Single channel speech enhancement using MMSE estimation of short-time modulation magnitude spectrum

Kuldip Paliwal, Belinda Schwerin, Kamil Wójcicki

Signal Processing Laboratory, School of Engineering,
Griffith University, Australia

k.paliwal@griffith.edu.au, belinda.schwerin@griffithuni.edu.au, kamil.wojcicki@gmail.com

Abstract

In this paper we investigate the enhancement of speech by applying MMSE short-time spectral magnitude estimation in the modulation domain. For this purpose, the traditional analysis-modification-synthesis framework is extended to include modulation domain processing. We compensate the noisy modulation spectrum for additive noise distortion by applying the MMSE short-time spectral magnitude estimation algorithm in the modulation domain. Subjective experiments were conducted to compare the quality of stimuli processed by the MMSE modulation magnitude estimator to those processed using the MMSE acoustic magnitude estimator and the modulation spectral subtraction method. The proposed method is shown to have better noise suppression than MMSE acoustic magnitude estimation, and improved speech quality compared to modulation domain spectral subtraction.

Index Terms: speech enhancement, MMSE short-time spectral magnitude estimator (AME), modulation spectrum, MMSE short-time modulation magnitude estimator (MME), modulation domain, analysis-modification-synthesis (AMS)

1. Introduction

Speech enhancement methods aim to improve the quality of noisy speech by reducing noise, while at the same time minimizing any speech distortion introduced by the enhancement process. Many enhancement methods are based on the short-time Fourier analysis-modification-synthesis framework. Some examples of these are the spectral subtraction method [1], the Wiener filter method [2], and the MMSE short-time spectral amplitude estimation method [3].

Spectral subtraction is perhaps one of the earliest and simplest methods, enhancing speech by subtracting a spectral estimate of noise from the noisy speech spectrum in either the magnitude or energy domain. Though this method is effective at reducing noise, it suffers from the problem of musical noise distortion, which is very annoying to listeners. To overcome this problem, Ephraim and Malah [3] proposed the MMSE short-time spectral amplitude estimator, referred to throughout this work as the acoustic magnitude estimator (AME). The AME method, even today, remains one of the most effective and popular methods for speech enhancement.

Recently, the modulation domain has become popular for speech processing. This has been in part due to the strong psychoacoustic and physiological evidence, which supports the significance of the modulation domain for the analysis of speech signals (a review of which can be found in [4]). At this point it is useful to differentiate the acoustic spectrum from the modulation spectrum as follows. The acoustic spectrum is the short-time Fourier transform (STFT) of the speech signal, while the modulation spectrum at a given acoustic frequency is the STFT of the time series of the acoustic spectral magnitudes at that frequency. The short-time modulation spectrum is thus a function of time, acoustic frequency (the axis of the first STFT of the input signal) and modulation frequency (the independent variable of the second STFT [5]).

Early efforts to utilize the modulation domain for speech enhancement assumed speech and noise to be stationary, and applied fixed filtering on the trajectories of the acoustic magnitude spectrum. However, speech and possibly noise are known to be nonstationary. To capture this nonstationarity, one option is to assume speech to be quasi-stationary, and process the trajectories of the acoustic magnitude spectrum on a short-time basis. This type of short-time processing in the modulation domain has been used in the past for automatic speech recognition (ASR) (*e.g.*, [6]), and also applied to objective speech quality evaluation (*e.g.*, [7]).

Short-time modulation domain processing was recently applied to speech enhancement in the modulation spectral subtraction method (ModSSub) of Paliwal *et al.* [8]. In the ModSSub method, the frame duration used for computing the short-time modulation spectrum was found to be an important parameter, providing a trade-off between quality and level of musical noise. Increasing the frame duration reduced musical noise, but introduced a slurring distortion. A somewhat long frame duration of 256 ms was recommended as a good compromise. The disadvantages of using longer modulation domain analysis window are as follows. Firstly, we are assuming stationarity which we know is not the case. Secondly, quite a long portion is needed for the initial estimation of noise, and thirdly, as shown by Paliwal *et al.* [9], speech quality and intelligibility is higher when the modulation magnitude spectrum is processed using short frame durations and lower when processed using longer frame durations. For these reasons, we aim to find a method better suited to the use of shorter modulation analysis window durations.

