Role of Phase Estimation in Speech Enhancement

Benjamin J. Shannon and Kuldip K. Paliwal

School of Engineering, Griffith University
Brisbane, QLD 4111, Australia
Ben.Shannon@student.griffith.edu.au, K.Paliwal@griffith.edu.au

Abstract

Typical speech enhancement algorithms that operate in the Fourier domain only modify the magnitude component. It is commonly understood that the phase component is perceptually unimportant, and thus, it is passed directly to the output.

In recent intelligibility experiments, it has been reported that the Short-Time Fourier Transform (STFT) phase spectrum can provide significant intelligibility when estimated using a window function lower in dynamic range than the typical Hamming window. Motivated by this, we investigate the role of the window function for STFT phase estimation in relation to speech enhancement.

Using a modified STFT Analysis-Modification-Synthesis (AMS) framework, we show that noise reduction can be achieved by modifying the window function used to estimate the STFT phase spectra. We demonstrate this through spectrogram plots and results from two objective speech quality measures.

Index Terms: speech enhancement, phase, windowing.

1. Introduction

Typical speech enhancement algorithms, such as Spectral Subtraction (SS) [1] [2] or the Ephraim-Malah algorithm [3], use short-time Fourier analysis-modification-synthesis (AMS) framework for speech enhancement. They process the corrupt speech signal by modifying (or correcting) the spectral magnitude component only and leave the phase component unchanged. This is due to the fact that the phase component is traditionally considered perceptually unimportant and has been shown not to contribute much towards speech enhancement [4].

Recently in [5, 6, 7], the relative importance of the STFT magnitude and phase spectra, in relation to speech intelligibility, has been investigated. In these studies, an Analysis-Modification-Synthesis (AMS) framework was used to create two types of stimuli for human listening tests, called *magnitude-only* and *phase-only*. To create the *magnitude-only* stimuli, the STFT phase spectrum was set to random values and the corresponding magnitude spectrum was left unmodified. A similar procedure was used to create the *phase-only* stimuli, but in this case, all of the detail in the STFT magnitude spectrum was removed by setting each magnitude component to one and leaving the phase spectrum unmodified. Sets of both types of test speech were then created using analysis frame sizes ranging from 16 ms to 512 ms. Subsequently, listeners were asked to identify the content of the artificial stimuli, which gave an identification rate score.

The experiments performed with "phase-only" stimuli by Lui *et al* [5] and Paliwal and Alsteris [6] [7] resulted in conflicting conclusions. In Lui's experiments, a Hamming window function was used in the estimation of the Fourier phase spectrum. This resulted in the "phase-only" stimuli having low intelligibility at short window lengths (window duration of about 32 ms). In Paliwal and Alsteris's experiments, a Rectangular window function was used in place of the Hamming window in Lui's experiments. As a result, significantly higher intelligibility was reported for the "phase-only" stimuli at short window lengths. It has been suggested that while the Hamming-type of window functions find application in magnitude spectrum estimation, their use for phase spectrum estimation results in bad distortion.

In the aforementioned research, estimation of STFT phase spectrum was investigated to evaluate its relative contribution to intelligibility compared to the magnitude spectrum. The two main conclusions from these works seem to be, 1) the Fourier phase spectrum contributes significantly to intelligibility when long analysis frames are used and 2) significant intelligibility is also observed from Fourier phase when a Rectangular window function is used during its estimation.

Earlier, Wang and Lim [4] have investigated the importance of the Fourier phase spectrum specifically in the context of speech enhancement. They used an oracle-type of experiment where the clean speech as well as the corresponding noisy speech was available for processing. In their investigation, a modified AMS framework was employed that consisted of two analysis blocks and one synthesis block. The role of one of the analysis blocks was to estimate the STFT magnitude spectra, while the other was used for the phase spectra. This modification allowed the signal-to-noise ratio (SNR) of the degraded speech feeding into each analysis block to be controlled independently. Hence a range of listening stimuli was created from different combinations of magnitude and phase spectra derived from varying quality speech. Listening tests were then performed using both the artificial stimuli and unmodified degraded speech. Subjects were asked to score the quality of the artificial speech by matching it with degraded speech they considered to be of equivalent quality. The SNR of the unmodified speech was then defined as the equivalent SNR score.

