TRADE-OFF EVALUATION FOR SPEECH ENHANCEMENT ALGORITHMS WITH RESPECT TO THE A PRIORI SNR ESTIMATION

Pei Chee Yong, Sven Nordholm, Hai Huyen Dam

Curtin University, Kent Street, Bentley, WA 6102, Australia
Peichee.Yong@postgrad.curtin.edu.au {S.Nordholm, H.Dam}@curtin.edu.au

ABSTRACT

In this paper, a modified *a priori* SNR estimator is proposed for speech enhancement. The well-known decision-directed (DD) approach is modified by matching each gain function with the noisy speech spectrum at current frame rather than the previous one. The proposed algorithm eliminates the speech transient distortion and reduces the impact from the choice of the gain function towards the level of smoothing in the SNR estimate. An objective evaluation metric is employed to measure the trade-off between musical noise, noise reduction and speech distortion. Performance is evaluated and compared between a modified sigmoid gain function, the state-of-the-art log-spectral amplitude estimator and the Wiener filter. Simulation results show that the modified DD approach performs better in terms of the trade-off evaluation.

***Index Terms*—** Speech enhancement, decision-directed approach, SNR estimation, musical noise, objective evaluation

1. INTRODUCTION

Single channel speech enhancement is widely used in mobile phones or listening devices. Among the vast amount of short-time spectral-domain noise reduction algorithms published in literature, the best known methods are the Wiener filter and the MMSE log-spectral amplitude (LSA) estimator [1]. The LSA approach is usually preferred against Wiener filtering as it can reduce unnatural artifacts known as *musical noise*.

A dominant point behind the reduction of musical noise by the LSA approach is the decision-directed (DD) approach for the *a priori* SNR estimation [2]. The DD approach performs a linear combination of two components: one being an estimate of previous *a priori* SNR and another being the maximum-likelihood (ML) SNR estimate. By applying a weighting factor closed to unity to the past *a priori* SNR estimate, the DD approach corresponds to a highly smoothed version of the *a posteriori* SNR, which reduces the musical noise. The drawback of reducing the variance in the *a priori* SNR estimate is that it cannot react quickly to abrupt changes in the instantaneous SNR. This bias leads to a performance degradation in speech enhancement schemes due to the speech transient distortion.

In addition to the speech transient distortion, a large degree of smoothing by the DD approach also yields two undesired effects: a reduced noise reduction and a reverberation effect [3]. Although the level of smoothing can be controlled in DD approach, it yields a trade-off between musical noise, noise reduction and speech distortion. Furthermore, such smoothing effect also depends highly on the choice of the gain function. For instance, not much smoothing was observed for Wiener filter compared to the LSA approach [3]. Accordingly, the LSA approach has less musical noise whilst the Wiener filter is reported to provide higher noise suppression.

In this paper, a modified *a priori* SNR estimator is proposed to reduce both the speech transient distortion and the effect of different gain functions towards the level of smoothing of the SNR estimate. To highlight this, a comparison is performed between the Wiener filter, the LSA approach and a sigmoid function. The sigmoid function has been developed as a spectral weighting gain function as it has several parameters that can be adjusted to achieve a balanced trade-off between noise reduction and speech distortion [4]. In this work, the sigmoid function is modified to map with the *a priori* SNR estimate. An analysis towards the trade-off between musical noise, noise reduction and speech distortion is performed by using an objective evaluation metric.

The remainder of this paper is organized as follows. Section 2 gives a system overview. Section 3 shows the proposed SNR estimate. Section 4 outlines the objective evaluation metric and Section 5 presents the results. Section 6 concludes the paper.