Since the AME method has been found to be more effective than spectral subtraction in the acoustic domain, in this paper, we explore the effectiveness of this method in the short-time modulation domain. The advantage of the MMSE-based method using decision-directed *a priori* SNR estimation is that it does not introduce musical noise and hence can be used with shorter frame durations in the modulation domain. The proposed approach, referred to as the modulation magnitude estimator (MME), is demonstrated to give better noise removal than the AME approach, without the musical noise of the spectral subtraction type approach, or the spectral smearing of the ModSSub method.

The rest of the paper is organised as follows. Secion 2 describes the proposed MME approach. Section 3 describes speech enhancement experiments and Section 4 evaluates the performance of the MME method. Conclusions are drawn in Section 5.

2. Minimum mean-square error short-time spectral modulation magnitude estimator

The minimum mean-square error short-time spectral amplitude estimator of Ephraim and Malah [3], here referred to as the MMSE acoustic magnitude estimator (AME), has been employed in the past for speech enhancement in the acoustic frequency domain with much success. In the present work we investigate its use in the short-time spectral modulation domain and denote the proposed approach the MMSE modulation magnitude estimator (MME).

2.1. AMS framework

AME is generally implemented by processing the short-time acoustic magnitude spectrum within a short-time Fourier AMS framework. This framework consists of three stages: 1) the analysis stage, where the noisy speech is processed using the STFT analysis; 2) the modification stage, where the noisy spectrum is compensated for noise distortion to produce the modified spectrum; and 3) the synthesis stage, where an inverse STFT operation is followed by overlap-add synthesis to reconstruct the enhanced signal.

Recently in [8], the traditional AMS framework has been extended to facilitate enhancement in the modulation domain. In [8], a secondary AMS procedure is used to framewise process the time series of each frequency component of the acoustic magnitude spectra. A block diagram of the AMS–based framework for speech enhancement in the short-time spectral modulation domain is shown in Fig. 1. In the present work, the modulation magnitude spectrum of clean speech is estimated from the noisy modulation magnitude spectrum, while the noisy modulation phase spectrum is left unchanged. This is justifiable based on the results of a recent study by Paliwal *et al.* [9] which suggests that for short MFDs ($\leqslant$64 ms), the modulation phase spectrum does not significantly contribute towards speech intelligibility or quality. The reconstructed acoustic magnitude spectrum is then combined with the noisy acoustic phase spectrum to construct the modified acoustic spectrum, which is finally used to reconstruct the enhanced speech signal.

2.2. Modulation MMSE Method

In the MME method, the modulation magnitude spectrum of clean speech is estimated from noisy observations. The proposed estimator minimises the mean-square error between the modulation magnitude spectra of clean and estimated speech

$$\epsilon = \mathrm{E}\left[\left(|\mathcal{S}_\ell(k,m)| - \left|\hat{\mathcal{S}}_\ell(k,m)\right|\right)^2\right] \tag{1}$$

where $\mathrm{E}\big[\cdot\big]$ denotes the expectation operator. Closed form solution to this problem in the acoustic spectral domain has been reported by Ephraim and Malah [3] under the assumptions that speech and noise are additive in time domain, and that their individual short-time spectral components are statistically independent, identically distributed, zero-mean Gaussian random variables. In the present work we make similar assumptions, namely that 1) speech and noise are additive in the short-time acoustic spectral magnitude domain (justification for this assumption is provided elsewhere [8]), and 2) the individual short-time spectral components of $\mathcal{S}_\ell(k,m)$ and $\mathcal{D}_\ell(k,m)$ are independent, identically distributed Gaussian random variables.