The results of Wang and Lim's study [4] were consistent with previous reports [8]. At relatively small window durations (approx. 50 ms), no significant improvement in equivalent speech quality was observed when the magnitude component was matched with a phase component computed from speech with a higher SNR. Likewise, for analysis windows of the order of 400 ms, a significant improvement in equivalent SNR was noted. These observations are also reflected in the aforementioned intelligibility tests reported by Lui *et al*. An important aspect of these experiments though, is the same window was used to estimate both the magnitude and phase spectra.

From the literature, there is a consistent relationship between window functions and the resulting phase spectrum. In

the experiments where a smooth window function (Hamming, Hanning) was used to estimate the STFT phase spectrum, the estimate contributed little to either the intelligibility or speech quality at short window lengths (approx. 32 ms) compared to the STFT magnitude spectrum. In the intelligibility testing, where a Rectangular window was used, a significant improvement was observed. Motivated by this observation, we apply a Rectangular window function to the STFT phase spectrum estimation problem for speech enhancement. We further expand the investigation and explore the role played by the dynamic range of the window function using a range of Chebyshev windows.

2. Evaluation Framework

To evaluate the effect of using an alternative window function to estimate the STFT phase spectrum for speech enhancement applications, we used a modified STFT Analysis-Modification-Synthesis (AMS) framework similar to the one proposed by Wang and Lim [4]. We conduct here an oracle-type experiment where we assume that the clean speech as well as the corresponding noisy (or degraded) speech is available to us for processing. A block diagram is shown in Fig.1 for reference.

Processing begins with an analysis stage in both the magnitude and phase spectra estimation branches. In this stage, the speech signal is decomposed into short-time overlapping frames. For this work, we have chosen the frame size to be 32 ms and the frame shift to be one eighth of a frame, which is 4 ms. Here we use m to denote the index of the frame.

Following the frame-blocking step, a window function is applied to the frames in each branch. In the magnitude spectrum estimation branch, a Hamming window is used, while in the phase spectrum estimation branch, $w_a(n)$ can be a Hamming, Rectangular or a Chebyshev window. After the frames have been windowed, we compute a STFT for each.

Before computing the Fourier transform of each length N frame, N zeros are appended. This step minimises the potential for time domain aliasing during resynthesis. At each frame index m, four spectra are produced from the two windowed time frames. These include a magnitude ($|X_a|$) and phase (ϕ_a) spectrum from the frame windowed with $w_a(n)$ and a magnitude ($|X_b|$) and phase (ϕ_b) spectrum from the frame windowed with a Hamming window.

Using the magnitude spectrum $|X_b|$ and the phase spectrum ϕ_a corresponding to frame index m, we construct an artificial STFT spectrum $\widehat{X}$ as shown in (1). An inverse STFT is then computed for each complex artificial frame resulting in a new real valued time domain frame. The new speech signal is then synthesised by applying the overlap-add algorithm.

$$\widehat{X} = |X_b|\mathrm{e}^{j\phi_a} \tag{1}$$

3. Experiments

In these evaluations, our aim was to investigate the effect of using different window functions to estimate the STFT phase spectrum using an AMS based speech enhancement systems. To do this, we performed experiments using six different window functions to estimate the phase spectra. The first of these windows was a Rectangular window. In addition to the Rectangular window, we used four different Chebyshev windows that ranged in dynamic range from 10 dB to 40 dB in 10 dB steps [9] and a Hamming window.

To conduct the evaluations, we used speech from the NOIZEUS database. This database is composed of gender and phonetically balanced utterances [10]. The sampling rate of the database is 8 kHz, and includes speech that has been corrupted using the noise samples from the Aurora II database. The corrupt speech on the database includes four noise levels, 0 dB, 5 dB, 10 dB and 15 dB, and eight different noise sources, Airport, Babble, Car, Exhibition, Restaurant, Station, Street and Train. Out of the samples provided, we used only the clean samples. In addition to these, a set containing artificial Gaussian white noise was also created.

