ACOUSTICAL LETTER

Effect of pole/zero manipulation in estimating the group delay spectrum

Husne Ara Chowdhury* and M. Shahidur Rahman

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

(*Received 8 January 2022, Accepted for publication 22 February 2022*)

Keywords: Magnitude spectrum, Phase spectrum, Group delay spectrum, Spikiness, Poles/Zeros

1. Introduction

Most speech signal analysis tasks ignore the use of phase spectrum and usually exploit the magnitude spectrum. However, some studies showed the importance of phase spectrum to preserve speech quality and intelligibility. The group delay (GD) spectrum is the alternative and meaningful representation of the phase spectrum. The negative derivative of the phase spectrum is called the GD spectrum. Researchers [1–4] use the GD spectrum to extract the basic features for speech recognition, synthesis, coding, voicing detection, etc.

The GD spectrum resembles the magnitude spectrum and related to it through cepstral coefficients [1]. Thus it is reasonable to extract features from the GD spectrum, likewise the magnitude spectrum. Some attractive properties of the GD spectrum [2] make it a high-resolution spectrum, which becomes good enough for estimating the formant frequencies.

Spikiness is the prime problem of the GD spectrum estimation. Bozkurt *et al.* in [3] estimated the GD spectrum from the magnitude spectrum and employed the analysis window other than the unit circle to avoid the spikes. However, this method shows lower performance for noisy speech owing to the noise effect on the poles/zeros near the unit circle.

In our previous research [4], we proposed a modified GD spectrum to extract the formant frequencies effectively, where we assured the system's stability by shifting the poles/zeros inside the unit circle and employed the magnitude spectrum for the GD spectrum calculation. As a result, the noise affected the GD spectrum at the poles/zeros near the unit circle. Thus the method is applicable for clean speech only.

In this letter, we compute the GD spectrum from the causal part of the magnitude spectrum and propose enhancing the GD spectrum by gradually shifting the poles/zeros inside the unit circle. Then we employed the enhanced GD spectrum for formant estimation. We also studied the poles/zeros manipulations in estimating the GD spectrum from noisy speech signals. We observed that the second-order and third-order GD spectrum and the poles/zeros radii from less than 1 down to 0.95 is more noise-robust than the first-order one.

2. Problem identification

2.1. Issues of magnitude spectrum

The discrete Fourier transform (DFT) of a windowed speech segment $x(n)$ is defined as

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

or alternatively

$$X(k) = |X(k)|e^{\phi_{X(k)}},$$

where

$$\phi_{X(k)} = \arctan\frac{X_{\mathrm{I}}(k)}{X_{\mathrm{R}}(k)}.$$

Here L is the frame length and $1 < k < N$. The $|X(k)|$ and $\phi_{X(k)}$ signify the magnitude and phase spectrum, respectively. The $X_{\mathrm{R}}(k)$ and $X_{\mathrm{I}}(k)$ represent the real and imaginary parts of the spectrum $X(k)$.

The magnitude spectrum is a well-received mathematical tool when analyzing the speech signal. Nonetheless, it suffers from different problems. The spectral leakage and low-resolution spectrum are two examples. The lack of integer multiple cycles of component spectra when segmenting the speech utterance causes leakage on the spectrum. Again, the multiplication of component spectra forms the magnitude spectrum, which reduces the overall resolution of the spectrum. The low-resolution magnitude spectrum can't identify the formant peaks at lower amplitude values. In the case of additive noise, the noise spectrum becomes multiplicative with the magnitude spectrum, which forms the noisy spectrum.

2.2. Motivation of using the GD spectrum

The main problem of using the GD spectrum is its spikiness. To evade this, we employed the magnitude spectrum for its calculation. Nevertheless, some spikes dominate at some poles/zeros having a radius equal to 1 [4]. The causal part of the magnitude spectrum mitigates this problem. Shifting the poles/zeros to a radius less than 1 can improve the performance of spectral estimation.

The GD spectrum is additive with its component spectra which provide a high-resolution spectrum. The poles of the transfer function show peaks, and zeros show valleys on the GD spectrum. Thus the harmonics are more explicit in the GD spectrum than the magnitude spectrum. The vocal-tract-dominated harmonics are manifested even in low amplitude [2]. In the case of additive noise, the noise spectrum is additive with it, which reveals noise immunity to some extent.

