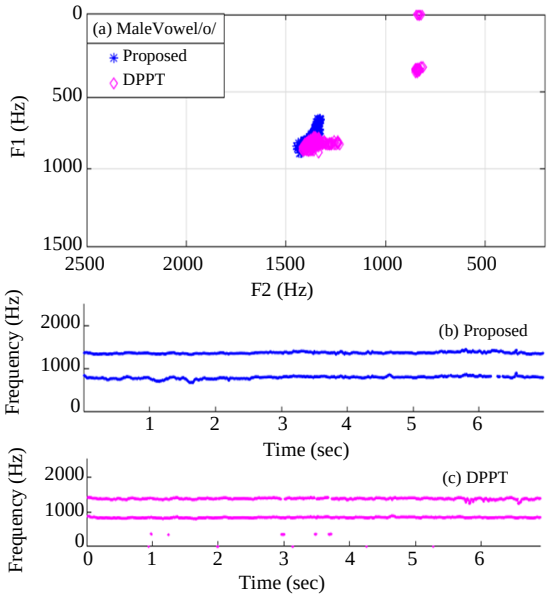

Fig. 13: Vowel /a/ spoken by a male speaker. (a) F2-F1 plot. Formant contour of F1, F2 (b) Proposed method (c) DPPT method.

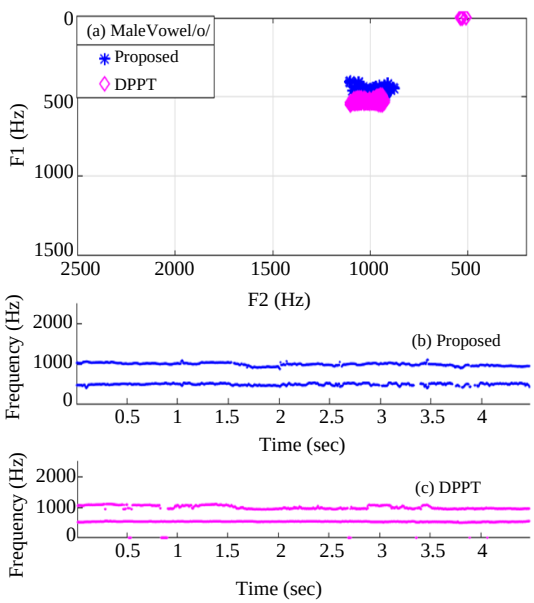

Fig. 14: Vowel /o/ spoken by a male speaker. (a) F2-F1 plot. Formant contour of F1, F2 (b) Proposed method (c) DPPT method.

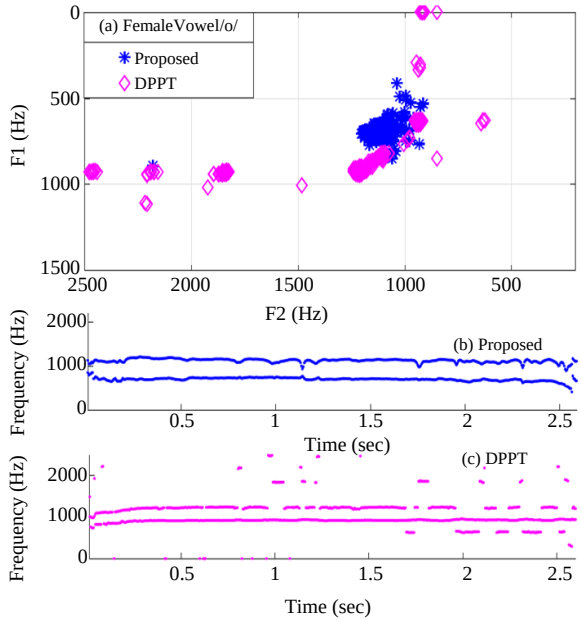

Fig. 15: Vowel /a/ spoken by a female speaker. (a) F2-F1 plot. Formant contour of F1, F2 (b) Proposed method (c) DPPT method.

