# Objective Evaluation of Magnitude and Phase Only Spectrum-based Reconstruction of the Speech Signal

Erfan Loveimi, *Student Member, IEEE* and Seyed Mohammad Ahadi, *Senior Member, IEEE*

***Abstract*—The aim of this paper is to investigate the effects of window shape and its length on the quality of phase-only and magnitude-only reconstructed speech. Speech signal is reconstructed via Least Square Error Estimation (LSEE) and Overlap Add (OLA) methods from its magnitude-only and phase-only spectra. The effects of using Hamming and rectangular windows have been studied, too. To evaluate the quality of reconstructed speech, we employed three objective measures: Log Likelihood Ratio (LLR), Weighted Spectral Slope (WSS) and PESQ. Results show that in case of using Hamming window, phase-only reconstructed speech surpasses its magnitude-only counterpart in frames longer than 256 ms qualitatively. Rectangular window seems to be a better choice in comparison to Hamming window in case of phase-only signal reconstruction, while for magnitude-only reconstructed speech, Hamming window results in better quality and intelligibility. LSEE has better performance than OLA in case of magnitude-only speech reconstruction while OLA outperforms LSEE in case of phase-only speech reconstruction.**

***Index Terms*—**Magnitude spectrum, phase spectrum, speech reconstruction, objective evaluation.

## I. INTRODUCTION

It is well established that the ear does not have preference among changes in the phase of sinusoidal signals or changes in the relative phase in the sinusoidal components of a signal [1]. This has gradually resulted in a common belief in speech processing that phase spectrum is not important. We can only observe the significance of phase in speech coding. In primary algorithms of speech coding [1], for lowering the bit rate, speech was coded from its magnitude spectrum. At the decoding stage, with minimum-phase assumption, phase was derived from magnitude and finally speech was reconstructed. Reconstructed speech did not have acceptable quality and intelligibility because the speech signal was not minimum phase intrinsically. As a result researchers tend to utilize phase in coding algorithms.

In other fields of speech processing, we can see no significant role for phase. For example, in spectral subtraction method [2], the enhancement process is focused on magnitude spectrum. At the end, the enhanced magnitude spectrum is used together with the noisy phase spectrum to reconstruct the speech signal. Similar story goes on in the field of speech recognition. Most of feature extraction methods such as MFCC or LPC only utilize magnitude spectrum information. From the signal processing viewpoint, both magnitude and phase spectra are needed to reconstruct the signal perfectly. However, in some situations magnitude-only reconstructed speech, along with random or zero phase, is intelligible to some extent. Is this reason enough to discard phase information and do not use them in speech recognition process? Is there any situation in which the role of phase becomes more important than the role of magnitude? In this paper, the aim is to find some answers to such these questions.

Oppenheim and Lim [3] conducted the first comprehensive study about the importance of phase in signal processing. In speech signal case, from the perception viewpoint, by performing some informal tests, the authors showed that if frame duration is chosen to be more than 1 sec, phase-only reconstructed speech will be intelligible.

Liu *et al* [4], by performing a series of subjective tests, came to more details about the importance of phase in speech recognition. They recorded 6 stop-consonants from 10 speakers in vowel-consonant-vowel context. Using these records, they created magnitude-only and phase-only stimuli. Phase-only stimuli were created by analyzing the original recordings through STFT, setting magnitude spectrum of each frame to unity. In case of magnitude-only stimuli, phase spectrum was replaced by uniformly distributed random numbers in the range of $(-\pi, \pi)$. At the next step, speech signals were reconstructed by OLA and finally they were played to subjects in order to recognize corresponding consonants. The stimuli were created for Hamming window and various frame durations from 16 to 512 ms with 50% overlap. Their results show that the intelligibility of phase-only reconstructed speech increases as the frame duration becomes longer while the opposite is true for magnitude-only reconstructed speech. In frame durations longer than about 128 ms, phase-only reconstructed speech surpasses its magnitude-only counterpart qualitatively.

Alsteris and Paliwal [5], [6] have carried out remarkable research in this field. Their framework is similar to that of Liu *et al* with some modifications. They chose the frame shift to be one eighth of frame length (= 87.5% overlap), numbers of consonants were 16 and rectangular window as well as Hamming window were applied. They showed that the phase spectrum, even in short length frames such as 32 ms, can contribute to speech intelligibility if the analysis window holds appropriate shape. In fact, rectangular (non-tapered) window is an excellent choice to reconstruct the speech signal from its phase-only spectrum even in short frames (e.g. 32 ms) and leads to results comparable to magnitude-only case. Hamming window is a suitable option to reconstruct the speech from its magnitude-only spectrum. For more detailed study of Alsteris and Paliwal experiments, readers are referred to [6].

