Phase-only Speech Reconstruction Using Very Short Frames

Erfan Loweimi, *Student Member, IEEE*, Seyed Mohammad Ahadi, *Senior Member, IEEE* and Hamid Sheikhzadeh, *Senior Member, IEEE*

Speech Processing Research Laboratory
Electrical Engineering Department, Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran
{eloveimi, sma, hsheikh}@aut.ac.ir

ABSTRACT

In this paper we deal with a question which is not answered since 1979[1], "why the quality and intelligibility of phase-only reconstructed speech improves with frame length extension?" Hilbert transform relations state that by the use of phase spectrum one can compute the signal up to a scale. The scale error of phase-only reconstructed frames is not the same within adjacent frames. As a result, the reconstructed speech is synthesized from frames which have not compatible scales. We show quantitatively that the scale error decreases by frame length expansion. That is why the quality of phase-only reconstructed speech improves by frame length extension. At the end, based on Hilbert transform relations, we propose a method to overcome this problem. Phase-only reconstructed speech based on the proposed method surpasses its magnitude-only counterpart in all the frame lengths, particularly 16 and 32 ms, qualitatively.

Index Terms- Phase spectrum, speech reconstruction, Hilbert transform, scale error.

1. INTRODUCTION

It is well-known that the phase spectrum of the speech signal does not play a significant role in speech processing. It is believed that in the speech signal, almost all the intelligibility information exists in magnitude spectrum. That is why the majority of speech processing algorithms focus on magnitude spectrum. For example, in nearly all speech enhancement algorithms, such as [2], all the process is focused on magnitude spectrum. At the end of the process, phase spectrum of the noisy signal is combined with the enhanced magnitude spectrum and enhanced speech is reconstructed. The same is true for speech recognition. Most of the feature extraction methods such as MFCC and LPCC, only utilize magnitude spectrum and discard phase spectrum. Only in speech coding, one can see a more notable role for phase spectrum [3].

Generally, speech signal is not a minimum or maximum phase signal. As a result, both phase and magnitude spectra are needed to reconstruct the original signal. In such cases, one spectrum (e.g. phase) cannot be constructed from the other spectrum (e.g. magnitude). This implies that the information of primary signal is divided between the phase and magnitude spectra. Now the question is that how we can measure and compare the amount of information which exists in each spectrum. To answer this question, the speech signal should be reconstructed from phase- and magnitude-only spectra. Then, by the use of appropriate subjective or objective measures, the deal of information which exists in the reconstructed signals could be evaluated.

Oppenheim and Lim [1,4] were the first to carry out a remarkable study to investigate the importance of phase spectrum in a few types of signals such as image and speech from the signal processing viewpoint. In case of speech signal, they observed that using frame lengths of more than 1 sec, the phase-only reconstructed speech will be intelligible. It is obvious that the magnitude-only reconstructed speech using such long frames is not intelligible because of non-stationarity of speech signal. In fact, as speech is a local-stationary signal, it should be decomposed into segments of length 20-40 ms. As a result, long frames (1 sec or more) were not practical, so this point did not attract researchers' attention.

Later, Liu et al. [5] conducted a notable study in order to investigate the importance of phase spectrum in speech recognition. They decomposed the speech signals (stop consonants in intervocalic context) into frame lengths of 16 to 512 ms with 50% overlap and windowed the segments with Hamming window. They reconstructed the phase-only and magnitude-only stimuli via the overlap add method and played them for a number of listeners. Results showed that the recognition rate of phase-only reconstructed speech in frames longer than 128 ms exceeded the recognition rate of their magnitude-only counterpart.

More recently, Alsteris and Paliwal have shown that the window shape has a notable role in the intelligibility information which exist in phase spectrum [6]. They found that rectangular window was a more suitable option in comparison with Hamming window when dealing with phase spectrum [6].

All the aforementioned works demonstrate that the intelligibility information existing in the speech phase spectrum increases with frame length extension. However, in none of the cited works [1-6] the reason for this phenomenon has been discussed. On the other hand, it is clear that by increasing the frame length, the quality of the magnitude-only reconstructed speech decreases due to non-stationarity of the speech signal.