The reasoning for the first assumption is that at high SNRs the phase spectrum remains largely unchanged by additive noise distortion [10]. The second assumption can be justified, since (for large frequency analysis lengths) the Fourier expansion coefficients can be modeled as statistically independent Gaussian random variables [3]. With the above assumptions in mind, the modulation magnitude spectrum of clean speech can be estimated from the noisy modulation spectrum under the MMSE criterion [3] as

$$\begin{aligned}\left|\hat{\mathcal{S}}_\ell(k,m)\right| &= \mathrm{E}\Big[\,|\mathcal{S}_\ell(k,m)|\,\Big|\,\mathcal{X}_\ell(k,m)\Big] &(2)\\ &= \mathcal{G}_\ell(k,m)\,|\mathcal{X}_\ell(k,m)| &(3)\end{aligned}$$

where $\mathcal{G}_\ell(k,m)$ is the MMSE-MME spectral gain function given by

$$\mathcal{G}_\ell(k,m) = \frac{\sqrt{\pi}}{2}\,\frac{\sqrt{\nu_\ell(k,m)}}{\gamma_\ell(k,m)}\,\Lambda\Big[\nu_\ell(k,m)\Big] \tag{4}$$

in which $\nu_\ell(k,m)$ is defined as

$$\nu_\ell(k,m) \triangleq \frac{\xi_\ell(k,m)}{1+\xi_\ell(k,m)}\,\gamma_\ell(k,m) \tag{5}$$

and $\Lambda\big[\cdot\big]$ is the following function

$$\Lambda\big[\theta\big] = \exp\!\left(-\frac{\theta}{2}\right)\left[(1+\theta)\,I_0\!\left(\frac{\theta}{2}\right) + \theta\,I_1\!\left(\frac{\theta}{2}\right)\right] \tag{6}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. In the above equations $\xi_\ell(k,m)$ and $\gamma_\ell(k,m)$ are interpreted as the *a priori* signal-to-noise ratio (or the *a priori* SNR), and the *a posteriori* signal-to-noise ratio (or the *a posteriori* SNR) [11]. These quantities are defined as

$$\xi_\ell(k,m) \triangleq \frac{\mathrm{E}\big[|\mathcal{S}_\ell(k,m)|^2\big]}{\mathrm{E}\big[|\mathcal{D}_\ell(k,m)|^2\big]} \tag{7}$$

and

$$\gamma_\ell(k,m) \triangleq \frac{|\mathcal{X}_\ell(k,m)|^2}{\mathrm{E}\big[|\mathcal{D}_\ell(k,m)|^2\big]}. \tag{8}$$

Since in practice only noisy speech is observable, the $\xi_\ell(k,m)$ and $\gamma_\ell(k,m)$ parameters have to be estimated. For this task we apply the decision-directed approach [3] in the short-time spectral modulation domain. In the decision-directed method the *a priori* SNR is estimated by recursive averaging as follows

$$\hat{\xi}_\ell(k,m) = \alpha\,\frac{\left|\hat{\mathcal{S}}_{\ell-1}(k,m)\right|^2}{\hat{\lambda}_{\ell-1}(k,m)} + (1-\alpha)\,\max\Big[\hat{\gamma}_\ell(k,m)-1,\,0\Big] \tag{9}$$

where α controls the trade-off between noise reduction and transient distortion [12, 3], $\hat{\lambda}_\ell(k,m)$ is an estimate of $\lambda_\ell(k,m) \triangleq \mathrm{E}\Big[\left|\mathcal{D}_\ell(k,m)\right|^2\Big]$, and the *a posteriori* SNR estimate is obtained by

$$\hat{\gamma}_\ell(k,m) = \frac{|\mathcal{X}_\ell(k,m)|^2}{\hat{\lambda}_\ell(k,m)}. \tag{10}$$

