Noise Estimation based on Standard Deviation and Sigmoid Function Using *a Posteriori* Signal to Noise Ratio in Nonstationary Noisy Environments

Soo-Jeong Lee and Soon-Hyob Kim

Abstract: In this paper, we propose a new noise estimation and reduction algorithm for stationary and nonstationary noisy environments. This approach uses an algorithm that classifies the speech and noise signal contributions in time-frequency bins. It relies on the ratio of the normalized standard deviation of the noisy power spectrum in time-frequency bins to its average. If the ratio is greater than an adaptive estimator, speech is considered to be present. The propose method uses an auto control parameter for an adaptive estimator to work well in highly nonstationary noisy environments. The auto control parameter is controlled by a linear function using *a posteriori* signal to noise ratio (SNR) according to the increase or the decrease of the noise level. The estimated clean speech power spectrum is obtained by a modified gain function and the updated noisy power spectrum of the time-frequency bin. This new algorithm has the advantages of much more simplicity and light computational load for estimating the stationary and nonstationary noise environments. The proposed algorithm is superior to conventional methods. To evaluate the algorithm's performance, we test it using the NOIZEUS database, and use the segment signal-to-noise ratio (SNR) and ITU-T P.835 as evaluation criteria.

Keywords: Noise reduction, noise estimation, speech enhancement, sigmoid function.

1. INTRODUCTION

Noise estimation algorithm is an important factor of many modern communications systems. Generally implemented as a preprocessing component, noise estimation and reduction improve the performance of speech communication system for signals corrupted by noise through improving the speech quality or intelligibility. Since it is difficult to reduce noise without distorting the speech, the performance of noise estimation algorithm is usually a trade-off between speech distortion and noise reduction [1].

Current single microphone speech enhancement methods belong to two groups, namely, time domain methods such as the subspace approach and frequency domain methods such as the spectral subtraction (SS), and minimum mean square error (MMSE) estimator [2,3]. Both methods have their own advantages and drawbacks. The subspace methods provide a mechanism to control the tradeoff between speech distortion and residual noise, but with the cost of a heavy computational load [4]. Frequency domain methods, on the other hand, usually consume less computational resources, but do not have a theoretically established mechanism to control tradeoff between speech distortion and residual noise. Among them, spectral subtraction (SS) is computationally efficient and has a simple mechanism to control tradeoff between speech distortion and residual noise, but suffers from a notorious artifact known as "musical noise" [5]. These spectral noise reduction algorithms require an estimate of the noise spectrum, which can be obtained from speech-absence frames indicated by a voice activity detector (VAD) or, alternatively, with the minimum statistic (MS) methods [6], i.e., by tracking spectral minima in each frequency band. In consequence, they are effective only when the noise signals are stationary or at least do not show rapidly varying statistical characteristics.

Many of the state-of-the-art noise estimation algorithms use the minimum statistic methods [6-9]. These methods are designed for unknown nonstationary noise signals. Martin proposed an algorithm for noise estimation based on minimum statistics [6]. The ability to track varying noise levels is a prominent feature of the minimum statistics (MS) algorithm [6]. The noise estimate is obtained as the minima values of a smoothed power estimate of the

Manuscript received November 4, 2007; revised October 31, 2008; accepted November 3, 2008. Recommended by Guest Editor Phill Kyu Rhee.

Soo-Jeong Lee is with the BK 21 program of Sungkunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea (e-mail: leesoo86@sorizen.com).

Soon-Hyob Kim is with the Department of Computer Engineering, Kwangwoon University, 447-1, Wolgye-dong, Nowon-gu, Seoul 139-701, Korea (e-mail: kimsh@kw.ac.kr).

noisy signal, multiplied by a factor that compensates the bias. The main drawback of this method is that it takes somewhat more than the duration of the minimum-search windows to update the noise spectrum when the noise level increases suddenly [7]. Cohen proposed a minima controlled recursive algorithm (MCRA) [8] which updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. These regions are found by comparing the ratio of the noisy speech to the local minimum against a threshold. However, the noise estimate delays by at most twice that window length when the noise spectrum increases suddenly [7]. A disadvantage to most of the noise-estimation schemes mentioned is that residual noise is still present in frames in which speech is absent. In addition, the conventional noise estimation algorithms are combined with a noise reduction algorithm such as the SS and MMSE [2,3].