In this paper we intend to provide an answer to the question that why the quality of phase-only reconstructed speech increases in longer window lengths. Based on the Hilbert transform, we will analyze this phenomenon and will propose a modification in the reconstruction algorithm. Results show that the quality of phase-only reconstructed speech using the proposed method surpasses that of its magnitude-only counterpart even for short frame lengths such as 16 and 32 ms.

The organization of this paper is as follows: in Section 2 we will briefly review Analysis-Modification-Synthesis (AMS) framework which is used for signal reconstruction. Section 3 first offers a new insight into the problem of phase-only speech reconstruction. Then we will propose

modifications to the phased-based speech reconstruction algorithm and present results. Section 4 concludes the paper.

2. ANALYSIS-MODIFICATION-SYNTHESIS (AMS) FRAMEWORK

As speech is a local-stationary signal, short frames of speech are chosen for analysis to ensure the stationarity within frames. Then Fourier transform is applied to each segment to obtain the short-time Fourier transform (STFT). Let *X(m,ω)* be a STFT of *x(n)* where *m* is the frame number and *ω* denotes the frequency. Since *X(m,ω)* is a complex quantity, it can be written in a polar form as follows

$$X(m,\omega) = |X(m,\omega)|e^{j\angle X(m,\omega)}, \quad (1)$$

where $|X(m,\omega)|$ is the short-time magnitude spectrum and $\angle X(m,\omega)$ indicates the short-time phase spectrum. From here on, the short-time modifier is implied wherever Fourier transform is mentioned.

The next step is modification. The operation which is done in this stage is highly task-dependent. To investigate the importance of phase spectrum, one should reconstruct the signal only from its phase spectrum, discarding its magnitude spectrum. After comparing the reconstructed speech with the original speech by suitable measures, one can evaluate the information content of the phase spectrum. Usually, magnitude spectrum is replaced with a constant like 1,

$$X^1(m,\omega) = 1.e^{j\angle X(m,\omega)}, \quad (2)$$

where superscript 1 refers to initialization of an iterative reconstruction algorithm which will be discussed in this section.

Similarly, one can reconstruct the signal only from its magnitude spectrum. In this case, phase spectrum is discarded and typically replaced by a sequence (φ) of random numbers, uniformly distributed in the range of ($-\pi$, π)

$$X^1(m,\omega) = |X(m,\omega)|e^{j\varphi}. \quad (3)$$

Instead of random numbers, the phase spectrum could be simply replaced by zero.

The next step is synthesis in which the speech signal should be reconstructed from its segments. Overlap Add (OLA) and Least Square Error Estimation (LSEE) [7] are two well-known methods for synthesizing the signal. In the OLA the synthesis signal is obtained as

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}^i(mM,\omega)e^{j\omega n}d\omega}{\sum_{m=-\infty}^{\infty} w\,(mM-n)}, \quad (4)$$

where *M* is decimation factor, *i* indicates the iteration number and *w(n)* is the window. However, after modifying the spectrum of an arbitrary signal, there is no guarantee that the modified spectrum still remains a valid spectrum. Griffin and Lim [7] proposed the LSEE method in order to find a signal with nearest spectrum to the modified spectrum in the sense of mean square error (MSE). The basic equation of LSEE is [7]

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mM-n)\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}^i(mM,\omega)e^{j\omega n}d\omega}{\sum_{m=-\infty}^{\infty} w^2\,(mM-n)}. \quad (5)$$

In case of reconstructing the signal from its phase spectrum, after initializing the algorithm with Eq. (2), the following substitution should be performed in every iteration

$$\hat{X}^i(m,\omega) = \left|X^i(m,\omega)\right|\angle X(m,\omega). \quad (6)$$

Eq. (6) results in modifying the magnitude spectrum from its initialized value. Obviously, the obtained magnitude spectrum will be different from the original one because the speech signal is a mixed-phase signal. Similarly, in case of reconstructing the signal from its magnitude spectrum, the following substitution should be performed per iteration

$$\hat{X}^i(m,\omega) = |X(m,\omega)|\angle X^i(m,\omega). \quad (7)$$

3. PROBLEM STATEMENT, OUR SOLUTION AND PROPOSED MODIFICATON

As stated before, speech is not a minimum or maximum phase signal. Consequently, we cannot construct the phase (or magnitude) spectrum from magnitude (or phase) spectrum. This property shows that the information content of phase and magnitude spectra are not the same. In order to compare the information content of each spectrum, we can reconstruct the signal only from that spectrum and then compare the intelligibility or quality of the reconstructed signal with the original signal. It has been shown that the information content of each spectrum depends on the frame length, window type, frame shift (overlap) and number of iterations [5, 6, 10, 11].