Frameworks of both aforementioned tasks were based on Analysis-Modification-Synthesis (AMS). In section 2 of this paper, we will briefly cover this framework and review OLA and LSEE. In section 3, the objective measures which have been used to evaluate the quality of the reconstructed speech will be introduced. In section 4, we will discuss the conducted experiments, their results and analysis of them.

Manuscript received November 9, 2009. This work was supported in part by the Iran Telecommunication Research Center (ITRC) under Grant T-500-21804.

The authors are with Speech Processing Research Laboratory, Electrical Engineering Department, Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran. (phone: +982164543392; e-mail: {eloveimi, sma}@aut.ac.ir).

## II. ANALYSIS-MODIFICATION-SYNTHESIS FRAMEWORK

Speech is a non-stationary signal, so it should be analyzed in a frame-wise manner. In this case, short-time Fourier transform can be used to analyze speech signal assuming that it is quasi-stationary. Let *x(n)* be a given speech sequence and *X<sub>n</sub>(ω)* its short-time Fourier transform (STFT) after applying an analysis window *w(n)* on the speech signal *x(n)* [7]:

$$X_n(\omega) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\omega m}. \qquad (1)$$

$X_n(\omega)$ is a complex quantity, so it can be decomposed as follows:

$$X_n(\omega) = |X_n(\omega)|e^{j\psi_n(\omega)}, \qquad (2)$$

where *n* represents the index of the short time over which the Fourier transform is evaluated, $|X_n(\omega)|$ is the short-time magnitude spectrum and $\psi_n(\omega)$ is the short-time phase spectrum[1].

In modification step, a specific change is made in the spectrum of signal to check the effect and importance of certain parameter(s). For example, the importance of phase can be studied by reconstructing the signal from its phase-only spectrum. The same is true for magnitude.

To reconstruct the speech from its magnitude spectrum we chose a sequence of random uniformly distributed numbers in the range of (-π, π), $\varphi$, as the phase sequence:

$$\hat{X}_n^1(\omega) = |X_n(\omega)|e^{j\varphi}, \qquad (3)$$

where $\hat{X}_n(\omega)$ is the modified speech spectrum and *n* denotes STFT index. Superscript 1 indicates to the initialization of the iterative signal reconstruction algorithms which will be discussed in the next lines.

However, there is another alternative: one can substitute phase spectrum with zero. In the meantime, for preserving compatibility with former researches [4]-[6], random phase sequence has been used here.

To reconstruct speech only from its phase spectrum, typically, magnitude spectrum is set to unity:

$$\hat{X}_n^1(\omega) = e^{j\psi_n(\omega)}. \qquad (4)$$

In the next step we should synthesize the signal from its modified spectrum. Two methods have been employed here: Overlap Add Method (OLA) and Least Square Error Estimation (LSEE). OLA is a well-known method to reconstruct the signal from its short time spectrum:

$$x^{i+1}(n) = \frac{\sum_{p=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}_{pL}^i(\omega)e^{j\omega n}d\omega}{\sum_{p=-\infty}^{\infty} \mathrm{w}\,(pL-n)}, \qquad (5)$$

where *L* is the decimation factor and *w(n)* is the window applied in analysis stage. Note that $\hat{X}_{pL}^i(\omega) = |X_{pL}(\omega)|\angle\hat{X}_{pL}^i(\omega)$ in case of magnitude-only and $\hat{X}_p^i(\omega) = |X_{pL}^i(\omega)|\angle X_{pL}(\omega)$ in case of phase-only speech signal reconstruction. Previous works [4]-[6] utilized OLA.

LSEE is another alternative of OLA to reconstruct the signal from its modified spectrum [8]. It is written as:

$$x^{i+1}(n) = \frac{\sum_{p=-\infty}^{\infty} w(pL-n)\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}_{pL}^i(\omega)e^{j\omega n}d\omega}{\sum_{p=-\infty}^{\infty} w^2\,(pL-n)}, \qquad (6)$$

which is derived by minimizing the following distance measure (*D*) between $X_{Rec}$ (reconstructed speech spectrum) and $X_{Mod}$ (modified speech spectrum):

$$D[X_{Mod}(pL,\omega), X_{Rec}(pL,\omega)] = \sum_{p=-\infty}^{\infty} \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} |X_{Mod}(pL,\omega) - X_{Rec}(pL,\omega)|^2 \,\mathrm{d}\omega. \qquad (7)$$

In fact, after modifying the spectrum, it would not necessarily result in a valid STFT. LSEE tries to find a signal with the most similar spectrum to the modified spectrum in sense of MSE.