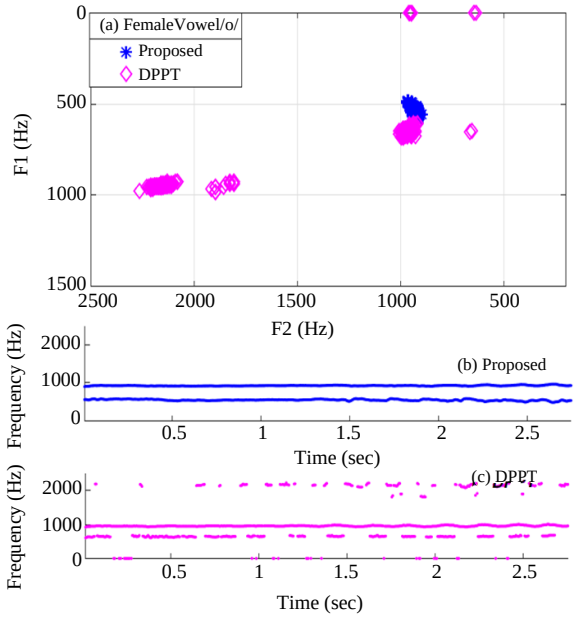

Fig. 16: Vowel /o/ spoken by a female speaker. (a) F2-F1 plot. Formant contour of F1, F2 (b) Proposed method (c) DPPT method.

From a close inspection on formant contour in all of the cases, F1 and F2 values are almost stable in the proposed method. In the DPPT method formants are shifted towards the next lower and higher formant regions. From these observations, we can say that the proposed method is more reliable to analyze the speech signal for both male and female speech.

Conclusion

In this research, we explored the evolution of the phase spectrum to elevate the accuracy of speech signal analysis. Based on pure phase domain analysis a method has been developed to deconvolve the signal into vocal tract filter and glottal source. For phase domain analysis we converted the speech signal to stable and minimum phase signal. We used the GD spectrum as a representative of PS in this research. From the analysis throughout the paper, it is evident that the properly estimated GD spectrum shows high-resolution pitch harmonics even with low amplitude. This forms the basis of source-filter separation. By parametric modeling of the signal obtained from this GD spectrum, the resonance frequencies can be extracted with elevated accuracy. All of the experiments conducted in this study shows that the proposed GD spectrum can be a trustworthy tool to analyze the speech signal.

Acknowledgement

We are grateful to Information and Communication Technology (ICT) Division of Bangladesh Government for their grant to conduct this research.

Authors Contribution

Husne Ara Chowdhury: Original conception, Literature, Data analysis and algorithm design, drafted the article and produce the figures used in the manuscript.

Mohammad Shahidur Rahman: Contributed in the conception of the designed algorithm, reviewed the manuscript sincerely and gave the approval of the final version of the manuscript.

Ethics

This manuscript has not been published anywhere. The confirmation from the corresponding author is that all of the other authors have read and approved the manuscript and there are no ethical issues involving this manuscript.

References

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. The journal of the acoustical society of America, 50(2B), 637-655.

Baldi, A., Bertolino, F., & Ginesu, F. (2000). Phase unwrapping algorithms: a comparison. In Interferometry in Speckle Light (pp. 483-490). Springer, Berlin, Heidelberg.

Bozkurt, B. (2005). Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals. Université Polytechnique de Mons, Belgium and LIMSI-CNRS, France (October 2005).

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

Bozkurt, B., Dutoit, T., Doval, B., & d'Alessandro, C. (2004a). Improved differential phase spectrum processing for formant tracking. In Eighth International Conference on Spoken Language Processing.

Bozkurt, B., Doval, B., d'Alessandro, C., & Dutoit, T. (2004b, September). Appropriate windowing for group delay analysis and roots of z-transform of speech signals. In 2004 12th European Signal Processing Conference (pp. 733-736). IEEE.

Bozkurt, B., Dutoit, T., Doval, B., & d'Alessandro, C. (2004c). A method for glottal formant frequency estimation. In Eighth International Conference on Spoken Language Processing.

