A Modified *A Priori* SNR for Speech Enhancement Using Spectral Subtraction Rules

Md. Kamrul Hasan, *Senior Member, IEEE*, Sayeef Salahuddin, and M. Rezwan Khan, *Senior Member, IEEE*

***Abstract*—This letter addresses the problem of single-channel speech enhancement using spectral subtraction method. The proposed approach is directed toward finding a self-adaptive averaging factor to estimate the *a priori* SNR. Performance of the modified averaging factor is evaluated using conventional spectral-subtraction algorithms in which the averaging factor is otherwise kept as a constant. Improved results are obtained in terms of speech quality measures for various types of noise and at different SNR levels when the time-frequency varying averaging factor, proposed in this letter, is adapted in the conventional subtraction rules.**

***Index Terms*—*A priori* signal-to-noise ratio (SNR), power subtraction, self-adaptive averaging factor, speech enhancement.**

I. Introduction

VARIOUS SPEECH processing systems have found their way in our everyday life through their vivid use in voice communication, speech and speaker recognition, aid for the hearing impaired, and numerous other applications [1]. However, in most of the cases the ambient environment is noisy which degrades the performance of the speech processing systems drastically. Therefore, speech enhancement has been a challenging topic of research for many years. Approaches to retrieve enhanced speeches are plentiful. Among them the spectral subtraction methods are the most widely used due to the simplicity of implementation and also due to low computational load, which makes them the primary choice for real time applications [2]. In general, using the family of *subtraction-type* algorithms, the enhanced speech spectrum is obtained by subtracting an average noise spectrum from the noisy speech spectrum. The phase of the noisy speech is kept unchanged, since it is assumed that the phase distortion is not perceived by human ear. However, the *subtraction-type* algorithms have a serious draw back in that the enhanced speech is accompanied by unpleasant *musical noise* artifact which is characterized by tones with random frequencies. Apart form being extremely annoying to the listeners, the *musical* noise also hampers the performance of the speech-coding algorithms to a great extent. Accordingly, various remedies have been proposed to subvert this effect [2]–[6].

This letter is directed toward finding an improved estimate of the *a priori* SNR, which in turn has the affect of reducing the *musical noise* produced by the *subtraction-type* algorithms. It has been shown in [4] that the key point behind the reduction of *musical noise* by the minimum-mean-squared-error (MMSE) estimator [7] is the use of *a priori* SNR. In this letter, a method for calculating the time-frequency varying averaging factor, used in estimating the *a priori* SNR, is proposed. It is shown that a time- and frequency-varying averaging factor, which is otherwise set to a constant value, can significantly remove the background noise and improve the speech quality.

Manuscript received January 29, 2003; revised June 25, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. P. C. Ching.

The authors are with the Department of Electrical and Electronic Engineering. Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh (e-mail: khasan@eee.buet.ac.bd).

Digital Object Identifier 10.1109/LSP.2004.824017

II. Subtraction-Type Algorithms

Let us consider that the noisy speech $y(t)$ consists of the clean speech $x(t)$ and additive uncorrelated random noise $d(t)$ such that

$$y(t) = x(t) + d(t). \tag{1}$$

Now, if at nth frame and kth frequency bin the noisy speech short-time spectral magnitude is $|Y_{n,k}|$ and the noise spectral magnitude is $|D_{n,k}|$, then an estimate of the clean speech short-time spectral magnitude can be obtained as

$$|\widehat{X}_{n,k}| = G_{n,k}\cdot|Y_{n,k}| \tag{2}$$

where

$$G_{n,k} = \sqrt{\max\left\{0, \left(1 - \frac{E\left\{|D_{n,k}|^2\right\}}{|Y_{n,k}|^2}\right)\right\}} \tag{3}$$

and $E\{\cdot\}$ is the expectation operator. Numerous variations of the original subtraction rules have been proposed. In this letter, we intend to limit ourselves only to the class of power subtraction-algorithms, since they decidedly produce the most annoying *musical noise*. In (3), the gain function takes on a value depending on the *a posteriori* SNR defined by

$$\mathrm{SNR}_{\mathrm{post}}(n,k) = \gamma_{n,k} = \frac{|Y_{n,k}|^2}{\sigma_d^2(n,k)} \tag{4}$$

where $\sigma_d^2(n,k) = E\{|D_{n,k}|^2\}$. On the other hand, *a priori* SNR is defined as

$$\mathrm{SNR}_{\mathrm{prior}}(n,k) = \xi_{n,k} = \frac{E\left\{|X_{n,k}|^2\right\}}{\sigma_d^2(n,k)}. \tag{5}$$

