Phase-only Speech Reconstruction Using Very Short Frames

Erfan Loweimi, Seyed Mohammad Ahadi and Hamid Sheikhzadeh

Speech Processing Research Laboratory
Electrical Engineering Department, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran
{eloveimi, sma, hsheikh}@aut.ac.ir

Abstract

This paper aims to investigate potentials existing in speech phase spectrum. We observed that the window shape and scale incompatibility error (SIE) are two important factors which deeply influence the quality of phase-only reconstructed speech. After evaluating effects of different windows, we found Chebyshev window with dynamic range of 25 to 30 dB the best option. Inspiring from Hilbert transform relations, we removed the SIE and found the reason for quality improvement of ordinary phase-only reconstructed speech by frame length extension. Results show that phase spectrum, even in very short frame lengths such as 16 ms, can be highly informative.

Index Terms: phase spectrum, Chebyshev window, Hilbert transform, scale incompatibility error (SIE)

1. Introduction

It is well accepted among the speech technologists that phase spectrum does not play a significant role in speech processing. Taking a glance on the speech enhancement algorithms and speech recognition feature extraction methods proves this point. According to some well-known researches there are two main reasons for dismissing phase spectrum in speech processing: phase wrapping as well as the speech signal phase spectrum being informative only in long frame lengths [1], [2]. Phase wrapping complicates the interpreting and processing of the phase spectrum. On the other hand, owing to non-stationarity of speech signal, long frame lengths are not applicable. As a result, phase spectrum has not been appealing for the researchers. Due to these two problems, the researchers opted to work on other representations of phase spectrum such as group delay function [3].

Oppenheim and Lim [2] showed that in frames as long as one second, the speech phase spectrum is informative. In this case the phase-only reconstructed speech is intelligible. However, the quality of the magnitude-only reconstructed speech in similar situations is very low due to non-stationarity of the signal. Liu, He, and Palm [4], through speech recognition tests, showed that the intelligibility of phase-only reconstructed speech increases by extending the frame length. They observed that the intelligibility of phase-only reconstructed speech surpasses the intelligibility of magnitude-only reconstructed speech in frames longer than 128 ms. Alsteris and Paliwal [5], [6] within the same framework of Liu *et al*. [4], showed that the intelligibility of phase-only reconstructed speech can be notably influenced by the window type. They compared the results of applying rectangular window with those of Hamming window and observed that using rectangular window improves the intelligibility of phase-only reconstructed speech, even in frame lengths as short as 32 ms.

Liu, He, and Palm [4] as well as Alsteris and Paliwal [5], [6] studied the information content of phase spectrum in a non-iterative manner. They constructed phase-only stimuli by discarding the magnitude spectrum information through replacing the magnitude spectrum by unity. Similarly, via substituting the phase spectrum with a sequence of uniformly distributed random numbers in range of $(-\pi, \pi)$ or $(0, 2\pi)$, magnitude-only stimuli were reconstructed. Then, they played the stimuli to a set of listeners and computed the recognition rates in different situations. In order to further investigate the potentials of phase spectrum, in contrast with previous works ([4], [5], and [6]), we reconstructed the phase and magnitude-only stimuli in an iterative manner instead of just discarding one of these two spectra. It appears that this approach provides a better way for determining and estimating the information content of phase and magnitude spectra. The reason behind this will be discussed more in Section 2.

Since the speech signal is local-stationary, it must be processed in a frame-wise manner. The length of frames typically is in the range of 20 to 40 ms. The applicability of phase spectrum in speech processing depends on the answer of the question that whether phase spectrum is informative in short frame lengths within this range. As already mentioned, it was known since 1979 [1] that in long frame lengths it is informative. However, there has not been any positive answer for short frame lengths. The approach used to investigate the information content of phase spectrum has been discarding the magnitude spectrum information and evaluating the quality and/or intelligibility of phase-only reconstructed speech. Hence, we follow this approach in this paper. Besides, we compare the results with those of the magnitude-only reconstructed speech due to its dominant role in speech processing.

It appears that there are some points that should be investigated in more detail in order to provide a comprehensive answer for the aforementioned question. There has been no explanation yet on why the information content of phase-only reconstructed speech increases by extending the frame length, although this trend has been observed in [1], [2], [4], [5], and [6]. Incidentally, the influence of window shape on the intelligibility of phase-only stimuli in [5] and [6] motivates search for other types of windows which can further improve the quality and/or intelligibility of phase-only reconstructed speech.