We have described in [12] that Hamming window along with LSEE is almost the best choice for reconstructing the speech from its magnitude spectrum. In case of phase-only speech reconstruction, the Chebyshev window with dynamic range of 25 dB along with OLA is almost the optimal choice [12]. We have used all of the 30 signals of NOIZEUS database [13] in our experiments. Fig. 1 shows the quality of the reconstructed speech for different frame lengths. The overlap and number of iterations are set to 87.5% [6] and 100, respectively. FFT size is *2N* where *N* is the number of samples of each frame. In order to evaluate the quality of the reconstructed speech we have used the PESQ [8] objective measure which, according to Hu and Loizou [9], has the highest correlation with subjective tests.

As seen in Fig. 1, the crossover point of the qualities of phase-only and magnitude-only reconstructed speech lies in frame length of 75 ms. More details can be found in [12]. In this research, we deal with a question which has not been answered since 1979 [1]: “Why the quality of phase-only reconstructed speech increases with frame length extension?” In order to answer the question, let us review the Hilbert relations [14]

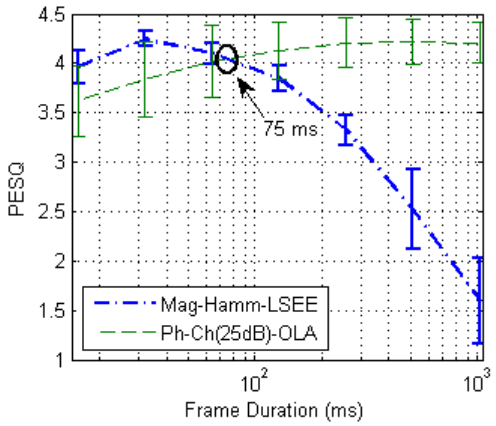

Figure 1*: PESQ of phase-only and magnitude-only reconstructed speech versus frame length (16, 32, 64, 128, 256, 512 and 1024 ms).*

$$arg[X(\omega)] = \frac{1}{2\pi}\rho\int_{-\pi}^{\pi}\ln|X(\omega)|\cot\left(\frac{\omega-\theta}{2}\right)d\theta, \quad (8)$$

$$\ln|X(\omega)| = \hat{x}(0) + \frac{1}{2\pi}\rho\int_{-\pi}^{\pi} arg[X(\omega)]\cot\left(\frac{\omega-\theta}{2}\right)d\theta, \quad (9)$$

$$\hat{x}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln|X(\omega)|\,d\omega, \quad (10)$$

where ln and ρ denote the natural logarithm and Cauchy principle value of the integrals, respectively.

As speech signal is not a minimum or maximum phase signal, the above equations cannot be applied[1], but they could be useful. In fact, they could be considered as a limit of information content of each spectrum. Eq. (8) implies that in order to reconstruct the signal from its magnitude spectrum we do not require additional information because the phase spectrum could be directly calculated from magnitude spectrum. On the other hand, Eq. (9) implies that having only the phase spectrum, we can reconstruct the magnitude spectrum only within a scale factor. In order to determine the magnitude spectrum exactly, the value of $\hat{x}(0)$ must be known.

In speech quality or intelligibility evaluation, the scale is not a significant factor. In other words, the scale of speech has no importance from the perceptual view point. However, in the analysis stage, speech signal is segmented with a specific overlap. In the synthesis step by overlapping and adding frames, the signal will be reconstructed. The point is that there is no guarantee for the scale factor to be equal in the adjacent frames which must be overlapped and added. Fig. 2 shows the variation of $\hat{x}(0)$ for an arbitrary speech signal. As depicted, $\hat{x}(0)$ changes noticeably over different speech frames. Assuming that the overlap is 75%, each frame is added with six adjacent frames (3 before and 3 after). The scale error due to variation of $\hat{x}(0)$ for each frame is different from others. This will perceptually degrade the quality of reconstructed speech because the reconstructed speech is synthesized by overlapping and adding frames which have no scale compatibility.