2.3. Calculation of GD spectrum

Suppose, $x_{\min}$ is the time domain signal obtained from the magnitude spectrum. The causal part of the magnitude spectrum shows the minimum phase signal property. To calculate the effective GD spectrum, we evaluated the signal

[*] e-mail: husna-cse@sust.edu

[doi:10.1250/ast.43.232]

$x_{\min}$ from the causal part of the magnitude spectrum and then shifted poles/zeros inside the unit circle. Taking the negative derivative of the phase spectrum of $x_{\min}$, we have the GD spectrum as

$$\tau(k) = -\frac{d\{arg(X_{\min}(k))\}}{dk},$$

where $X_{\min}(k)$ is the DFT of $x_{\min}$. Since the negative values of the GD spectrum violate the causality [5], we used only positive values as effective GD spectral values denoted by $\tau_1(k)$, which indicates the first-order GD spectrum. The second and third-order GD spectrum can be defined, respectively, as

$$\tau_2(k) = (\tau_1(k))^2,$$
$$\tau_3(k) = (\tau_1(k))^3.$$

2.4. LP (Linear Prediction) spectrum calculation from GD spectrum

A time-domain equivalent signal is obtained by applying the inverse discrete Fourier transform (DFT) on the GD spectrum. Levinson-Durbin's algorithm [6] is then employed to find the AR (auto-regressive) coefficients. Finally, the DFT of the AR coefficients yields the LP spectrum.

3. Effects of pole/zero manipulation

Since the poles signify the spectral peaks and zeros indicate the spectral deep, the GD spectrum calculated from the causal part of the magnitude spectrum reflects the harmonics more clearly. As mentioned earlier, some spikes dominate at some poles/zeros having a radius equal to 1. When the poles/zeros radii decrease gradually from 1 down to 0.95, the vocal tract dominating harmonics become elicited, suppressing the other harmonics. The respective LP spectrum also shows accurate formant peaks when decreasing the radius, as shown in Fig. 1. The LP spectrum in Figs. 1(a) and 1(b) corresponding to poles/zeros radii at 1 and 0.99, respectively, show the shifted formant peaks, whereas the same in Figs. 1(c) and 1(d) corresponding to pole/zero radius at 0.98 and 0.97, respectively, are estimated more reliably.

Figure 2 shows the effect of pole/zero manipulation when estimating the GD spectrum from noisy speech. The LP spectrum plotted on the respective GD spectrum fails to emphasize the formant peaks in Figs. 2(a) and 2(b). However, in Figs. 2(c) and 2(d), formant peaks are apparent with a reduced noise effect. The accuracy of formant peaks improves after decreasing the poles/zeros radii from less than 1 down to 0.95. If we further decrease the poles/zeros radii to less than 0.95, the formant peaks become obscured. To reduce the number of illustrations, all figures in this letter show the pole/zero manipulation effect corresponding to the radius from 1 down to 0.97. Thus it is evident that the reduction of poles/zeros radii suppresses the noise and emphasizes the formant peaks.

4. Experimental results

4.1. Synthetic clean speech analysis

For generating the synthetic vowels, we simulated the source using the Liljencrants-Fant glottal model [7]. The formant frequencies used for synthesizing the vowels are given in Table 1 in [4]. Bandwidths of the five formants of all the five vowels are fixed to 60, 100, 120, 175, and 281 Hz, respectively. The sampling frequency is 10 kHz. We employed a Gaussian window of 30 ms for framing the speech. A pre-emphasis filter $1 - z^{-1}$ is applied to each speech segment. We used a 1,024 point DFT with the analysis order 12 to analyze the spectrum. We calculated the AR coefficients from the time domain equivalent signal of the GD spectrum employing Levinson-Durbin's algorithm [6]. After root-solving, we obtained the formant values. We analyzed by shifting every frame by 5 ms and then calculated the average of the first three formant estimation errors in percentage using the formula