In this paper, we explain a method to enhance speech by improving its overall quality while minimizing residual noise. The proposed algorithm is based on the ratio of the normalized standard deviation (STD) of the noisy power spectrum in the time-frequency bin to its average and a sigmoid function (NTFAS). This technique, which we call the "NTFAS noise reduction algorithm," determines that speech is present only if the ratio is greater than the adaptive threshold estimated by the sigmoid function. In the case of a region where a speech signal is strong, the ratio of STD will be high. This is not high for a region without a speech signal. Specifically, our method uses an adaptive method for tracking the threshold in a nonstationary noisy environment to control the trade-off between speech distortion and residual noise. The adaptive method uses an auto control parameter to work well in highly nonstationary noisy environments. The auto control parameter is controlled by a linear function using *a posteriori* signal to noise ratio (SNR) according to the increase or the decrease of the noise level.

The clean speech power spectrum is estimated by the modified gain function and the updated noisy power spectrum of the time-frequency bin. We tested the algorithm's performance with the NOISEUS [10] database, using the segment signal-to-noise ratio (SNR) and ITU-T P.835 [11] as evaluation criteria. We also examined its adaptive tracking capability in nonstationary environments. We show that the performance of the proposed algorithm is superior to that of the conventional methods. Moreover, this algorithm produces a significant reduction in residual noise .

The structure of the paper is as follows. Section 2 introduces the overall signal model. Section 3 describes the proposed noise reduction algorithm, while Section 4 contains the experimental results and discussion. The conclusion in Section 5 looks at future research directions for the algorithm.

2. SYSTEM MODEL

Assuming that speech and noise are uncorrelated, the noisy speech signal $x(n)$ can be represented as

$$x(n) = s(n) + d(n), \qquad (1)$$

where $s(n)$ is the clean speech signal and $d(n)$ is the noise signal. The signal is divided into the overlapped frames by window and the short-time Fourier transform (STFT) is applied to each frame. The time-frequency representation for each frame is as follows. $X(k,l) = S(k,l) + D(k,l)$, where $(k = 1, 2, ..., L)$ are the frequency bin index and $(l = 1, 2, ..., L)$ are the frame index. The power spectrum of the noisy speech $|X(k,l)|^2$ can be represented as

$$|X(k,l)|^2 \approx |S(k,l)|^2 + |\hat{D}(k,l)|^2, \qquad (2)$$

where $|S(k,l)|^2$ is the power spectrum of the clean speech signal and $|\hat{D}(k,l)|^2$ is the power spectrum of the noise signal.

The proposed algorithm is summarized in the block diagram shown in Fig. 1. It is consists of seven main components: window and fast Fourier transform (FFT), standard deviation of the noisy power spectrum and estimation of noise power, calculation of the ratio, adaptive threshold using the sigmoid function, classification of speech presence and absence in time-frequency bins and updated gain function, updated noisy power spectrum, and product of the modified gain function and updated noisy power spectrum.