Deepak, K. T., & Prasanna, S. R. M. (2015). Epoch extraction using zero band filtering from speech signal. Circuits, Systems and Signal Processing, 34(7), 2309-2333.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014, May). COVAREP—A collaborative voice analysis repository for speech technologies. In 2014 IEEE international conference on acoustics, speech and signal processing (icassp) (pp. 960-964). IEEE.

Deller, J. R., Proakis, J. G., & Hansen, J. H. (2000). Discrete-time processing of speech signals. Institute of Electrical and Electronics Engineers.

Duncan, G., Yegnarayana, B., & Murthy, H. A. (1989, May). A nonparametric method of formant estimation using group delay spectra. In International Conference on Acoustics, Speech and Signal Processing, (pp. 572-575). IEEE.

Fant, G., Liljencrants, J., & Lin, Q. G. (1985). A four-parameter model of glottal flow. STL-QPSR, 4(1985), 1-13.

Gowda, D. N., Pohjalainen, J., Kurimo, M., & Alku, P. (2013, September). Robust formant detection using group delay function and stabilized weighted linear prediction. In INTERSPEECH (pp. 49-53).

Loweimi, E. (2018). Robust Phase-based Speech Signal Processing From Source-Filter Separation to Model-Based Robust ASR (Doctoral dissertation, University of Sheffield).

Loweimi, E., Ahadi, S. M., & Drugman, T. (2013, May). A new phase-based feature representation for robust speech recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7155-7159). IEEE.

Loweimi, E., Barker, J., & Hain, T. (2015, September). Source-filter separation of speech signal in the phase domain. In 16TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (INTERSPEECH 2015), VOLS 1-5 (pp. 598-602). ISCA.

Magi, C., Pohjalainen, J., Bäckström, T., & Alku, P. (2009). Stabilised weighted linear prediction. Speech Communication, 51(5), 401-411.

Makhoul, J. (1975). Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4), 561-580.

Murthy, H. A., & Gadde, V. (2003, April). The modified group delay function and its application to phoneme recognition. In 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003. Proceedings.(ICASSP'03). (Vol. 1, pp. I-68). IEEE.

Murthy, H. A., & Yegnanarayana, B. (1991). Formant extraction from group delay function. speech communication, 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, H. A., Murthy, K. M., & Yegnanarayana, B. (1989). Formant extraction from phase using weighted group delay function. Electronics Letters, 25(23), 1609-1611.

Noll, A. M. (1967). Cepstrum pitch determination. The journal of the acoustical society of America, 41(2), 293-309.

Oppenheim, A. V., Buck, J. R., & Schafer, R. W. (2001). Discrete-time signal processing. Vol. 2. Upper Saddle River, NJ: Prentice Hall.

Rabiner, L. R. (1978). Digital processing of speech signal. Digital Processing of Speech Signal.

Rahman, M. S., & Shimamura, T. (2005). Formant frequency estimation of high-pitched speech by homomorphic prediction. Acoustical science and technology, 26(6), 502-510.

Rahman, M. S., & Shimamura, T. (2007). Linear prediction using refined autocorrelation function. EURASIP Journal on Audio, Speech and Music Processing, 2007, 1-9.

Vijayan, K., & Murty, K. S. R. (2015). Analysis of phase spectrum of speech signals using allpass modeling. IEEE/ACM Transactions on Audio, Speech and Language Processing, 23(12), 2371-2383.

Yegnanarayana, B. (1978). Formant extraction from linear-prediction phase spectra. The Journal of the Acoustical Society of America, 63(5), 1638-1640.

Yegnanarayana, B., Duncan, G., & Murthy, H. A. (1988, September). Improving formant extraction from speech using minimum-phase group delay spectra. In Proc. of European Signal Processing Conference (EUSIPCO) (Vol. 1, pp. 5-8).

Yegnanarayana, B., Saikia, D., & Krishnan, T. (1984). Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Transactions on Acoustics, Speech and Signal Processing, 32(3), 610-623.

Zhu, D., & Paliwal, K. K. (2004, May). Product of power spectrum and group delay function for speech recognition. In 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (Vol. 1, pp. I-125). IEEE.