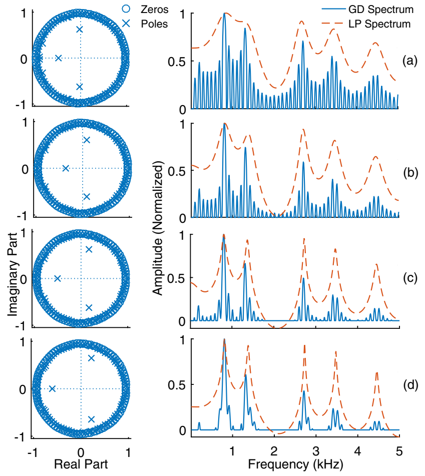

Fig. 1 Poles/zeros plot and the GD spectrum along with respective LP spectrum of a synthetic vowel /a/ in case of the poles/zeros (a) at radius = 1, (b) at radius = 0.99, (c) at radius = 0.98, and (d) at radius = 0.97.

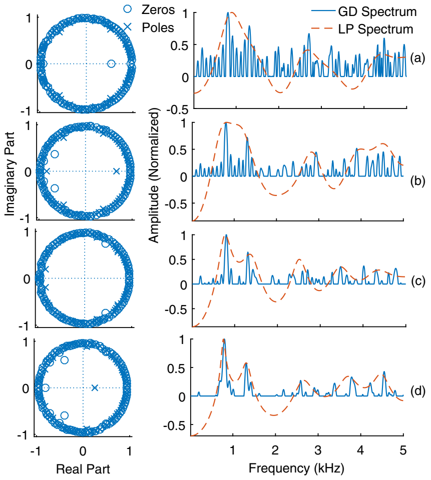

Fig. 2 Poles/zeros plot and the GD spectrum along with respective LP spectrum of a noisy synthetic vowel /a/ at SNR = 15 dB, in case of the poles/zeros (a) at radius = 1, (b) at radius = 0.99, (c) at radius = 0.98, and (d) at radius = 0.97.

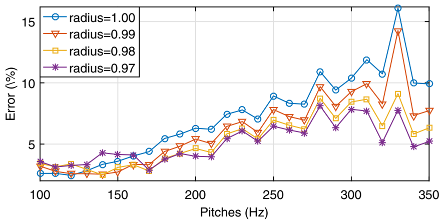

Fig. 3 The average of the first three formant estimation errors (%) of five synthetic vowels at different pitches up to 350 Hz was calculated using the first-order GD spectrum at different poles/zeros radii.

$$E_{\text{avg}} = \left(\frac{1}{5}\frac{1}{3}\sum_{i=1}^{5}\sum_{i=1}^{3}|\hat{F}_{ij} - F_{ij}|/F_{ij} \right) * 100,$$

where F_{ij} indicates the *ith* formant frequency of the *jth* vowel and $\hat{F}_{ij}$ is the estimated value.

Figure 3 shows the experimental results on clean synthetic speech. The average of the first three formant frequencies is obtained from the GD spectrum after changing the poles/zeros radii from 1 down to 0.97. From Fig. 3, it is explicit that the formant estimation error reduces gradually with the decrease in pole/zero radius.

4.2. Synthetic noisy speech analysis

We added white Gaussian noise with the speech signal that corrupts the poles/zeros near the unit circle of the noisy signal. As a result, spectral distortion occurs, which causes deviation of the formant peaks at poles/zeros radii equal to 1. Shifting the poles/zeros inside the unit circle enhances the spectrum. We have analyzed the noise effect on the first-order, second-order, and third-order GD spectrum by estimating the formant frequencies in the case of poles/zeros radii variation at different SNRs. From the results in Fig. 4, we observed that the noise effect is higher in the case of the first-order GD spectrum, and the accuracy of formant frequencies improves in the second and third-order GD spectrum at the reduced poles/zeros radii (less than 1).