2. SYSTEM OVERVIEW

Let the noisy signal in discrete-time domain be expressed as $y(n) = x(n) + v(n)$, where $x(n)$ is the clean speech signal and $v(n)$ is the uncorrelated additive noise. By using the short-time Fourier transform (STFT), the spectral coefficients $Y(k,m)$ can be obtained by

$$Y(k,m) = \sum_{n=1}^{N} y(mR+n)\, w(n) \exp\left(\frac{-j2\pi kn}{N}\right) \quad (1)$$

where k is the frequency bin index, m is the frame index, R is the oversampling point and $w(n)$ is a window function. The clean speech spectrum estimate $\hat{X}(k,m)$ is then obtained by

$$\hat{X}(k,m) = G(k,m)Y(k,m) \quad (2)$$

where $G(k,m)$ is a non-linear gain function mapped with the *a priori* SNR $\xi(k,m)$ and/or the *a posteriori* SNR $\gamma(k,m)$, defined as

$$\xi(k,m) = \frac{E\left\{|X(k,m)|^2\right\}}{E\left\{|V(k,m)|^2\right\}} = \frac{\lambda_x(k,m)}{\lambda_v(k,m)} \quad (3)$$

$$\gamma(k,m) = \frac{|Y(k,m)|^2}{E\left\{|V(k,m)|^2\right\}} = \frac{|Y(k,m)|^2}{\lambda_v(k,m)} \quad (4)$$

where $\lambda_x(k,m)$ and $\lambda_v(k,m)$ denote clean speech power spectral density (PSD) and noise PSD, respectively.

The gain function can be derived from MMSE optimization criteria. One of those is the Wiener filter, which minimizes the expected value $\mathrm{E}\{|X(k,m) - \hat{X}(k,m)|^2\}$. Another widely used

This research was supported in part by Sensear Pty Ltd and Linkage Grant LP100100433 by Australian Research Council (ARC).

algorithm is the LSA estimator, which is obtained by minimizing $\mathrm{E}\{[\log(|X(k,m)|) - \log(|\hat{X}(k,m)|)]^2\}$ [1]. The resulting gain functions for the Wiener filter and the LSA approach are obtained respectively as

$$G_{\text{Wiener}}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)} \tag{5}$$

and

$$G_{\text{LSA}}(k,m) = \min\left(\varsigma, \frac{\xi(k,m)}{1+\xi(k,m)}\left\{\frac{1}{2}\int_{\nu(k,m)}^{\infty}\frac{e^{-t}}{t}dt\right\}\right) \tag{6}$$

where $\varsigma = 10$ denotes the upper bound for the LSA estimator and $\nu(k,m) = \gamma(k,m)\frac{\xi(k,m)}{1+\xi(k,m)}$.

As an alternative gain function to the MMSE approaches, a sigmoid function mapped with the *a posteriori* SNR has been proposed in [4]. Here, the sigmoid function is modified to map with the *a priori* SNR. The modification is done by multiplying the original logistic function in [4] with a hyperbolic tangent function, as

$$G_{\text{MSIG}}(k,m) = \begin{array}{l}\frac{1-\exp[-a_1\xi(k,m)]}{1+\exp[-a_1\xi(k,m)]} \\ \times\frac{1}{1+\exp(-a_2[\xi(k,m)-c])}.\end{array} \tag{7}$$

Finally, the enhanced speech signal $\hat{x}(n)$ is obtained by transforming $\hat{X}(k,m)$ back to the time domain using an inverse STFT.

3. A PRIORI SNR ESTIMATION

3.1. Traditional Decision-Directed Approach

Since the clean speech signal is practically unavailable, the *a priori* SNR from Eq. (3) has to be estimated. The most widely used method is the DD approach, given by [5]

$$\hat{\xi}_{\text{DD}}(k,m) = \beta\frac{|\hat{X}(k,m-1)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1] \tag{8}$$

where $\hat{\lambda}_v(k,m)$ and $\hat{X}(k,m-1)$ denote, respectively, the estimated noise PSD and the estimated clean speech spectrum from the preceding frame. In this work, the noise PSD is estimated by using the step-size controlled noise estimator in [6]. The parameter β denotes the smoothing factor and $P[.]$ denotes the half-wave rectification. The advantage of the DD approach is its capability to eliminate musical noise based on the choice of β in the conditional smoothing procedure [2].