In this paper, we will show that in light of applying the appropriate window type the quality of phase-only reconstructed speech would surpass the quality of its magnitude-only counterpart in frames longer than 75 ms. We will also analyze the reason for quality improvement of phase-only reconstructed speech by frame length extension. Finally, we will show that by applying appropriate window type and inspiring from Hilbert transform relations, phase spectrum even in short and very short frame lengths such as 32 and 16 ms can be notably more informative than magnitude spectrum.

The organization of the rest of this paper is as follows. In Section 2 the advantages of iterative speech reconstruction will be discussed. Section 3 explains the utilized quality assessment method. Section 4 presents the

first part of our simulation results, investigating the importance of window type on the quality of phase and magnitude-only reconstructed speech. In Section 5, the reason of quality improvement of phase-only reconstructed speech by frame length extension is discussed and Section 6 concludes the paper.

2. Iterative speech reconstruction

Generally, both phase and magnitude spectra are required for unique reconstruction of a typical signal from its Fourier transform components. Under certain conditions, Hilbert transform relations provide a close form for specifying each spectrum from another and consequently determining the signal in time domain [7]. In case of minimum or maximum phase signals, these relations work [7]. However, speech signal is neither minimum nor maximum phase signal. The complex cepstrum of a minimum phase signal must be causal and similarly maximum phase signal must have anti-causal complex cepstrum [7]. Figure 1 shows the complex cepstrum of a typical speech signal. As seen, the speech signal is a mixed-phase signal, so the Hilbert transform relations cannot be applied.

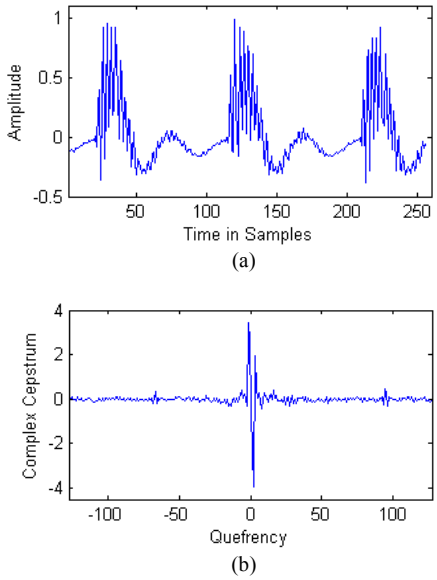

Figure 1: *(a) Signal of a vowel (/a/) and (b) complex cepstrum of it.*

If the phase (or magnitude) spectrum can be reconstructed from magnitude (or phase) spectrum, it implies that the information content of both spectra is identical. However, for a mixed-phase signal such as speech, the information content of phase and magnitude spectra are not the same. As said before, reconstructing the signal from each spectrum and comparing the result with the original signal through appropriate measures determines the information content of that spectrum. In contrast to previous studies ([4], [5], and [6]), we have used iterative reconstruction algorithm. It has been shown that the distance between the original signal and the iteratively reconstructed one decreases in each iteration in mean square sense [8]. Here, we have examined the quality of phase-only reconstructed speech through PESQ [9] objective measure which has the maximum correlation with subjective tests and is highly reliable [10]. As seen in Figure 2, the quality of the reconstructed speech rapidly decreases for a number of iterations. It also shows that the information content of the phase (and similarly magnitude) spectrum as well as its potential for reconstructing the original signal is high and just performing one iteration (non-iterative manner) does not reveal this potential. Iterative signal reconstruction is mathematically simple and in case of phase-only reconstruction it does not need unwrapped phase spectrum, as needed in Hilbert transform relations.

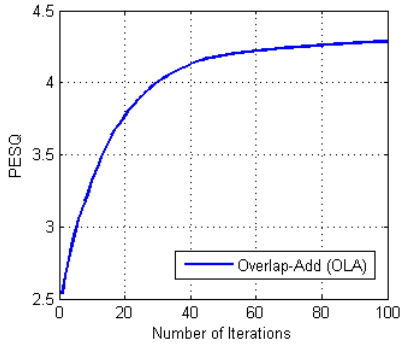

Figure 2: *PESQ versus number of iterations in case of phase-only speech reconstruction via OLA.*

3. Quality assessment

To evaluate the quality of speech signals, the ideal solution is systematic subjective tests on a large population. However, in practice, usually the available population is restricted. This leads to highly variable and sometimes ambiguous results. In addition, the whole process may take too long, which is not usually practical. As well, in subjective recognition tests each listener classifies the played signal correctly or incorrectly, so the intelligible speeches with different qualities categorized in the same class (correctly recognized), however, their different quality is an indicator of their different information content. A reliable objective measure does not classify two intelligible signals with different qualities within the same class. Therefore, it is more illuminating in comparison with human recognition tests, in order to investigate the information content of each spectrum. Regarding these points, we have used PESQ [9] objective measure because of its high reliability and correlation with subjective tests [10].