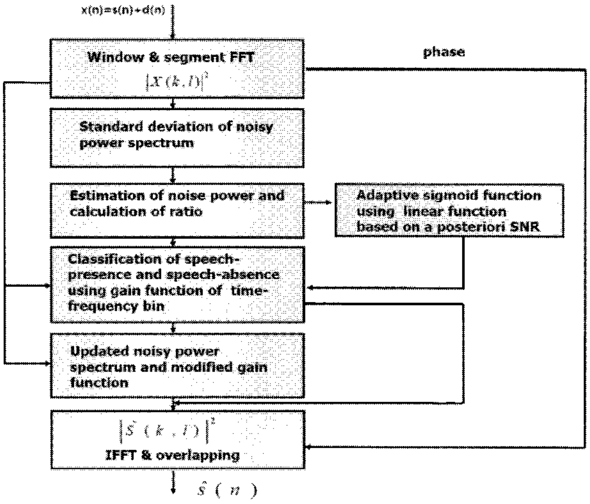

x(n)=s(n)+d(n)
Window & segment FFT
$|X(k,l)|^2$
phase
Standard deviation of noisy power spectrum
Estimation of noise power and calculation of ratio
Adaptive sigmoid function using linear function based on a posteriori SNR
Classification of speech-presence and speech-absence using gain function of time-frequency bin
Updated noisy power spectrum and modified gain function
$|\hat{S}(k,l)|^2$
IFFT & overlapping
$\hat{s}(n)$

Fig. 1. Flow diagram of proposed noise reduction algorithm.

3. PROPOSED NOISE ESTIMATION AND REDUCTION ALGORITHM

The noise reduction algorithm is based on the STD of the noisy power spectrum in a time and frequency-dependent manner as follows:

$$\bar{x}_t(l)=\frac{1}{K}\sum_{k=1}^{K}|X(k,l)|^2,\ \bar{x}_f(k)=\frac{1}{L}\sum_{l=1}^{L}|X(k,l)|^2,\tag{3}$$

$$v_t(l)=\sqrt{\left[\frac{1}{K}\sum_{k=1}^{K}\left(|X(k,l)|^2-\bar{x}_t(l)\right)^2\right]},\tag{4}$$

$$v_f(k)=\sqrt{\left[\frac{1}{L}\sum_{l=1}^{L}\left(|X(k,l)|^2-\bar{x}_f(k)\right)^2\right]},\tag{5}$$

$$\hat{\sigma}_t^2=\frac{1}{L}\sum_{l=1}^{L}v_t(l),\ \hat{\sigma}_f^2=\frac{1}{K}\sum_{k=1}^{K}v_f(k),\tag{6}$$

$$\gamma_t(l)=\frac{v_t(l)}{\hat{\sigma}_t^2},\ \gamma_f(k)=\frac{v_t(k)}{\hat{\sigma}_f^2},\tag{7}$$

where $\bar{x}_t(l)$ is the average noisy power spectrum in the frequency bin, $\bar{x}_f(k)$ is the average noisy power spectrum for the frame index, and $\hat{\sigma}_t^2$ and $\hat{\sigma}_f^2$ are the assumed estimate of noise power. (7) gives the ratio of the (STD) for the noisy power spectrum in the time-frequency bin to its average. In the case of a region in which a speech signal is strong, the STD ratio by (7) will be high. The ratio is generally not high for a region without a speech signal. Therefore, we can use the ratio in (7) to determine speech-presence or speech-absence in the time-frequency bins [12].

3.1. Classification of speech-presence and speech-absence in frames using an adaptive sigmoid function based on *a posteriori* SNR

Our method uses an adaptive algorithm with a sigmoid function to track the threshold and control the trade-off between speech distortion and residual noise:

$$\psi_t(l)=\left[\frac{1}{1+\exp\left(10\cdot(\gamma_t(l)-\delta_t)\right)}\right],\tag{8}$$

where $\psi_t(l)$ is the adaptive threshold using the sigmoid function. We defined a control parameter δ_t. This threshold $\psi_t(l)$ is adaptive in the sense that it changes depending on the control parameter δ_t. The control parameter δ_t is derived from the linear function using the *a posteriori* signal to noise ratio (SNR) in frame index.

$$\delta_t=\delta_s\cdot SNR(l)+\delta_{off},\tag{9}$$

$$SNR(l)=10\cdot\log\left[\frac{norm\left(|X(k,l)|^2,2\right)}{norm\left(\left|\hat{D}(k)\right|^2,2\right)}\right],\tag{10}$$

where $\left|\hat{D}(k)\right|^2\approx 1/5\sum_{l=1}^{5}|X(k,l)|^2$ is the average of the $|X(k,l)|^2$ initial 5 frames during the period of the first silence and *norm* is the Euclidean length of a vector.