3.2. Modified Decision-Directed Approach

The drawback of the traditional DD approach is the extra one-frame delay during speech transients, e.g. speech onsets and offsets, resulting in a degradation of speech quality. This is due to the fact that the *a priori* SNR estimate depends on the estimation of the clean speech spectrum in the previous frame. As a consequence, the gain function matches the previous frame instead of the current one. Thus, we propose to reduce the delay in speech transients by matching both estimates of the clean speech spectrum and the *a priori* SNR estimate with the current noisy speech spectrum. This is done by modifying the first term of the DD approach such that the gain function at previous frame is mapped with the current noisy speech spectrum. The modified approach is given by

$$\hat{\xi}_{\text{MDD}}(k,m) = \beta\frac{|G_{(.)}(k,m-1)Y(k,m)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1] \tag{9}$$

where $G_{(.)}$ denotes the gain function used in the speech enhancement scheme, such as the afore-mentioned LSA, Wiener or MSIG. Since the first term of Eq. (9) does not contain an estimate of the *a priori* SNR at previous frame, the modified approach can no longer represent a conditional first order recursive averaging algorithm as in Eq. (8). As such, it increases the sensitivity of the *a priori* SNR estimate towards the abrupt changes in speech signal, which directly reduces the speech transient distortion. However, such variance in the *a priori* SNR estimate can again lead to audible musical noise. In order to reduce, or eliminate the musical noise, the proposed method is smoothed by modifying the *a posteriori* SNR in Eq. (4) as [4]

$$\bar{\gamma}(k,m) = \frac{\lambda_y(k,m)}{\hat{\lambda}_v(k,m)} \tag{10}$$

where $\lambda_y(k,m) = \alpha_y\lambda_y(k,m-1) + (1-\alpha_y)|Y(k,m)|^2$ denotes the noisy speech PSD. The parameter $\alpha_y = \exp\left(\frac{-2.2R}{t_y f_s}\right)$ is the time averaging constant.

4. REPRESENTATIVE OBJECTIVE MEASURES

The performance of the speech enhancement scheme has a trade-off between musical noise, speech distortion and noise reduction. Any performance evaluation can be meaningless if it does not represent results from all of these trade-offs. Therefore, an objective evaluation metric is utilized to evaluate and compare the results between the amount of musical noise, speech distortion and noise reduction generated from the speech enhancement schemes.

First of all, the musical noise and the noise reduction should only be calculated during noise-only periods in short-time spectral domain. Since in practical situations the true noise PSD is often not known, a reference VAD is required for performance evaluation without the knowledge of noise characteristics. In order to obtain the VAD decisions at different frames and frequency bins, the multi decisions sub-band VAD (MDSVAD) is utilized [7]. Given two hypotheses, $\mathcal{H}_0\,(k,m)$ and $\mathcal{H}_1\,(k,m)$, which indicate speech absence and presence respectively in the k^{th} frequency bin of the m^{th} frame, the MDSVAD decisions are given by

$$D\,(k,m) = \begin{cases} 1 & \mathcal{H}_0\,(k,m) \\ 0 & \mathcal{H}_1\,(k,m)\,. \end{cases} \tag{11}$$

The amount of musical noise is believed to be highly correlated with the number of isolated spectral components and their level of isolation [8]. Since such components have relatively high power, they can be perceived as tonal sound that is strongly related to the weight of skirt of the probability density function (PDF). A signal with skirt can be identified using higher-order statistics, i.e. kurtosis. However, in order to identify only the musical-noise components, a kurtosis ratio (KurtR) is used to measure the change in kurtosis between the noisy signal and enhanced signal. In contrast to the approach in [8], which involved a musical noise assessment theory for spectral subtraction function, this measure is defined in this paper as

$$\text{kurtR} = \mathrm{E}\left\{\frac{\mathcal{K}_{\hat{x}}(k)}{\mathcal{K}_y(k)}\right\} \tag{12}$$

where $\mathcal{K}_{\hat{x}}(k)$ and $\mathcal{K}_y(k)$ denote the kurtosis of the enhanced signal and the noisy signal, respectively at k-th frequency bin. Both of them are computed only during speech absence periods, as given by

$$\mathcal{K}_{\hat{x}}(k) = \frac{\sum_{m=1}^{M}|\hat{X}\,(k,m)\,D(k,m)|^4}{\left\{\sum_{m=1}^{M}|\hat{X}\,(k,m)\,D(k,m)|^2\right\}^2} - 2. \tag{13}$$

$$\mathcal{K}_y(k) = \frac{\sum_{m=1}^{M} |Y(k,m)\,D(k,m)|^4}{\left\{\sum_{m=1}^{M} |Y(k,m)\,D(k,m)|^2\right\}^2} - 2. \qquad (14)$$

A smaller value of KurtR in Eq. (12) indicates less musical noise.