4. Investigating the importance of window shape on the quality of phase and magnitude-only reconstructed speech

In this Section, we will discuss the importance of window shape on the quality and intelligibility of phase and magnitude-only reconstructed speech. We have used 10 speech signals of NOIZEUS database [11] in this part of simulations from 5 male and 5 female speakers. The speech signals are segmented with frame lengths of 32 to 1024 ms. Frames overlap and number of iterations are set to 75% and 100, respectively. The FFT length is set to $2N$ where N is the number of samples of each frame. Effects of many types of windows such as rectangular, triangular (Bartlett), Hanning, Hamming, and Chebyshev with different dynamic ranges from 20 to 110 dB have been studied. Overlap-add (OLA) and least square error estimation (LSEE) [8] have been used as synthesis methods for phase and magnitude-only speech reconstruction, respectively, due to our previous studies [12], [13]. More details about the applied iterative reconstruction process could be found in [12] and [13].

As seen in Tables 1 and 2, the window shape can play a notable role on the quality of reconstructed speech for both phase and magnitude spectra. In case of magnitude-only speech reconstruction, Hamming window is the best option among the windows studied here. This shows why it is

frequently used in speech processing where most of the algorithms focus on magnitude spectrum. However, it is not a suitable window for phase spectrum at all. As seen, in frame length of 32 ms, applying appropriate window instead of Hamming window increases the quality of phase-only reconstructed speech from 2.4 to 3.8 in PESQ scale. Table 2 shows that the most appropriate window for speech reconstruction based on phase spectrum is Chebyshev with a dynamic range of 25 to 30 dB. It seems that smear-leakage trade-off which is provided by Hamming and Chebyshev (25 to 30 dB) windows is the main reason of making them the best options for working with magnitude and phase spectra, respectively.

Table 1. *Quality of the Magnitude-Only Reconstructed Speech via LSEE in PESQ.*

	Frame Length (ms)					
	32	64	128	256	512	1024
Rect.	4.00	3.91	3.61	3.05	2.36	1.90
Tri.	4.18	3.99	3.54	2.73	1.95	1.18
Han.	3.34	3.37	3.19	3.18	0.81	0.50
Ham.	**4.25**	4.09	3.83	3.28	2.66	1.43
Ch-20	2.42	2.44	2.29	2.18	1.91	1.95
Ch-25	2.54	2.54	2.40	2.29	2.10	1.88
Ch-30	2.86	2.65	2.53	2.43	2.27	1.77
Ch-35	3.40	2.90	2.61	2.57	2.45	1.87
Ch-40	3.78	3.36	2.88	2.72	2.37	1.53
Ch-45	4.08	3.60	3.30	2.93	2.58	1.73
Ch-50	4.18	3.86	3.50	3.09	2.63	1.48
Ch-80	4.19	4.10	3.95	3.53	1.64	0.77
Ch-110	3.61	3.54	3.44	3.13	0.75	0.53

Table 2. *Quality of Phase-Only Reconstructed Speech via OLA in PESQ.*

	Frame Length (ms)					
	32	64	128	256	512	1024
Rect.	3.51	3.55	3.66	3.78	3.83	3.59
Tri.	1.85	1.96	2.03	2.33	2.58	2.53
Han.	1.37	2.25	1.92	2.48	0.89	0.77
Ham.	2.41	2.47	2.58	2.80	2.99	2.88
Ch-20	3.78	3.98	4.04	4.20	4.12	4.04
Ch-25	**3.78**	**3.98**	**4.09**	**4.24**	**4.17**	**4.21**
Ch-30	3.73	3.96	4.05	4.20	4.16	4.24
Ch-35	3.63	3.88	3.97	4.12	4.05	4.08
Ch-40	3.36	3.65	3.81	3.92	3.86	3.80
Ch-45	2.92	3.26	3.54	3.64	3.60	3.44
Ch-50	2.57	2.83	3.14	3.32	3.33	3.08
Ch-80	1.88	1.85	2.02	2.36	2.72	2.63
Ch-110	1.88	2.10	2.29	2.53	2.31	1.54

In order to come in more reliable results for comparing the information content of phase and magnitude spectra we used all 30 speech signals of NOIZEUS, increased the overlap to 87.5%, and applied the best window type and synthesis method over different frame lengths. As seen in Figure 3, cross over point occurs in frame length of 75 ms.