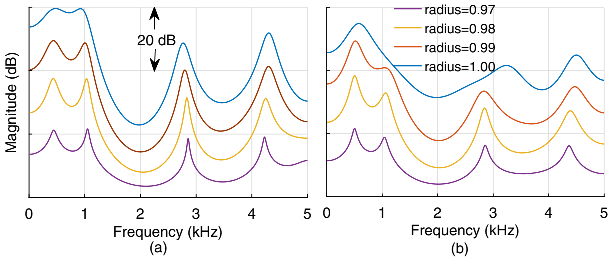

Fig. 5 The LP spectrum estimated employing the GD spectrum after manipulating the poles/zeros of different radius values for the vowel /o/ spoken by a male speaker (a) in a clean case and (b) in a noisy (SNR = 10 dB) case.

4.3. Natural vowel analysis

We set all preconditions the same as synthetic speech analysis for natural vowel analysis. Figure 5 shows the result of the analysis of a vowel /o/ spoken by a male speaker. For the clean case, we employed a first-order GD spectrum. On the other hand, a third-order GD spectrum is utilized for noisy vowel analysis. Since we manipulated the poles/zeros to increase the radii from 0.97 up to 1, the formant peaks on the LP spectrum of clean speech deviate, whereas obscured along with the deviation for noisy vowel as shown in Figs. 5(a) and 5(b). Thus it is evident that the formant estimation of vowels using the GD spectrum at poles/zeros radii less than 1 is more reliable both in clean and noisy cases.

4.4. Natural utterance analysis

We analyzed an utterance from TIMIT [8] database maintaining all preconditions as natural vowel analysis. The Left and the Right column in Fig. 6 shows the results on the

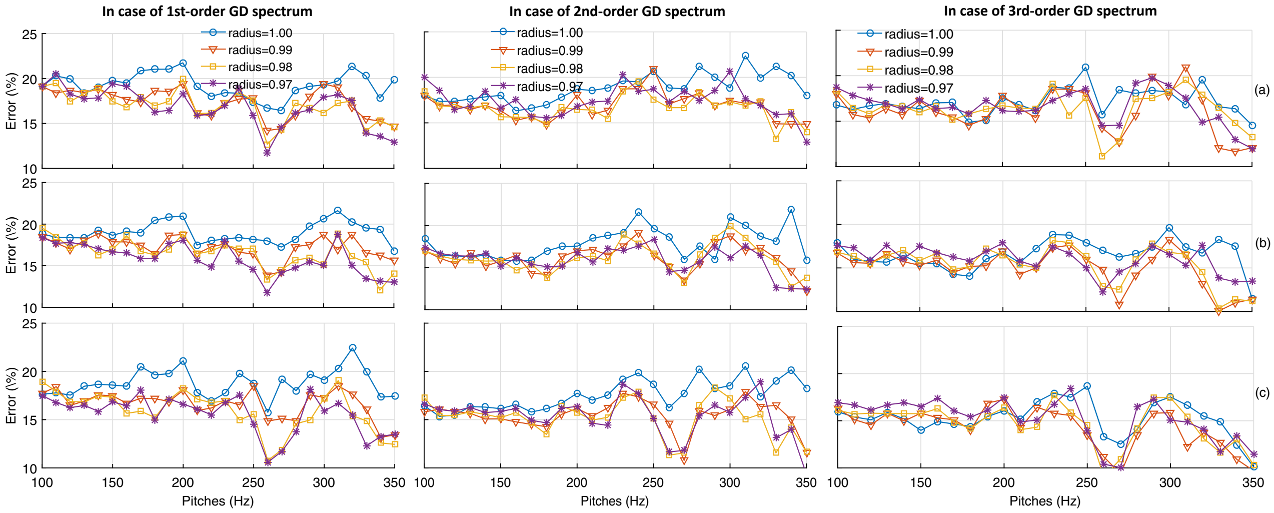

Fig. 4 The average of the first three formant estimation error (%) of five noisy synthetic vowels at (a) SNR = 10 dB, (b) SNR = 15 dB, and (c) SNR = 20 dB was calculated at different poles/zeros radii. The first column, second column, and third column indicate the poles/zeros manipulation effect on the first-order, second-order, and third-order noisy GD spectrum, respectively.