Ephraim and Malah [7] proposed a "decision-directed" method for the estimator $\widehat{\xi}_{n,k}$ of $\xi_{n,k}$ as

$$\widehat{\xi}_{n,k} = \alpha\frac{|\widehat{X}_{n-1,k}|^2}{\widehat{\sigma}_d^2(n-1,k)} + (1-\alpha)P[\gamma_{n,k} - 1] \tag{6}$$

where $|\widehat{X}_{n-1,k}|$ denotes the amplitude estimate of the kth speech spectral component, $\widehat{\sigma}_d^2(n-1,k)$ is the estimate of

1070-9908/04$20.00 © 2004 IEEE

variance of the kth noise spectral component in the $(n-1)$th analysis frame, the operator $P[\cdot]$ denotes half wave rectification, and α denotes an averaging parameter. Considering the maximum-likelihood estimate of the *a priori* SNR, we have $\xi_{n,k} = E\{\gamma_{n,k-1}\}$ [7]. Accordingly, the gain function of the original subtraction rule can be modified by

$$G_{n,k}^{PE} = \sqrt{\frac{\widehat{\xi}_{n,k}}{(1+\widehat{\xi}_{n,k})}} \tag{7}$$

and an estimate of the short-time clean speech spectral magnitude is obtained as

$$|\widehat{X}_{n,k}| = G_{n,k}^{PE}.|Y_{n,k}|. \tag{8}$$

This has been proposed in [4] and is denoted as power spectral estimate (PE) in this letter. In the recent years, efforts were directed to find an optimum parametric estimator [5] by using statistical distribution of speech and noise signals and optimizing in the MMSE sense. The corresponding subtraction rule is defined by

$$|\widehat{X}_{n,k}| = \sqrt{\frac{\widehat{\xi}_{n,k}^2}{0.5+\widehat{\xi}_{n,k}^2}}\sqrt{\frac{\gamma_{n,k}-1}{\gamma_{n,k}}}.|Y_{n,k}|. \tag{9}$$

However, to reduce spectral distortion that results form this subtraction rule, the authors proposed the inclusion of a spectral floor

$$|\bar{X}_{n,k}| = \begin{cases} |\widehat{X}_{n,k}|, & \text{if } |\widehat{X}_{n,k}| > \mu|Y_{n,k}| \\ f(\mu,|Y_{n,k}|), & \text{otherwise} \end{cases} \tag{10}$$

where $f(\mu,|Y_{n,k}|) = 0.5(\mu|Y_{n,k}| + |\bar{X}_{n-1,k}|)$ and now $|\bar{X}_{n,k}|$ is the estimate for short-time spectral magnitude of clean speech. This is denoted as parametric subtraction rule (PARA) in this letter. In the above subtraction rule, the value of α was proposed to be 0.96–0.995 and the value of μ to be 0.05–0.2.

III. Proposed Modification of *A Priori* SNR

In the expression of $\widehat{\xi}_{n,k}$ given by (6), the choice of α is critical. In general, α is given a value very close to 1. It has been shown [8] that the closer the value of α is to 1, the lesser is the *musical noise*, but there is more "transient distortion" to the resulting signal. Balancing these two effects, reported results in the literature usually set α a constant value in the range 0.95–0.99 with a few exceptions [9], [10]. But using a constant α has certain drawbacks. Consider an example as a test case where $\alpha = 0.98$, and the $\text{PSNR}_{\text{post}}$ shows a pulse-like behavior, i.e., for $n < n_1$ and $n > n_2$, it is very low as compared to its values in the interval n_1 to n_2, where n_1 and n_2 denote, respectively, the frames of rising and falling edges. At n_1, a signal component suddenly goes high such that $\gamma_{n_1,k} \gg \gamma_{n_1-1,k}$. Since $\widehat{\xi}_{n,k}$ contains 98% of the previous frame estimated SNR, it will fail to respond to this change. Rather, $\widehat{\xi}_{n,k}$ will rise slowly and ultimately begin to follow $\text{PSNR}_{\text{post}}$ in this high SNR region (n_1 to n_2) with some delay. Similarly at n_2, $\widehat{\xi}_{n,k}$ fails to respond to the abrupt downfall of $\text{PSNR}_{\text{post}}$ and only after a certain delay converges to the low SNR level. Therefore, it will be logical to use a much smaller value of α in these transitional areas. In [10], an adaptation scheme for α has been defined based on the assumption that the additive noise is stationary and the noise energy does not change significantly from frame to frame. A formulation using the frame energy is given by