Note that limiting the minimum value of the *a priori* SNR has a considerable effect on the nature of the residual noise [3, 12]. For this reason, a lower bound ξ_{min} is typically used to prevent *a priori* SNR estimates falling below its prescribed value, *i.e.*,

$$\hat{\xi}_\ell(k,m) = \max\Big[\hat{\xi}_\ell(k,m), \xi_{min}\Big]. \tag{11}$$

Spectral modulation domain estimates of the noise are also needed. Here, an initial estimate of the modulation power spectrum of noise is computed from leading silence frames. This estimate is then updated during speech absence using a recursive averaging rule (*e.g.*, [13]), applied in the modulation spectral domain as follows

$$\hat{\lambda}_\ell(k,m) = \varphi\,\hat{\lambda}_{\ell-1}(k,m) + (1-\varphi)\,|\mathcal{X}_\ell(k,m)|^2 \tag{12}$$

where φ is a forgetting factor chosen depending on the stationarity of the noise. The speech presence or absence is determined using a statistical model–based voice activity detection algorithm by [14], applied in the modulation spectral domain.

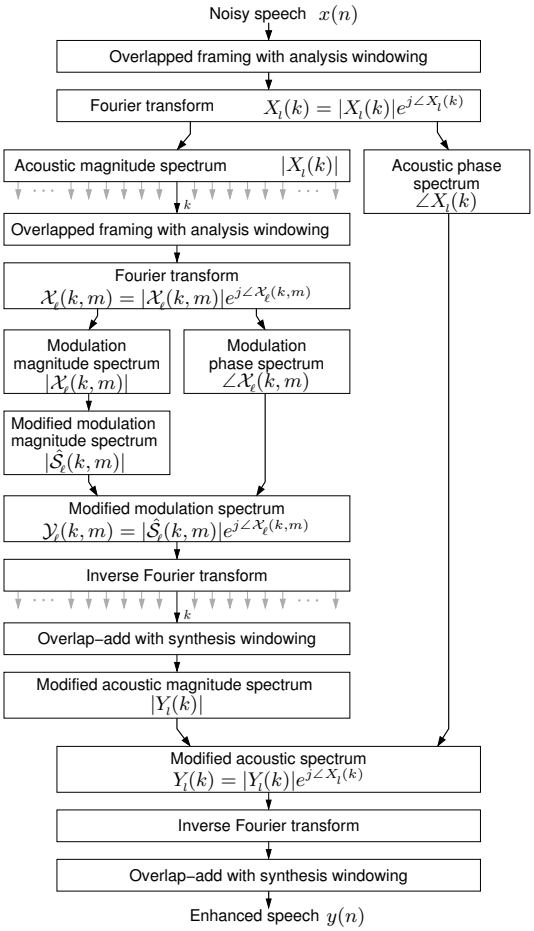

Figure 1: *Block diagram of the AMS–based framework for speech enhancement in the short-time spectral modulation domain.*

3. Experimental Procedure

3.1. Speech corpus

The Noizeus speech corpus [10] was used for the experiments presented in this work. The corpus contains 30 phonetically-balanced sentences belonging to six speakers (three males and three females). The recordings were sampled at 8 kHz and filtered to simulate the receiving frequency characteristics of telephone handsets. The corpus includes stimuli with non-stationary noises at different SNRs. For our experiments only the clean stimuli were used. Corresponding noisy stimuli were generated by degrading the clean stimuli with additive white Gaussian noise (AWGN) at 5 dB SNR. Since use of the entire corpus was not feasible for human listening tests, in our experiments four sentences—belonging to two male and two female speakers—were employed.

3.2. Parameters of the MME approach

One of the reasons for the good performance of the AME method of Ephraim and Malah [3] is that the parameters have been well tuned. In the current work this MMSE estimator is applied in the spectral modulation domain. Therefore, the parameters of the proposed MME method need to be retuned.