Meanwhile, the amount of noise reduction can be defined as the input noise power in dB minus the output noise power in dB. This noise reduction ratio (NRR) is defined during noise-only periods as

$$\text{NRR [dB]} = 10\log_{10} \frac{\sum_{m=1}^{M}\sum_{k=1}^{K} |Y(k,m)\,D(k,m)|^2}{\sum_{m=1}^{M}\sum_{k=1}^{K} |\hat{X}(k,m)\,D(k,m)|^2}. \qquad (15)$$

For speech distortion measure, the log-likelihood (LLR) measure is used. It is a spectral distance measure that models the mismatch between the formats of the clean and enhanced speech signals [9]. The LLR measure is defined as

$$d_{\text{LLR}}\left(\vec{l}_{\hat{x}}, \vec{l}_x\right) = \frac{\vec{l}_{\hat{x}} \mathbf{R}_x \vec{l}_{\hat{x}}^T}{\vec{l}_x \mathbf{R}_x \vec{l}_x^T} \qquad (16)$$

where $\vec{l}_x$ and $\vec{l}_{\hat{x}}$ are the linear predictive coding (LPC) coefficient vectors of the clean speech signal and the enhanced speech signal respectively, and R_s is the autocorrelation matrix of the clean speech signal. A lower LLR score indicates a better speech quality.

5. EXPERIMENTAL EVALUATION

Performance evaluation was done for each speech enhancement scheme with MSIG, LSA or Wiener. The speech sequences were taken from NOIZEUS speech database [9] and were added with pink noise for evaluation. By using the objective evaluation metric described in the previous section, the tests were done with 0.01 step for both $0 \le t_y \le 0.1$ and $0.9 \le \beta \le 0.99$. The reference decisions in Eq. (11) were generated from the same speech sequences but with 50 dB global SNR to reduce miss-detections of speech. The results were generated with $K = 512$ frequency bins. A square-root Hamming window was used for $w(n)$ with 50% overlap ($R = 256$).

Fig. 1 shows the gain curves of the MSIG function in Eq. (7), the LSA approach in Eq. (6) and the Wiener filter in Eq. (5) as functions of the *a priori* SNR. A noise floor $\epsilon = 0.1$ was used for each gain function, such that $G_{(.)}(k,m) = \max\{\epsilon, G_{(.)}(k,m)\}$. Two MSIG curves are plotted in the figure to demonstrate the flexibility of the gain function. For performance evaluation, MSIG1 with parameters $a_1 = 3$, $a_2 = 1$ and $c = 0.7$ was used, which gives heavier attenuation at low *a priori* SNR region. An advantage of MSIG is a larger gain value at $\xi(k,m) = 0$ dB when compared to other two methods. This allows more speech components to be preserved.

Fig. 2 shows that when $\alpha_y = 0$, a large β is preferred for low musical noise in conventional DD approach. However, for all the gain functions, $\hat{\xi}_{\text{DD}}(k,m)$ follows the $\gamma(k,m) - 1$ estimate with one frame delay in speech frames when β is close to 1 ($\beta = 0.98$). Such drawback can be eliminated by using the proposed MDD approach. Besides that, for $\hat{\xi}_{\text{DD}}(k,m)$, the degree of smoothing at noise-only frames varies for different gain functions. As shown in the figure, hardly any smoothing can be observed apart from the LSA approach. Although all gain functions have almost the same level of smoothing for $\hat{\xi}_{\text{MDD}}(k,m)$, the variations at noise-only frames remain large. In this case, α_y plays an important role in reducing such variations.