$$\alpha_n = \sqrt{\left(1 - \frac{|FE_n - FE_{n-1}|}{\max(FE_n, FE_{n-1})}\right)} \tag{11}$$

where $FE_n = \sum_k |Y_{n,k}|^2$. Clearly, the above expression for α is based on intuitive arguments. In this letter, we develop an expression for self-adaptive α based on MMSE criterion to account for the abrupt changes in the speech spectral amplitude. The proposed modification in the estimation of *a priori* SNR is given by

$$\widehat{\xi}_{n,k} = \alpha_{n,k}\tilde{\xi}_{n-1,k} + (1-\alpha_{n,k})P[\gamma_{n,k}-1] \tag{12}$$

where $\tilde{\xi}_{n-1,k} = |\widehat{X}_{n-1,k}|^2/\widehat{\sigma}_d^2(n-1,k)$. Since the estimate $\widehat{\xi}_{n,k}$ given by (12) should actually be as close as possible to *a priori* SNR $\xi_{n,k}$, we propose an MMSE estimator for $\alpha_{n,k}$ that minimizes the error

$$J_\alpha = E\left\{(\widehat{\xi}_{n,k} - \xi_{n,k})^2 | \tilde{\xi}_{n-1,k}\right\} \tag{13}$$

given $\tilde{\xi}_{n-1,k}$. Substituting (12) into (13), an expression for J_α can be obtained as

$$J_\alpha = \alpha_{n,k}^2(\tilde{\xi}_{n-1,k} - \xi_{n,k})^2 + (1-\alpha_{n,k})^2(\xi_{n,k}+1)^2. \tag{14}$$

Note that in the above derivation $E\{(\gamma_{n,k}-1)^2\} = 2\xi_{n,k}^2 + 2\xi_{n,k} + 1$ is substituted. This result is obtained assuming that both noise and speech spectral coefficients are statistically independent zero-mean complex Gaussian random variables, and using the relation $E\{|X_{n,k}|^4\}/\sigma_d^4(n,k) = 2\xi_{nk}^2$, which follows from the definition of the fourth moment with the assumption that speech spectral amplitude $|X_{n,k}|$ has a Rayleigh distribution [5].

Now equating $\partial J_\alpha/\partial\alpha_{n,k}$ to zero, we obtain an expression for optimum $\alpha_{n,k}$ as

$$\alpha_{n,k}^{opt} = \frac{1}{1+\left(\frac{\xi_{n,k}-\tilde{\xi}_{n-1,k}}{\xi_{n,k}+1}\right)^2}. \tag{15}$$

As $\xi_{n,k}$ is unknown, (15) cannot be used directly. Nevertheless, an approximate value of $\alpha_{n,k}$ can be obtained substituting $\bar{\xi}_{n,k} = P[\gamma_{n,k}-1]$ for $\xi_{n,k}$ in (15). This is a reasonable substitution as $E\{\bar{\xi}_{n,k}\} \cong \xi_{n,k}$. If $\text{PSNR}_{\text{post}}$ over a region shows uniform variation, $\alpha_{n,k}$ will attain a value close to 1. For any abrupt change, $\alpha_{n,k}$ attains a lower value enabling $\widehat{\xi}_{n,k}$ to respond to that change more suitably.