The adjustable parameters of the MME approach include the acoustic frame duration (AFD), acoustic frame shift (AFS), modulation frame duration (MFD), modulation frame shift (MFS), as well as the smoothing parameter α and the lower bound ξ_{min} used in *a priori* SNR estimation. Tuning of some of these parameters (such as AFD and AFS) can be done qualitatively from our knowledge of speech processing, and can be fixed without further investigation. For example, speech can be assumed to be approximately stationary over short durations, and therefore acoustic frameworks typically use a short AFD of around 10–40 ms (*e.g.*, [10]), which at the same time is long enough to provide reliable spectral estimates. Based on these qualitative reasons, an AFD of 32 ms was selected in this work. We have also chosen to use a 1 ms AFS to facilitate experimentation with a wide range of frame sizes and shifts in the modulation domain, and to increase the adaptability of the proposed method to changes in signal characteristics.

For other parameters, subjective listening tests were conducted to determine values that maximise the subjective quality of stimuli enhanced using the MME method. The tuned values for these parameters are a 32 ms MFD, 2 ms MFS, 0.998 for smoothing parameter α, and -25 dB for ξ_{min}. The MME stimuli produced with these settings were found to have good noise suppression without the introduction of musical noise or temporal slurring distortion.

3.3. Stimuli types

MME stimuli were constructed using the procedure detailed in Section 2, and the above parameters (see Section 3.2). The voice activity detector (VAD) threshold was set to 0.15, the forgetting factor φ for noise estimation was set to 0.98, the Hamming window was used for both the acoustic and modulation analysis windows, and the FFT-analysis length was set to $2N$ and $2\mathcal{N}$, respectively. Griffin and Lim's method for windowed overlap-add synthesis [15] was used for both acoustic and modulation domain syntheses.

Stimuli enhanced using the AME method [3] were generated using a publically available reference implementation [10]. Here, optimal estimates (in the minimum mean-square error sense) of the short-time acoustic spectral magnitudes were computed. The AMS procedure used a 20 ms AFD and a 10 ms AFS. The decision-directed approach was used for the *a priori* SNR estimation, with the smoothing factor set to 0.98, and the *a priori* SNR lower bound was set to -25 dB. Noise spectrum estimates were computed from non-speech frames using recursive averaging with speech presence or absence determined using a statistical VAD [14].

ModSSub stimuli were created used an AFD of 32 ms, with an 8 ms AFS, and MFD of 256 ms, and a 32 ms MFS. The spectral floor parameter β was set to 0.002, and γ was set to 2 for subtraction in the modulation magnitude-squared domain. Speech presence or absence was determined using a VAD algorithm based on segmental SNR. The speech presence threshold was set to 3 dB. The forgetting factor was set to 0.98.

In addition to the MME, AME, and ModSSub stimuli, clean and noisy speech stimuli were also included in our experiments. Example spectrograms for each of the stimuli types are shown in Fig. 3.

3.4. Listening test

Experiments in the form of AB listening tests were conducted to compare the subjective quality of stimuli described in Section 3.3. The actual test consisted of stimuli pairs played back in randomised order over closed circumaural headphones at a comfortable listening level, in a quiet room. For each stimuli pair, the listeners were presented with three labelled options on a computer and asked to make a subjective preference. The first and second options were used to indicate a preference for the corresponding stimuli, while the third option was used to indicate a similar preference for both stimuli. The listeners were instructed to use the

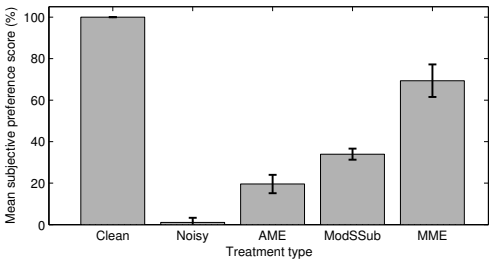

Figure 2: *Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated using the following treatment types: (c) AME [3]; (d) ModSSub [8]; and (e) MME (proposed).*

third option only when they did not prefer one stimulus over the other. Participants were familiarised with the task during a short practice session. Pairwise scoring was used, with a score of +1 awarded to the preferred treatment, and 0 to the other. For the similar preference response, each treatment was awarded a score of +0.5. Participants could re-listen to stimuli if required. Four Noizeus sentences belonging to two male and two female speakers and degraded with 5 dB AWGN were used for the experiment. The complete test included 80 comparisons and lasted approximately 15 minutes. Twelve listeners participated in the test.