[1] There are also some cases in which the Hilbert transform relations could be applied for mixed-phased signals [15]. However, speech signal is not of those kinds.

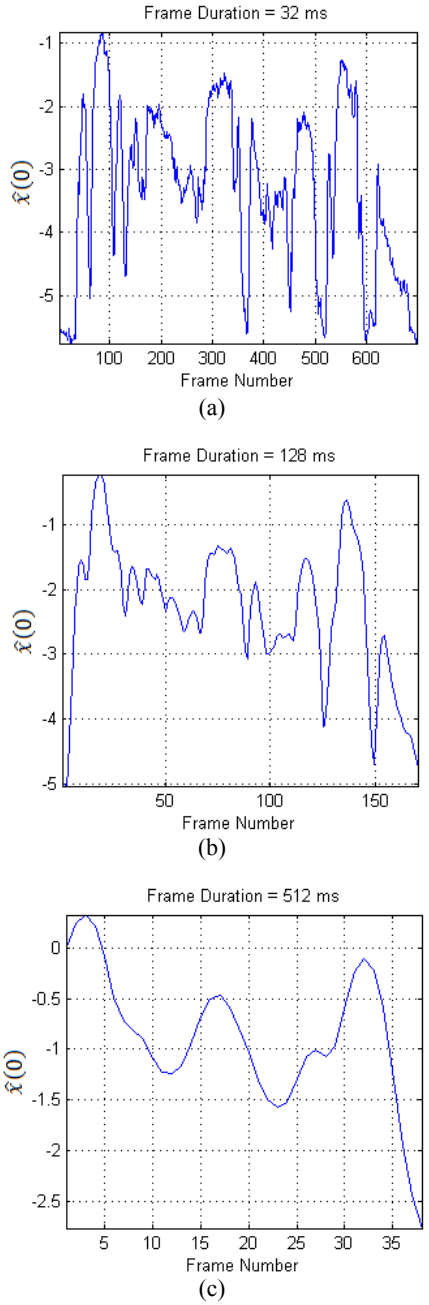

Figure 2*: variations of $\hat{x}(0)$ in different frames of a speech signal from NOIZEUS database (sp01) with different frame durations. (a) 32 ms, (b) 64 ms and (c) 512 ms.*

As seen in Fig.2 (a, b and c), by frame length extension, as the dynamics of $\hat{x}(0)$ is reduced, the scale error which is established by initializing the magnitude spectrum with 1 (Eq. (2)) decreases. For a better demonstration, the problem is better pursued quantitatively. We introduce the following measure of the scale error based on Eq. (2) and Eq. (9) to investigate this issue

$$Scale\ Error = \sum_m (1 - e^{\hat{x}(m,0)})^2. \quad (11)$$

Equation (9) implies that by initializing the magnitude spectrum with $e^{\hat{x}(m,0)}$ instead of 1 (Eq. 2), the scale incompatibility problem would be solved. As Fig. 3 shows, by frame length extension, the error which is introduced due to scale incompatibility is decreased

remarkably leading to speech quality improvement. We call this *scale error*. Finally, based on Eq. (9), we propose the following method for initializing the magnitude spectrum instead of Eq. (2)

$$X^1(m,\omega) = \exp\big(\hat{x}(m,0)\big).e^{j\angle X(m,\omega)}. \qquad (12)$$

Fig. 4 shows the results of initializing the magnitude spectrum of each frame with Eq. (12). Results show that the quality of phase-only reconstructed speech is higher than magnitude-only reconstructed speech over all frame lengths, even for frames as short as 16 and 32 ms.