IV. Performance Evaluation and Discussion

For effectiveness evaluation of the proposed modification, simulations were performed over several speech utterances taken from the TIMIT database. In this letter, simulation results are presented for the female utterance "Are you looking for employment" and male utterance "Would you please confirm government policy regarding waste removal." The speech signals were sampled at 8 kHz. For all cases, a frame size of

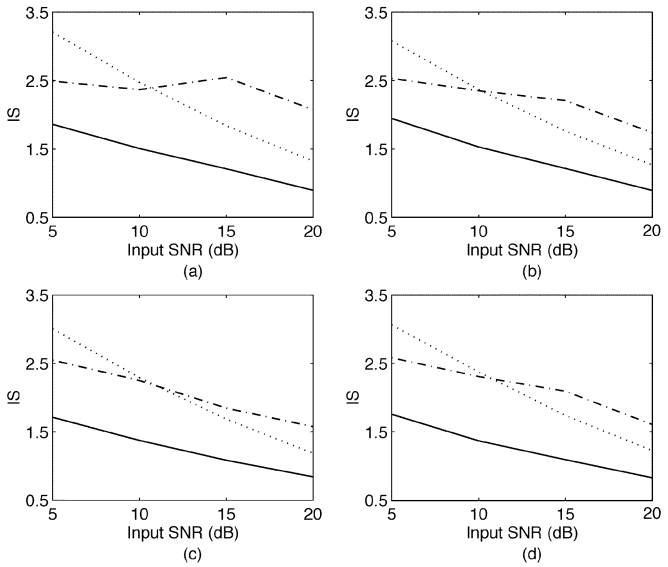

Fig. 1. Variations of IS with input SNR of PARA using (dashed–dotted line) $\alpha = 0.98$ and (solid line) the proposed $\alpha_{n,k}$ along with (dotted line) the IS of degraded speech (female) for (a) white, (b) babble, (c) vehicle, and (d) aircraft cockpit noise.

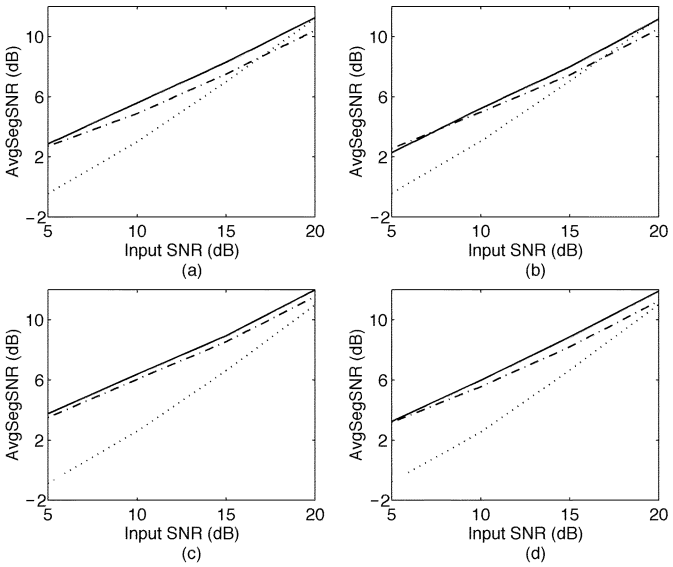

Fig. 2. Results on AvgSegSNR using the PARA method for (dashed–dotted line) $\alpha = 0.98$ and (solid line) the proposed $\alpha_{n,k}$ along with (dotted line) the AvgSegSNR of degraded speech (female) for (a) white, (b) babble, (c) vehicle, and (d) aircraft cockpit noise.

32 ms was used, and signal was reconstructed using standard overlap–add method with a 50% overlap. The performance of the PARA algorithm was examined using the conventional value of $\alpha = 0.98$ and the proposed self-adaptive $\alpha_{n,k}$, and the performance of the PE algorithm was investigated using the averaging parameter α_n defined in [10] and the proposed $\alpha_{n,k}$. Various comparative results are presented in Figs. 1–4.

A. Noise Estimation

For a stationary or slowly varying noise, a noise estimate $[\widehat{\sigma}_d(n,k)]^\beta$ can be obtained from the speech pauses as [11]

$$[\widehat{\sigma}_d(n,k)]^\beta = \lambda_D\,[\widehat{\sigma}_d(n-1,k)]^\beta + (1-\lambda_D)|Y_{n,k}|^\beta \quad (16)$$

where $0.5 \le \lambda_D \le 0.9$. In this letter, we have used $\lambda_D = 0.9$ and $\beta = 2$ for all cases.