There are some points about the AMS framework that should be addressed here:

(1) *Analysis window type*. We used Rectangular and Hamming windows here.
(2) *Frame duration*. Frame durations of 32, 64, 128, 256, 512 and 1024 ms have been investigated.
(3) *Overlap*. In order to minimize the aliasing during the reconstruction process, overlap must be at least 75% in case of Hamming window [9]. To be on the safer side, it is set to 87.5%. Although the rectangular window can be used with a lower overlap, we used the same overlap for consistency.

We implemented both OLA and LSEE and compared their performances. The algorithm which is used to reconstruct the signal from its modified spectrum is in accordance with what has been proposed in [8].

## III. OBJECTIVE MEASURES

To evaluate the quality of speech signals, although the ideal solution is the utilization of systematic subjective tests on a large population, unfortunately usually in practice the available population is restricted. This leads to highly variable and sometimes ambiguous results. In addition, the whole process may take too long to be practical. Therefore some mathematical objective quality measures are necessary. Ideally, these measures should be as highly correlated as possible to subjective measures. In other words, we should look for the measures to be good predictors of average subjective preferences.

### *A. Log Likelihood Ratio (LLR)*

LLR is a LPC-based objective method with the following formula [10]:

$$d_{LLR}(\vec{a}_p, \vec{a}_c) = log\left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T}\right), \qquad (8)$$

where $\vec{a}_p$ and $\vec{a}_c$ are LPC vectors of the reconstructed and the original speech frames and $R_c$ is the autocorrelation matrix of the original speech. LLR is first calculated on each frame and then averaged on all of them. In order to reduce the number of outliers, segmental LLR was limited to the range of (0,2). Final LLR value is the average over all frames. The lower the LLR value, the better the quality of reconstructed speech.

### *B. Weighted Spectral Slope (WSS)*

In this method, weighted distance between spectral slopes of the original and enhanced speech is calculated [11]:

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m) - S_p(j,m))^2}{\sum_{j=1}^{K} W(j,m)}, \qquad (9)$$

[1] From here on, the modifier ‘short-time’ is implied wherever mentioning the phase spectrum and magnitude spectrum.

where $S_p(j,m)$ and $S_c(j,m)$ are the spectral slopes of the $m$th frame of the reconstructed and original speeches in the $j$th bandwidth, respectively. $K$ is the number of bandwidths and $M$ is the number of frames. $K$ is typically chosen equal to 13 or 25. In our implementation, $K$ was set to 25. The lower the WSS value, the better the quality of reconstructed speech.

### *C. Perceptual Evaluation of Speech Quality (PESQ)*

This algorithm was proposed by ITU-T to evaluate the quality of speech in telephone handsets and narrowband speech codecs [12]. It represents an aggregation of PAMS and PSQM99. These two algorithms were the highest performing methods in ITU-T competition that was held to find a more robust objective speech quality measure. PESQ was proposed to predict the MOS (Mean Opinion Score) subjective test and therefore, because of high correlation with subjective tests, it can be used to evaluate the quality of the reconstructed speech. Among all of the objective measures PESQ has the highest computational cost and complexity. PESQ produces a score between 1.0 and 4.5, with high values indicating better quality.

## IV. EXPRIMENTS, SIMULATIONS RESULTS AND ANALYSIS

### *A. Experimental conditions*

Experiments were conducted on the clean version of 30 speeches of NOIZEUS database [13]. This database is composed of gender and phonetically balanced utterances. The speeches were originally sampled at 25 kHz and downsampled to 8 kHz with a precision of 16 bits per sample. The signals were reconstructed from Magnitude-only and phase-only spectrums in accordance to aforementioned methods. Frame lengths were changed in order of 32, 64, 128, 256 and 512 and 1024. Overlap was set to 87.5%. FFT size was chosen to be $2N$ where $N$ is the number of samples of each frame. After taking IFFT the first $N$ samples were retained and second $N$ points were discarded. Specified overlap value and FFT size were chosen to reduce aliasing effects [5], [6].