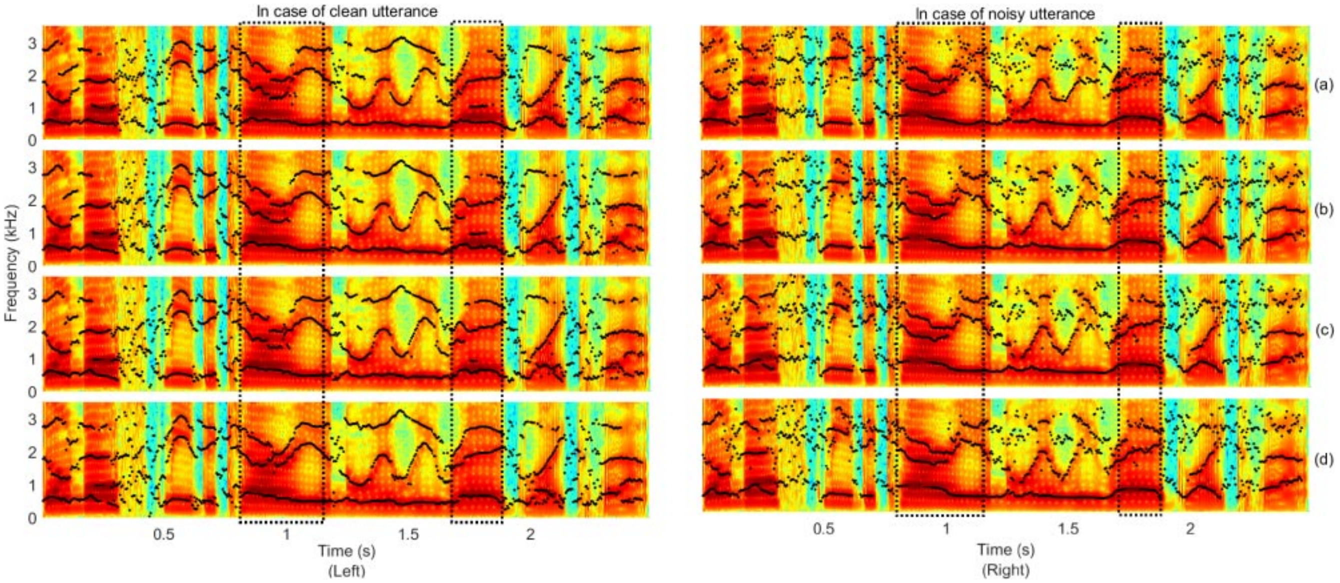

Fig. 6 The spectrogram along with three formant contours estimated from an utterance of "Don't ask me to carry an oily rag like that" in clean (Left) and noisy (Right) case at SNR = 15 dB. Formants are calculated from poles/zeros with (a) radius = 1, (b) radius = 0.99, (c) radius = 0.98, and (d) radius = 0.97.

clean and noisy utterance of "Don't ask me to carry an oily rag like that," respectively, where the formant contours are plotted on the respective spectrograms. The formants are estimated using the first- and third-order GD spectrum for clean and noisy cases. We added the Gaussian white noise to produce the noisy speech with the speech signal. In both cases, Fig. 6(a) represents the formant contours before poles/zeros manipulation (i.e., with radius 1), which exhibits some spurious values as depicted in the selected region. However, the accuracy improves gradually as the poles/zeros radii decrease from the radius of 1 down to 0.97, illustrated in Figs. 6(b), 6(c), and 6(d). The second and third formant contours become more consistent as the poles/zeros radii decrease.

5. Conclusion

In this letter, we analyzed the impacts of pole/zero manipulation on different-order GD spectra. Empirical results suggest that the poles/zeros radii to less than 1 down to 0.95 enhances the GD spectrum and reduces the noise effect resulting in improved formant estimation accuracy.

References

[1] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Commun.*, **10**, 209–221 (1991).

[2] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, **36**, 745–782 (2011).

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

[4] H. A. Chowdhury and M. S. Rahman, "Speech signal analysis in phase domain," *J. Comput. Sci.*, **16**, 1115–1127 (2020).

[5] E. Loweimi, "Robust phase-based speech signal processing from source filter separation to model based robust ASR.," *Ph.D. Dissertation, University of Sheffield* (2018).

[6] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, **28**(3), 233–244 (1960).

[7] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, **26**(4), 1–13 (1985).

[8] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, **9**, 351–356 (1990).