B. Objective Quality Measures

To measure quality of the enhanced signal, we have used the Itakura–Saito (IS) measure and the segmental SNR. Both the measures show high correlation with informal listening tests. The lower the IS measure for an enhanced speech, the better is its perceived quality. The frame-based segmental SNR is formed by averaging frame level SNR estimates and is defined by [12]

$$\text{AvgSegSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{t=Nm}^{Nm+N-1}x^2(t)}{\sum_{t=Nm}^{Nm+N-1}\left(x(t)-s(t)\right)^2}\ \text{dB} \quad (17)$$

where M denotes the number of frames, and $s(t)$ may be the reconstructed signal or the noisy signal. The lower and upper thresholds are selected to be -10 and 35 dB, respectively.

C. Discussion

Figs. 1 and 2 show the variation of IS measures and AvgSegSNRs, respectively, at different noise levels and for different noise types for the parametric spectral subtraction method referred to as PARA here. Note that at low SNR levels using the proposed $\alpha_{n,k}$ has more impact on the improvement in IS measure than on the improvement in AvgSegSNR. But at high SNRs, improvements in both the objective measures are noticeable. Also, note that the conventional PARA fails to improve the IS measure for the range of input SNR of 11 dB and above. But the use of the self-adaptive $\alpha_{n,k}$ proposed in this letter has ensured significant quality improvement. It has been observed that for a given utterance, the performance of the conventional PARA is highly dependent on the choice of α. Another point to be noted with this method is that if no "floor" is used, the improvement in SNR is reasonably good but with the cost paid in quality. This is the reason why the "floor" was introduced even at the cost of overall SNR. It is evident from Figs. 1 and 2 that the self-adaptive $\alpha_{n,k}$ has further improved the performance of the algorithm in quality terms. Therefore, there is a greater flexibility in changing the "floor" parameters, and accordingly higher quality at a comparatively higher SNR can be ensured.

A comparison of the adaptation rate of the smoothing parameter (α) proposed in this letter and in [10] is presented in Fig. 3. Note that α_n defined in [10] is invariant to frequency index k; it only varies from frame to frame. In contrast, the proposed $\alpha_{n,k}$ varies with speech frame and k. Though the behavior of the two adaptation parameters as shown in Fig. 3 is different, the *a priori* SNR estimates are fairly close for the frequency index $k = 37$. Fig. 4 shows comparative IS measures and improvements in AvgSegSNR (ImpAvgSegSNR) for the PE method at different noise levels and for two different noise environments. The results shown are obtained for the male utterance. It can be observed that the improvements in terms of quality measures