### *B. Simulation Results and Analysis*

Figs. 1-3 depict LLR, WSS and PESQ values after phase-only and magnitude-only reconstruction of speech using Hamming window. We can make the following observations from these three figures:

1. LLR measure claims that the quality of reconstructed speech is degraded with frame extension and quality of magnitude-only reconstructed speech is higher than its phase-only counterpart while informal subjective tests show different results, especially in frames longer than 128 ms. It seems that LLR is not a suitable and fair objective measure in long frames. In fact the LPC vector of the frame is not valid anymore if frame length exceeds a specific range, where the stationarity of the segment become questionable. As a result, it could be considered as a relatively reliable objective measure only in short frame lengths up to 64 ms.

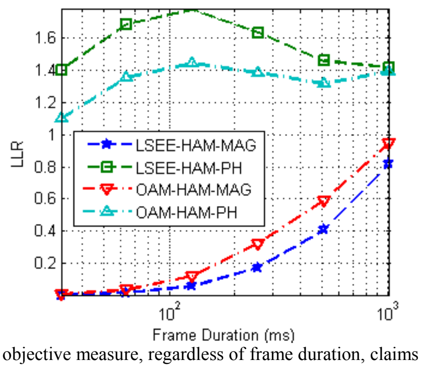

Fig. 1. LLR objective measure, regardless of frame duration, claims that the magnitude-only reconstructed speech has better quality than the phase-only one while subjective measures reject these results.

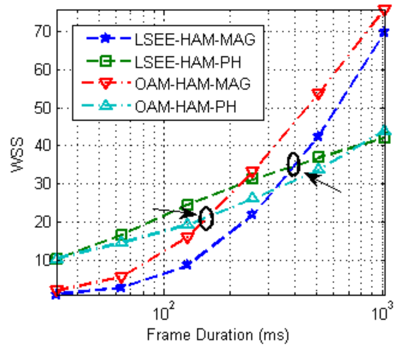

Fig. 2. WSS shows that in case of OLA, the quality of phase-only reconstructed speech surpasses magnitude-only reconstructed speech quality in frame length longer than 128 ms.

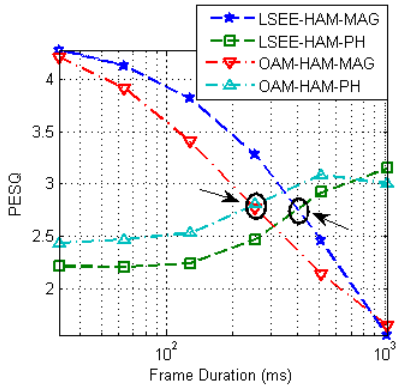

Fig. 3. PESQ has maximum correlation with subjective tests. It shows that for frame lengths longer than 256 ms phase-only reconstructed speech surpasses its magnitude-only counterpart qualitatively.

2. WSS and PESQ show more correlation with informal subjective tests. In both cases, magnitude-only reconstructed speech has better quality in shorter frame lengths such as 32 and 64 ms while phase-only reconstructed speech shows better quality in frame lengths more than 256 ms. Black circles denote the frame length in which phase-only reconstructed speech surpasses its magnitude-only counterpart, qualitatively. However, there is some difference between their assessment in the vicinity of frame duration of 256 ms. Informal subjective tests have more similarity to PESQ results. Generally, in such cases, PESQ takes precedence over other measures because of its highest correlation with subjective tests [14].

3. LSEE in comparison to OLA, in case of magnitude-only speech reconstruction, leads to better performance and results. This was expected from theoretical viewpoint [8]. However, in case of phase-only signal reconstruction, OLA shows better performance. This is why phase-only reconstructed speech signal via OLA surpasses its magnitude-only counterpart in shorter frame duration, qualitatively. It should be mentioned that the results do not show significant difference between these two methods, so they can be used alternatively.

Regarding to the type of window, according to what has been reported in [5] and [6], rectangular window should improve the quality of phase-only reconstructed speech and decrease the quality of magnitude-only reconstructed speech. Figs. 4-6 show the effect of employing rectangular window. Fig. 4 shows that magnitude-only reconstructed speech has better quality in all frame durations. In Figure 5, WSS objective measure claims that phase-only reconstructed speech has better quality in all frame durations. It seems that LLR and WSS are not fair measures. LLR tends to show that magnitude-only reconstructed speech has better quality than its phase-only counterpart and WSS tries to show the opposite result. In this case, we use PESQ as the referee [14]. It is interesting to note that PESQ shows that in frames longer than 128 ms phase-only reconstructed speech surpasses its rival. The second important result is that applying rectangular window increases the quality of phase-only reconstructed speech. It seems that for reconstructing the signal from its magnitude-only spectrum a compromise between both frequency resolution and leakage is required. However, in case of phase-only reconstructed signal, frequency leakage plays a more important role in comparison to magnitude-only case. Hence the best compromise is achieved in case of rectangular window.