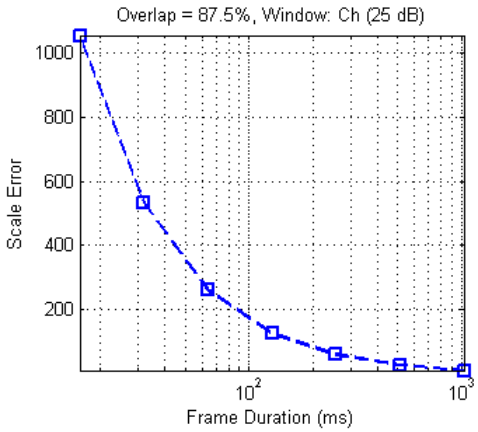

Figure 3: *Scale error (Eq. 11) versus frame duration (16, 32, 64, 128, 256, 512 and 1024 ms).*

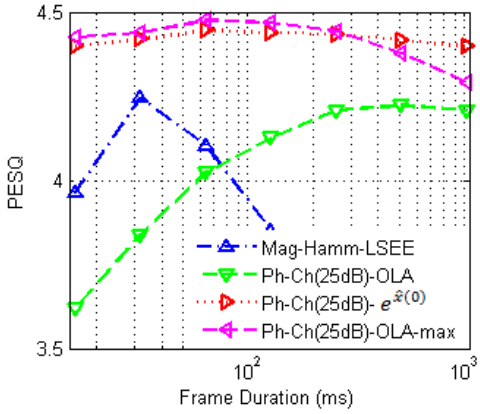

Figure 4: *Quality comparison of phase-only reconstructed speech based on proposed method with magnitude-only reconstructed speech versus frame duration (16, 32, 64, 128, 256, 512 and 1024 ms).*

It should be noted that the exact value of $\exp\big(\hat{x}(m,0)\big)$ is not crucial. If we replaced $\exp\big(\hat{x}(m,0)\big)$ with $\exp\big(\hat{x}(m,0)\big)/constant$ the quality of the reconstructed speech would be the same. Hence the important point is to maintain the proportion $\frac{\exp(\hat{x}(m,0))}{\exp(\hat{x}(p,0))}$ where *m* and *p* denote frames which have overlap. We can simply estimate $\exp\big(\hat{x}(m,0)\big)$ with $max_\omega\{|X(m,\omega)|\}$. This estimation almost maintains the aforementioned proportion and is justifiable considering the gradual changes in magnitude spectrum frame-by-frame. As Fig. 4 depicts, the results for phase-only reconstruction using $\exp\big(\hat{x}(m,0)\big)$ and the one using $max_\omega\{|X(m,\omega)|\}$ are almost the same.

4. CONCLUSION

In this paper we tried to answer a question which has been around since 1979, 'why the quality of phase-only reconstructed speech increases by frame length extension?' We showed that in phase-only speech reconstruction, the scales of the reconstructed frames which must be overlapped and added together are not compatible. This will perceptually decrease the quality of the synthesized speech. By frame length extension this problem is alleviated, so that the quality of the phase-only reconstructed speech improves. We proposed a method based on Hilbert transform to solve this problem. The results show remarkable quality improvement. The quality of phase-only reconstructed speech based on the proposed method is better than the magnitude-only reconstructed speech for all frame lengths and particularly for short frames. This work shows and proves the high potentials of the phase-based speech signal processing.

5. REFERENCES

[1] A. V. Oppenheim, J. S. Lim, G. E. Kopec, and S. C. Pohlig, "Phase in speech and pictures," in *Proc. IEEE Int. Con5 Acoust., Speech and SignalProcesring,* pp. 632-637, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[3] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *SpeechCoding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Marcel Dekker, 1991, ch. 4, pp. 165–172.

[4] A.V. Oppenheim, J.S. Lim, "The importance of phase in signals," Proc. IEEE 69 (1981) 529–541.

[5] 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.

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

[7] 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.

[8] "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.

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

[10] E. Loveimi and S.M. Ahadi, "Objective Evaluation of Magnitude and Phase Only Spectrum-based Reconstruction of the Speech Signal*",* in *Proc. Int. Symp.* On Communications, Control and Signal Processing (*ISCCSP2010*), Limassol, Cyprus, Mar. 2010.

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

[12] E.Loweimi and et. al., "On the Importance of Phase Spectrum in Speech Reconstruction", submitted to publish.

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

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

[15] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," IEEE Trans. Acoust., Speech, *Signal* Processing, vol. ASSP-28, pp. 672-680, Dec. 1980.