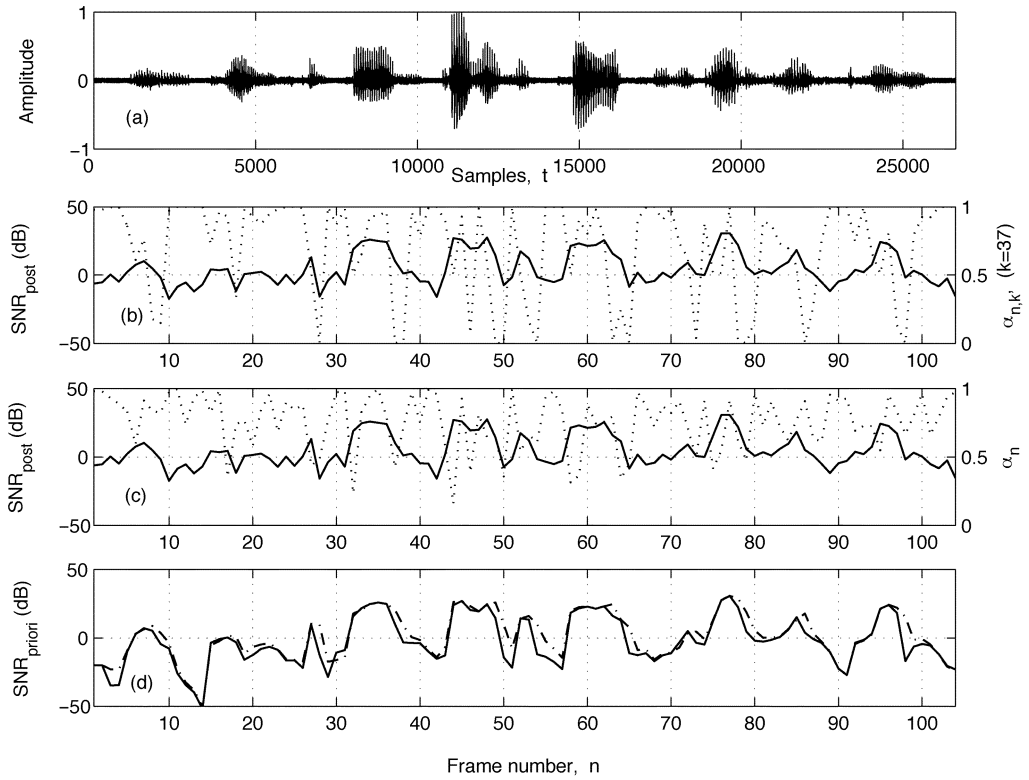

Fig. 3. Variation of adaptation parameter (α) when the speech (male) is corrupted by highway noise at SNR $= 10$ dB. (a) Noisy speech. (b) (Dotted line) Proposed $\alpha_{n,k}$, (solid line) *a posteriori* SNR. (c) (Dotted line) α_n of [10], (solid line) *a posteriori* SNR. (d) (Solid line) *a priori* SNR estimate using proposed $\alpha_{n,k}$ for $k = 37$, (dashed–dotted line) *a priori* SNR estimate using α_n of [10].

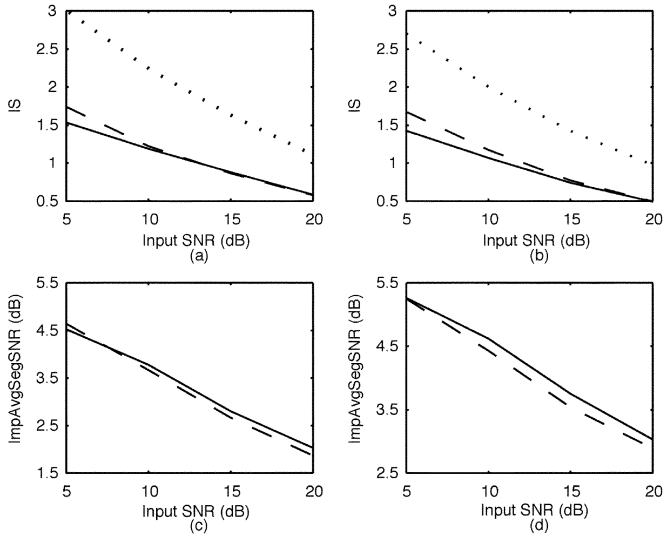

Fig. 4. Comparing performances of the PE method with adaptation parameter (α) proposed in this letter and in [10] for (a) white, (b) highway, (c) white, and (d) highway noise. (Dotted line) Degraded speech, (dashed line) using α_n of [10], (solid line) using proposed $\alpha_{n,k}$.

of the enhanced speech are relatively better using the proposed $\alpha_{n,k}$ than that of using the one defined in [10].

V. Conclusion

In this letter, an effort has been made to develop an optimal expression for the time-frequency varying averaging factor in the MMSE sense to estimate more accurately the *a priori* SNR. Comparative results have shown that incorporation of the proposed averaging factor in the conventional spectral subtraction-based algorithms produces impressive results in terms of quality measures of the enhanced speech.

References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113–120, Feb. 1979.

[3] M. Berouti, R. Scwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, Washington, DC, 1979, pp. 208–211.

[4] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, 1996, pp. 629–632.

[5] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 328–337, July 1998.

[6] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adpative averaging," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 799–807, Nov. 2001.

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

[8] O. Cappe, "Estimation of the musical noise phenomena with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.

[9] C. Beaugeant and P. Scalart, "Noise reduction using perceptual spectral change," in *Proc. EUROSPEECH Conf.*, vol. 6, 1999, pp. 2543–2546.

[10] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," *Proc. Inst. Elect. Eng., Vis. Image Signal Process.*, vol. 147, pp. 247–253, 2000.

[11] N. Virag, "Single channel speech enhancement system based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[12] J. H. L. Hansen and B. L. Pellom, "An effective evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, vol. 7, Sydney, Australia, 1998, pp. 2819–2822.