## V. CONCLUSION

In this paper, we investigated the capability of phase and magnitude spectra to reconstruct the speech in different situations i.e. different frame lengths and windows. We employed LSEE and OLA to reconstruct the speech signals. The quality of reconstructed speech was evaluated by LLR, WSS and PESQ objective measures. The results showed that in case of applying Hamming window, in frames longer than about 256 ms, phase-only reconstructed speech has better quality. Rectangular window improves the quality of phase-only reconstruction such that it surpasses magnitude-only reconstructed speech in the vicinity of 128 ms, qualitatively. In case of magnitude-only speech reconstruction LSEE has better performance while in case of phase-only speech reconstruction OLA shows better results. However, the difference is not very significant so they can be used alternatively.

## REFERENCES

[1] R. J. McAulay and T. F. Quatieri, *Sinusoidal coding, in Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 121–173.

[2] S. F. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction,” *IEEE Trans. on ASSP*, vol. 27, *No.2, pp.113-120*, 1979.

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

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

[5] K.K. Paliwal, L.D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” *Speech Commun.* 45 (2) (2005) 153–170.

[6] L. D. Alsteris and K. K. Paliwal, “Short-time phase spectrum in speech processing: A review and some experimental results,” *Digital signal processing*, vol. 17, pp. 578–616, May 2007.

[7] Rabiner, L. and Juang, B.H. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[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] J.B. Allen and L.R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis” *Proc. IEEE*, Vol. 65, No. 11, pp. 1558-1564, 1977.

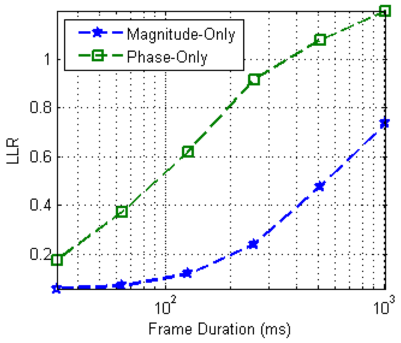

Fig. 4. LLR shows that even in case of using rectangular window and regardless of frame length, magnitude-only reconstructed speech has better quality than its phase-only counterpart. Subjective tests are not compatible with these results.

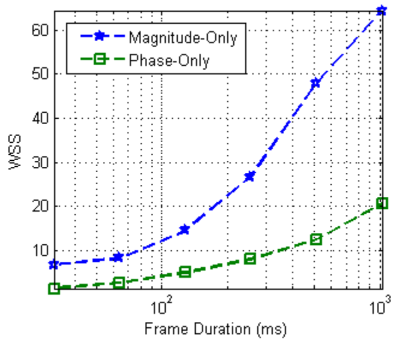

Fig. 5. WSS shows that phase-only reconstructed speech outperforms its magnitude-only counterpart qualitatively at all frame durations. It has acceptable correlation with informal subjective tests only in frames longer than about 128 ms.

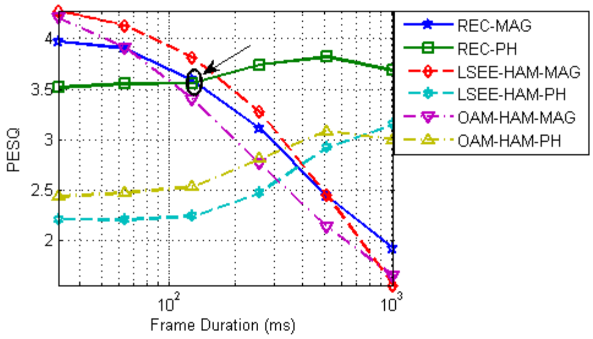

Fig. 6. PESQ shows interesting results with high correlation to informal subjective tests. Black circle indicate to the frame length in which phase-only reconstructed speech surpasses its magnitude-only counterpart qualitatively.

[10] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[11] D. Klatt, “Prediction of perceived phonetic distance from critical band spectra,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1982, vol. 7, pp. 1278–1281.

[12] “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.

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

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