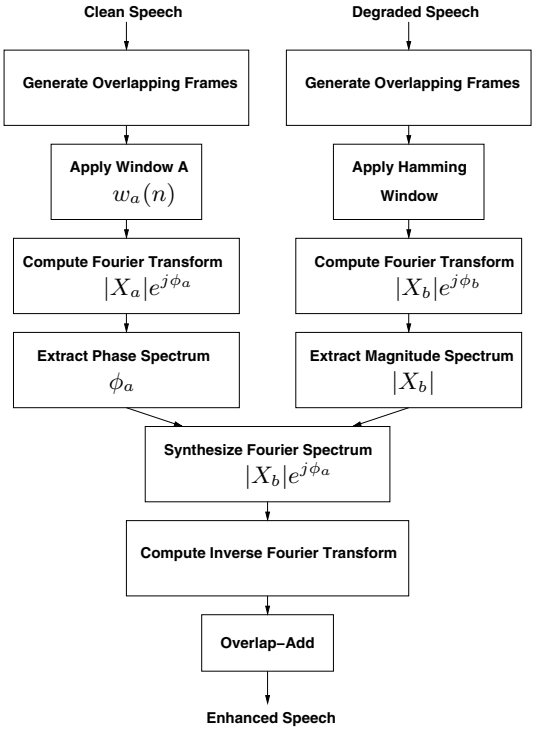

Figure 1: Block diagram of speech processing framework used to investigate the effect of alternative window functions on speech enhancement.

3.1. Spectrogram Analysis

In these experiments, we performed a spectrogram analysis of utterances that have been processed by the modified AMS system. Specifically, we took "sp01" from the NOIZEUS database [10] and added artificial Gaussian white noise so the resulting global SNR was 10 dB. We then processed the resulting utterance in conjunction with the clean version using the modified AMS system and a range of window functions for the phase spectrum estimator. Again, a Hamming window was used in all cases to estimate the magnitude spectra. The results from these evaluations can be seen in Fig.2.

3.2. Objective Speech Quality

In these experiments we estimated the quality of the processed speech using two objective speech quality measures, Perceptual Estimation of Speech Quality (PESQ) and Enhanced Modified Bark Spectral Distortion (EMBSD). The PESQ algorithm [11] represents an aggregation of two other techniques, PAMS and PSQM99. These two methods were the highest performing al-