4. Results and Discussion

The mean subjective preference scores for each enhancement method are shown in Fig. 2. MME scores are significantly higher than those of ModSSub and AME methods, indicating that listeners consider MME stimuli to have a higher quality than those of both AME and ModSSub types. This is an important finding as it demonstrates the efficiency of short-time modulation processing for speech enhancement, and demonstrates that the performance of existing acoustic domain approaches can be potentially improved when these are applied in the short-time modulation domain.

Spectrograms of the utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus are shown in Fig. 3. Spectrograms for clean and noisy (degraded with 5 dB AWGN) stimuli are shown in Fig. 3(a) and (b) respectively. Figure 3(c) shows the spectrogram of stimuli enhanced using AME, where much residual noise can be seen. The ModSSub stimuli, as shown by the spectrogram of Fig. 3(d), are much cleaner, but there is some spectral smearing and visible spectral artifacts. These distortions can be heard as a type of slurring and some low level musical-type noise. The spectrogram for the MME stimuli is shown in Fig. 3(e). As can be seen, MME stimuli have better noise suppression than AME without introducing the spectral artifacts visible in the ModSSub spectrogram. Informal listening confirms that speech quality of MME stimuli are improved without the introduction of the annoying residual distortions heard in the other stimuli types investigated.

5. Conclusions

In this paper we have proposed the MMSE modulation magnitude estimation method for speech enhancement. We have compared the performance of the proposed approach against that of two other methods, the MMSE acoustic magnitude estimator of Ephraim and Malah [3] and modulation spectral subtraction of Paliwal *et al.* [8]. The results of our experiments show that the proposed method achieves significantly higher subjective preference than the other two methods. This is because it uses a shorter modulation frame duration than modulation spectral subtraction and thus does not suffer from slurring distortion. In addition, it

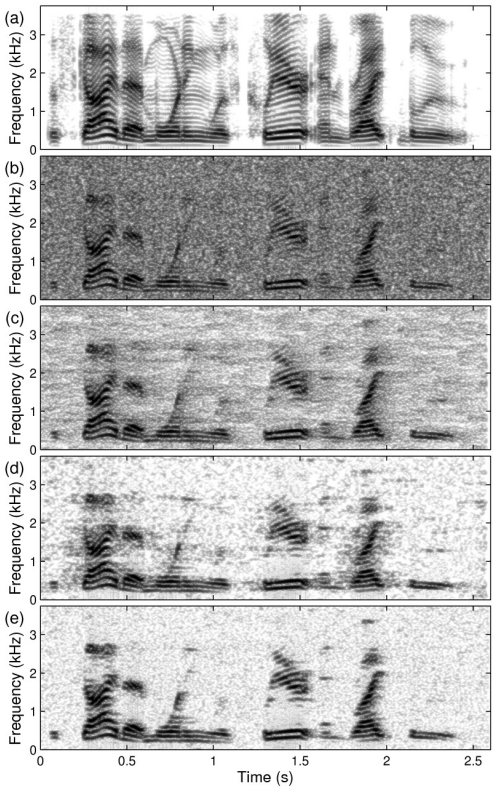

Figure 3: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) AME [3]; (d) ModSSub [8]; and (e) MME (proposed).*

achieves considerably better noise reduction than that associated with the MMSE acoustic magnitude estimator.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[2] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley, 1949.

[3] 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. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[4] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, Jan 2003.

[5] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, Montreal, Quebec, Canada, May 2004, pp. 761–764.

[6] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, Aug 1998.

[7] D. Kim, "Anique : An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821–831, Sep 2005.

[8] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.

[9] K. Paliwal, B. Schwerin, and K. Wójcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Communication*, vol. 53, no. 3, pp. 327–339, Mar 2011.

[10] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[11] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr 1980.

[12] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[13] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, Atlanta, Georgia, USA, May 1996, pp. 629–632.

[14] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan 1999.

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