Figs. 3-6 show the averaged KurtR, NRR and LLR measures for MSIG, LSA and Wiener. Figs. 3 and 4 show the results for DD and MDD, respectively at 0 dB SNR while Figs. 5 and 6 show the results

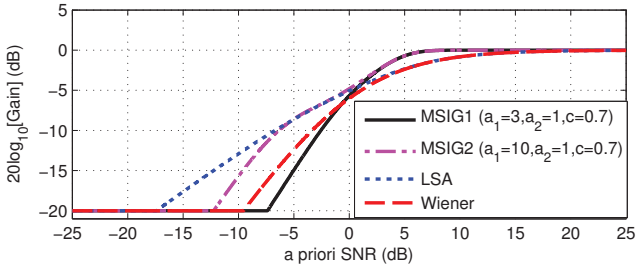

Fig. 1. Gain curve of MSIG1 (solid line), MSIG2 (dash-dotted line), LSA (dotted line), and Wiener (dashed line), as a function of the *a priori* SNR $\xi(k,m)$, where $\gamma(k,m) = \xi(k,m) + 1$.

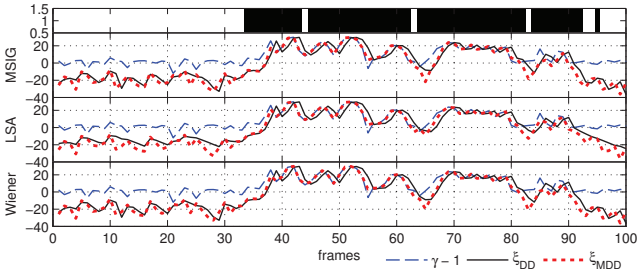

Fig. 2. Speech sequence with pink noise at 15 dB SNR: comparison between MDSVAD decisions, $\gamma(k,m) - 1$ (dashed line), $\hat{\xi}_{\text{DD}}(k,m)$ (solid line), and $\hat{\xi}_{\text{MDD}}(k,m)$ (dotted line) at 937.5 Hz, $\beta = 0.98$.

for DD and MDD, respectively at 15 dB SNR. The smoothing constant α_y is plotted instead of t_y, in conjunction to β for consistency in terms of the frame rate. As observed from the figures, the trade-off is not linear especially for the amount of speech distortion and musical noise generated from LSA method with DD approach. Thus, the advantage of the evaluation metric is a better understanding of the trade-off between those measures.

In overall, it can be seen that the MDD approach generates less musical noise, particularly for MSIG and Wiener, when compared to DD approach. In terms of the amount of noise reduction and speech distortion, the results for both the DD and MDD approaches are almost identical for all the gain functions at 0 dB SNR. Whilst at 15 dB SNR, both MSIG and Wiener have lower speech distortion for MDD approach when compared to the DD approach, except for the LSA approach. More speech distortion is observed for LSA with the MDD approach when $\beta < 0.97$. This is because when β is small, DD approach corresponds to a smoothed version of the ML estimate without much delay in speech frames. However, it is shown in Fig. 5 than when β decreases, the amount of musical noise generated for LSA remains almost identical but with less noise reduction. This indicates when β is small, more tonal sound will be perceived. In this case, NRR can be increased by increasing α_y. When α_y is large, the MDD approach introduces less musical noise and speech distortion when compared to the DD approach.

In terms of the performance of different gain functions, when $\alpha_y = 0$, the LSA approach gives the best performance with the smallest KurtR and LRR, but a smaller NRR. While α_y is considered and increased, both MSIG and Wiener can achieve less musical noise with larger NRR, particularly at 0 dB SNR. Furthermore, since MSIG has a steeper slope with heavier attenuation at low SNR region, the required degree of smoothing for β and α_y to eliminate musical noise is less than the Wiener filter. However, although MSIG

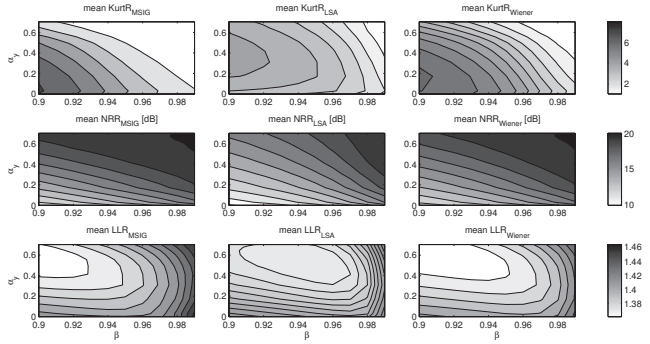

Fig. 3. Mean results with $\hat{\xi}_{\mathrm{DD}}$ at SNR = 0 dB.