Another important point is that the quality of phase-only reconstructed speech, contrary to that of its magnitude-only counterpart, does not change drastically by frame length extension. It shows the high potential of phase spectrum to be used in non-stationary signal processing. In addition, as seen, the quality of the magnitude-only reconstructed speech decreases by frame length extension because of non-stationarity of speech signal. However, the quality of phase-only reconstructed speech improves by frame length extension. This trend has been observed in [1], [2], [4], [5], and [6] but in none of them has been reasoned. In the next section we will explain why this happens.

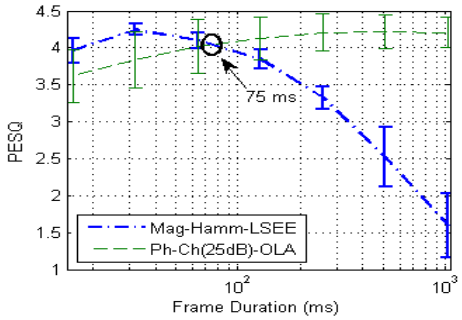

Figure 3: *PESQ of phase-only and magnitude-only reconstructed speech versus frame length (16, 32, 64, 128, 256, 512, and 1024 ms).*

5. Frame length and quality of phase-only reconstructed speech

As said before, the Hilbert transform relations cannot be applied to speech signal. However, they could be inspiring. As it is known, from phase spectrum, one can reconstruct the signal up to a scale error [7]. Hilbert transform relations show that the scale error depends on the value of the complex cepstrum in zero, i.e. $\hat{x}(0)$ [7]. In case of non-stationary signals segmentation into frames is necessary. When one reconstructs such a signal from its short-time phase spectra, the scale errors of different frames are not identical since the values of $\hat{x}(0)$ are not the same over different frames. Figure 4 shows the values of this parameter over different frames for a typical speech signal. As seen, this parameter, even for frames which have high overlap, is not the same. So, each phase-only reconstructed frame has a different scale error in comparison with the corresponding frame in the original signal. Hence, the adjacent frames which have overlap and should be added to each other in synthesis stage are not scale-compatible. This will have negative effect from psychoacoustic point of view and decreases the intelligibility and/or quality of the synthesized speech. Therefore, this may be accounted for as a reason for loss of quality in synthesizing the signal using phase spectrum, which otherwise may lead to a higher quality.

We inspire again from Hilbert transform relations in order to provide a suboptimal solution for scale-incompatibility problem. These relations show that the exact value for magnitude spectrum is achieved by multiplying what have been provided by phase spectrum in $e^{\hat{x}(0)}$ [7]. Keeping this point in mind, instead of initializing the magnitude spectrum with unity, we initialized the magnitude spectrum with a constant number which is $e^{\hat{x}(m,0)}$, where m is the frame number. Figure 5 shows the result of initializing the iterative phase-only speech reconstruction through the proposed method in case of applying the best window type and synthesis method over different frame lengths. As seen, the phase-only reconstructed speech has a very high quality even in case of short frame lengths such as 32 and 16 ms. We also observed that initializing the magnitude spectrum of each frame with the maximum value of the magnitude spectrum of that frame results in a high quality signal.

In order to come in a quantitative demonstration to check our argument from another point of view, we introduce the following error for different frame lengths between 1 (typical initial value for magnitude spectrum in

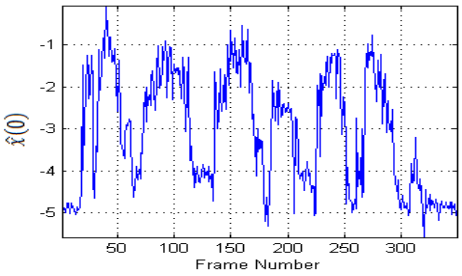

Figure 4: *variation of $\hat{x}(0)$ in frame length of 32 ms for a speech signal from NOIZEUS versus frame number. Overlap is 75%.*

phase-only speech reconstruction) and $e^{\hat{x}(m,0)}$ (inspired from Hilbert transform relations to remove the scale incompatibility) and call it scale incompatibility error (SIE)

$$\text{SIE} = \sum_{all\ frames}\left(1 - e^{\hat{x}(m,0)}\right)^2. \quad (1)$$

As Figure 6 shows, SIE decreases and becomes less severe with frame length increase. That is why the quality of the ordinary phase-only reconstructed speech (initialized by unit magnitude spectrum) improves by frame length extension.