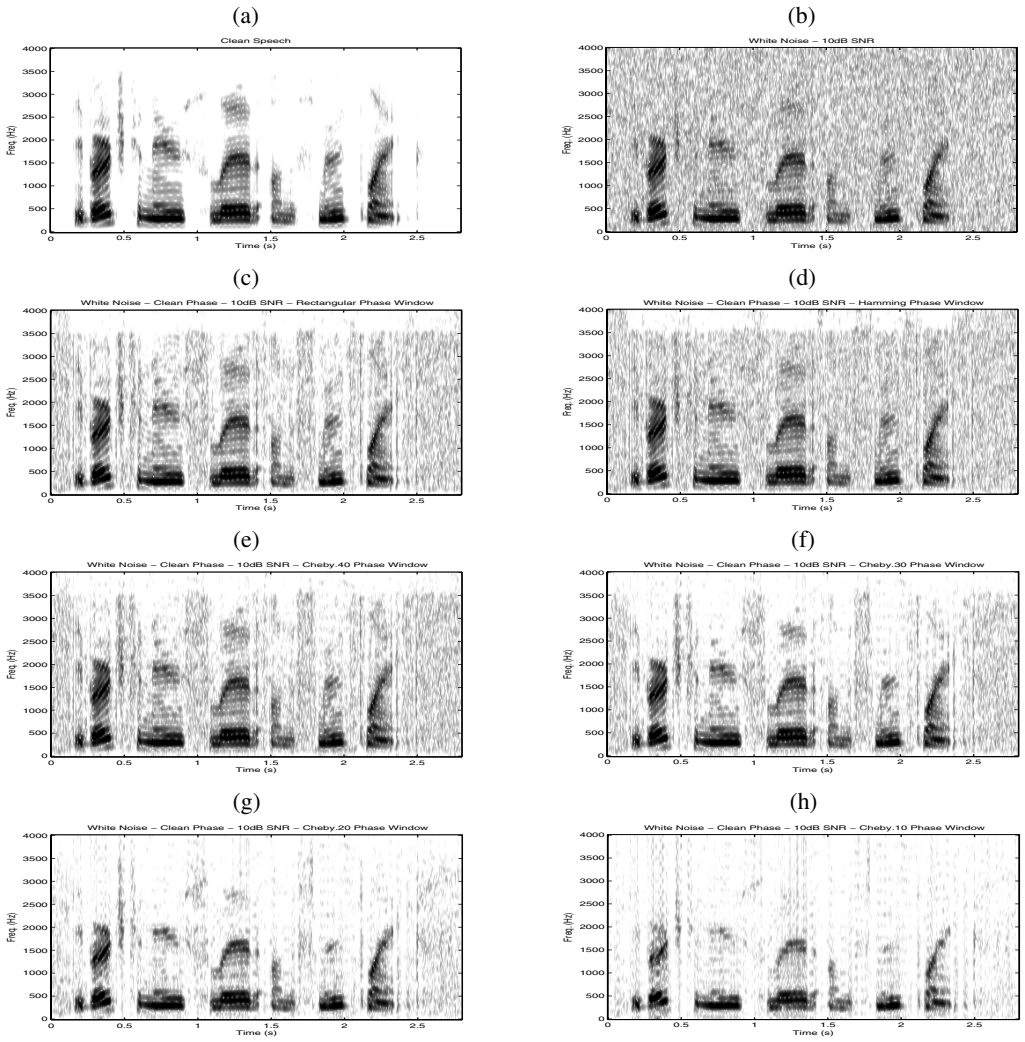

Figure 2: Spectrograms showing (a) clean speech (Noizeus - sp01) and (b) speech degraded with artificial Gaussian white noise at 10 dB SNR. Spectrograms are also shown for speech modified using the framework shown in Fig.1, where w_a is (c) Rectangular, (d) Hamming, (e) Cheby.40, (f) Cheby.30, (g) Cheby.20 and (h) Cheby.10.

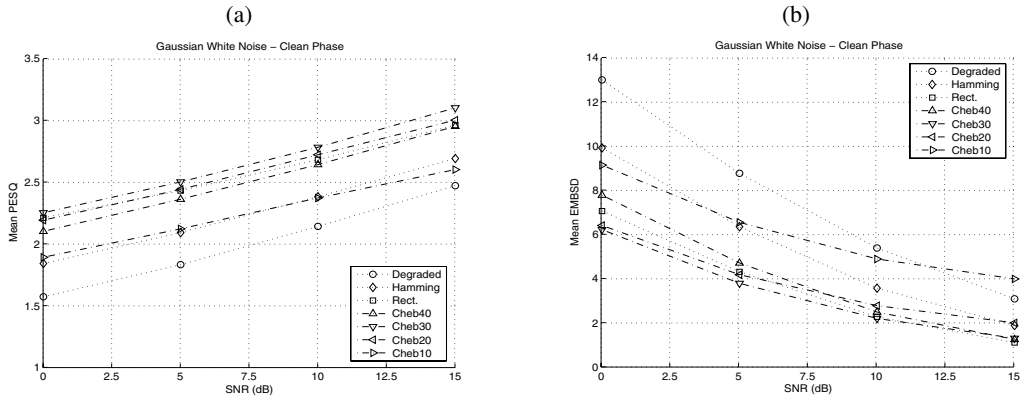

Figure 3: Objective speech quality results for Gaussian white noise. (a) Results using PESQ measure. (b) Results using EMBSD measure.

gorithms in an ITU-T competition that was held to find a more robust objective speech quality measure. The PESQ method attempts to overcome these limitations and produce more accurate scores in the presence of these disturbances. In our case, robust estimates of speech quality in the presence of background noise is of particular interest, since this is the primary source of corruption we are considering.