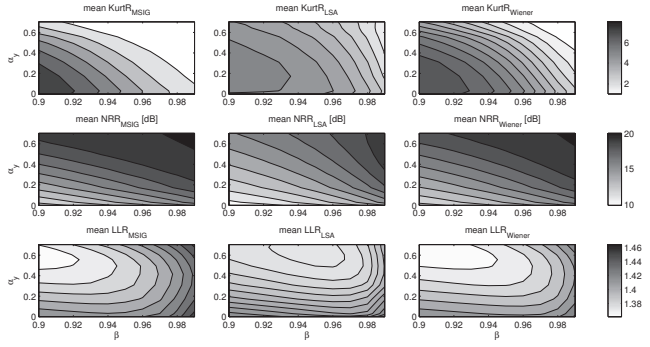

Fig. 4. Mean results with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 0 dB.

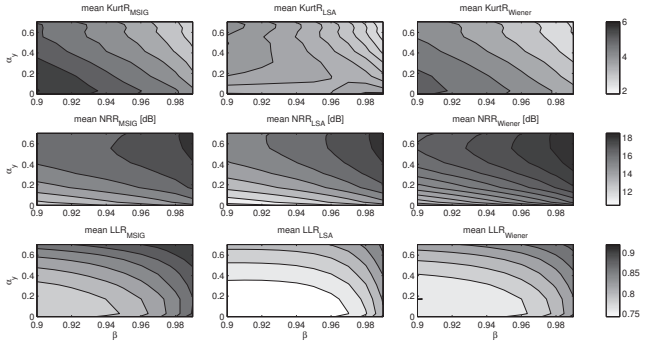

Fig. 5. Mean results with $\hat{\xi}_{\mathrm{DD}}$ at SNR = 15 dB.

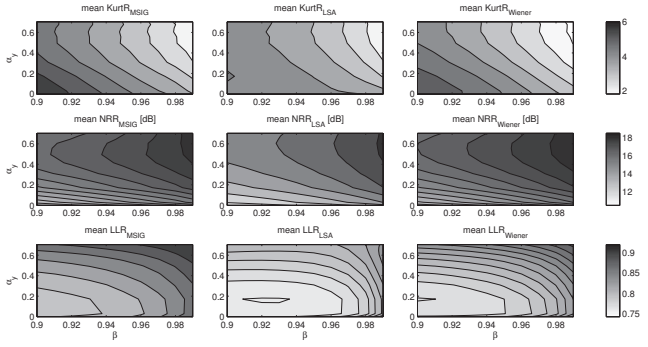

Fig. 6. Mean results with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 15 dB.

and Wiener can achieve approximately zero KurtR with sufficiently high NRR, the corresponding values of LRR become much larger compared to the LSA approach. In general, a smaller β is preferred in a speech enhancement scheme to obtain less speech distortion, while the value of α_y depends on the global SNR. In addition, a small α_y is required for high SNR conditions while a large α_y can be tolerated in low SNR conditions.

6. CONCLUSIONS

An objective evaluation metric is used to analyse the trade-off between musical noise, noise reduction and speech distortion for different noise reduction algorithms based on two *a priori* SNR estimation methods. A modified decision-directed approach is proposed to improve the performance by eliminating the speech transient distortions in the *a priori* SNR estimate. Based on the evaluation metric, the trade-off for different algorithms can be well-balanced by applying different level of smoothing using two parameters β and α_y.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[2] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[3] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[4] P. C. Yong, S. Nordholm, H. H. Dam, and S. Y. Low, "On the optimization of sigmoid function for speech enhancement," in *Proc. 19th European Signal Processing Conference (EUSIPCO'11)*, Barcelona, Spain, Aug. 2011, pp. 211–215.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] P. C. Yong, S. Nordholm, and H. H. Dam, "Noise estimation with low complexity for speech enhancement," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New York, USA, Oct. 2011.

[7] A. Davis, S. Nordholm, S. Y. Low, and R. Togneri, "A multi-decision sub-band voice activity detector," in *Proc. 14th European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, Sep. 2006.

[8] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'08)*, Seattle, USA, Sep. 2008.

[9] P. C. Loizou, *Speech Enhancement Theory and Practice*, CRC Press, 2007.