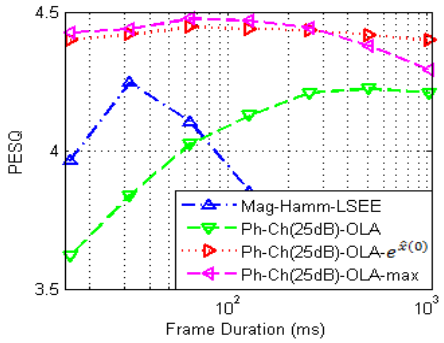

Figure 5: *Quality comparison of phase-only reconstructed speech based on proposed method with magnitude-only reconstructed speech versus frame duration (16, 32, 64, 128, 256, 512, and 1024 ms).*

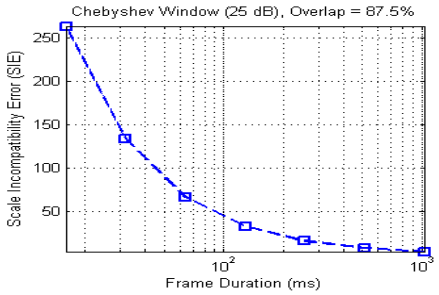

Figure 6: *Scale Incompatibility Error (SIE) versus frame duration (16, 32, 64, 128, 256, 512, and 1024 ms).*

Table 3. *Comparing the effect of Hamming and Chebyshev window with dynamic range of 25 dB.*

	Frame Duration (ms)					
	32	**64**	**128**	**256**	**512**	**1024**
Hamming	2.57	2.59	2.62	2.84	3.05	2.95
Che. (25 dB)	4.42	4.45	4.44	4.43	4.42	4.40

So far, we showed that in case of applying suitable window (Chebyshev with dynamic range of 25~30 dB) and appropriate initialization, the quality and/or intelligibility of phase-only reconstructed speech becomes highly notable. As seen in Table 3, applying Hamming window as well as the proposed initialization method results in a low quality signal, proving the importance of window shape.

The last point that should be discussed is the applicability of the phase spectrum. Although we only used the $\hat{x}(0)$ of each frame for initialization, one may see that the success of phase spectrum in high quality reconstruction of speech signal depends on the magnitude spectrum. It should be noted that the algorithms of speech processing are applied separately for each frame. The proposed method utilized the magnitude spectrum information just for removing the incompatibility scale error which occurs in synthesis process. However, in many applications such as feature extraction in speech recognition there is no synthesis stage, therefore, this information is not needed. Thus, the phase spectrum and its high information content in short frame lengths can be functional in that type of applications.

6. Conclusion

In this paper, we investigated the potentials of phase spectrum to be used in speech processing. The applicability of phase spectrum in this field depends on its information content in frame lengths in the range of 20 to 40 ms. We showed that through applying a suitable window, which is Chebyshev with dynamic range of 25 to 30 dB, and removing the scale incompatibility error (SIE) inspired from Hilbert transform relations, the intelligibility and/or quality of phase-only reconstructed speech becomes high, even in very short frame lengths such as 16 ms. We also argued that the decrease of SIE is the reason of quality improvement of ordinary phase-only reconstructed speech by frame length extension. Results show the high potential of phase spectrum to be used in speech and non-stationary signal processing.

7. References

[1] A. V. Oppenheim, J. S. Lim, G. E. Kopec, and S. C. Pohlig, "Phase in speech and pictures," in *Proc. IEEE Int. Conf. Acoust., Speech and SignalProcesring,* pp. 632-637, Apr. 1979.

[2] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, pp. 529-550, May 1981.

[3] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Process.*, vol. 17, pp. 141–150, 1989.

[4] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, pp. 403-417, 1997.

[5] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. of Eurospeech-2003*, pp. 2117–2120, 2003.

[6] L. D. Alsteris and K. K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 573–576, 2004.

[7] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[8] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process*. ASSP-32 No. 2, pp. 236–243, 1984.

[9] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[10] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement*," IEEE Trans. Audio, Speech, Lang. Process.* 16, 229–238, 2008.

[11] Y. Hu, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," http://www.utdallas.edu/loizou/speech/noizeus, 2005.

[12] E. Loveimi and S.M. Ahadi, "Objective Evaluation of Magnitude and Phase Only Spectrum-based Reconstruction of the Speech Signal*",* in *Proc. ISCCSP2010*, Cyprus.

[13] E. Loveimi and S.M. Ahadi, "Objective Evaluation of Magnitude and Phase Only Reconstructed Speech: new considerations*",* in *Proc. ISSPA 2010.*