The EMBSD measure was developed by Yang and Yantorno [12, 13] by altering the Modified Bark Spectral Distortion (MBSD) measure, which itself is an extension of the Bark Spectral Distortion (BSD) measure. The modification made to the BSD algorithm to develop the MBSD algorithm was the consideration of noise masking. Through the application of a Noise Masking Threshold (NMT) level, an attempt is made to remove the perceptually insignificant disturbances from the speech quality estimate, thus improving performance over conventional BSD [12].

Both PESQ and EMBSD are perceptual speech quality measures. Any change in these scores from one test scenario to another is indicative of a quality difference perceivable by a human listener. Two things to keep in mind about these measures though are PESQ is an opinion score and EMBSD is a distortion score. The ramification of this is a higher score is better for PESQ, while a lower score is better for EMBSD.

For this evaluation, quality scores from the two objective measures were computed from all thirty utterances at 15, 10, 5 and 0 dB SNR. A summary score was then created for each SNR by computing the mean of the thirty individual utterance scores. The window functions tested included Rectangular and four Chebyshev windows ranging in dynamic range from 40 to 10 in 10 dB steps. We also tested the degraded speech as a reference. The results for the object speech quality experiments are shown in Fig.3. Plot (a) and (b) show the plots for PESQ and EMBSD respectively.

4. Discussion

From the spectrogram results, where a Hamming window has been used to estimate the phase spectra, a degree of noise reduction over the degraded speech can be observed. The level of noise reduction in this case appears to be the lowest out of the six tested scenarios. In the other cases, as the dynamic range of the window function used to estimate the phase spectra decreases, the amount of noise observable in the spectrogram also decreases. In Fig.2(h), where a Chebyshev window with 10 dB dynamic range has been used to estimate the phase spectra, the spectrogram looks similar to the clean speech spectrogram (Fig.2(a)).

The objective speech quality results shown in Fig.3, also indicate that an improvement in speech quality occurs when the STFT phase spectrum was estimated using clean speech and a Hamming window. In Wang and Lim's work [4], they reported an equivalent SNR improvement for a similar scenarios of 1 dB. Using our modified framework, we estimated an equivalent SNR improvement of 3.75 dB. In an informal experiment, we modified our framework parameters to match Wang and Lim's more closely; 52 ms frame size, 50% overlap and no zero padding. This resulted in an estimated equivalent SNR improvement of approximately 1 dB.

When a Chebyshev 30 window function was used to estimate the STFT phase in conjunction with clean speech, it resulted in an estimated equivalent SNR improvement of 9.75 dB. This figure is 6 dB better than the Hamming window estimate, which indicates that reducing the dynamic range of the window function used to estimate the STFT phase spectrum can improve the perceived speech quality.

5. Conclusions

We have proposed and investigated estimating the STFT phase spectrum independently from the STFT magnitude spectrum for speech enhancement applications.

In our experiments where we used a Hamming window to estimate both a STFT magnitude spectrum from degraded speech and a STFT phase estimate from clean speech, little improvement in speech quality could be observed. This set-up was similar to Wang and Lim's [4] and confirmed their findings. When we replaced the window function used to estimate the phase spectrum to one with a lower dynamic range, a substantial increase in noise reduction and speech quality could be observed. This effect could be seen in both spectrogram plots and measured using two perceptual domain objective speech quality measures.

6. References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Acoust., Speech, Signal Process.*, 1979, pp. 208–211.

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

[4] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 679–681, 1982.

[5] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[6] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, pp. 153–170, 2005.

[7] L. D. Alsteris and K. K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2004, pp. 573–576.

[8] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[9] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," in *Proc. of the IEEE*, vol. 66, no. 1, 1978, pp. 51–83.

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

[11] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[12] W. Yang and R. Yantorno, "Comparison of Two Objective Speech Quality Measures: MBSD and ITU-T Recommendation P.861," in *Proc. Second Annual IEEE Signal Processing Multimedia Conference*, 1998.

[13] W. Yang, "Enhanced Modified Bark Spectral Distortion," Ph.D. dissertation, Temple University, 1999.