$$\delta_{off}=\delta_{\max}-\delta_s\cdot SNR_{\min},\tag{11}$$

$$\delta_s=\frac{\delta_{\min}-\delta_{\max}}{SNR_{\max}-SNR_{\min}},\tag{12}$$

where δ_s is the slope of the δ_t, δ_{off} is the offset of the δ_t. The constants $\delta_{\min}=0.1$, $\delta_{\max}=0.5$, $SNR_{\min}=-5dB$ and $SNR_{\max}=20dB$ are the experimental values we used. Consequently, the *a posteriori* SNR in (10) controls the δ_t. Fig. 2 shows that the more the *a posteriori* SNR increases, the more the δ_t decreases. Simulation results show that an increase in the δ_t parameter is good for noisy signals with a low SNR of less than 5 dB, and that a decrease in δ_t is good for noisy signals with a relatively high SNR of greater than 15 dB. We can thus control the trade-off between speech distortion and residual noise in the frame index using δ_t. Fig. 3 shows that the adaptive threshold using the sigmoid function allows for a trade-off between speech distortion and residual noise by controlling δ_t. If a speech signal is present, the $\psi_t(l)$ calculated by (8) will be extremely small (i.e., very close to 0). Otherwise, the value of $\psi_t(l)$ calculated by (8) will be approximately 1. Fig. 4 is a

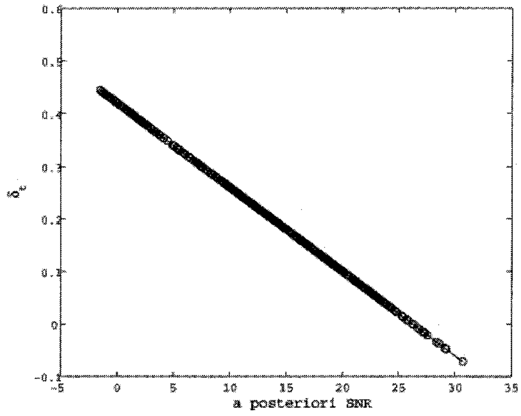


Fig. 2. The linear function using the *a posteriori* SNR for the control parameter δ_t.

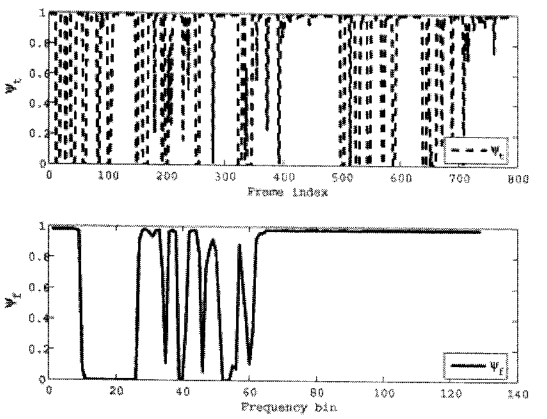


Fig. 3. Adaptive thresholds using a sigmoid function on the time-frequency bin index for 15dB car noise, 5dB car noise, 10dB babble noise, 0dB white noise, and 5dB SNR babble noise in a nonstationary environments. Top panel: the adaptive thresholds of the time index (dotted line). Bottom panel: the adaptive thresholds of the frequency bin index (heavy line).

good illustration of Fig. 3.

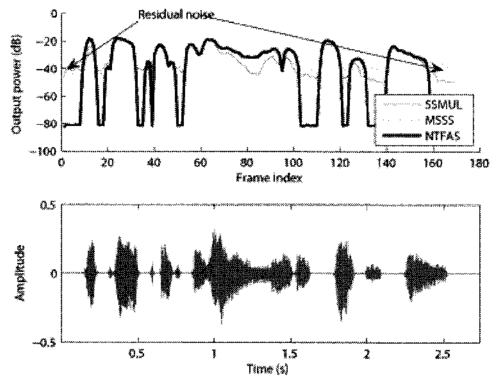


Fig. 4. Example of noise reduction by three enhancement algorithms with 5dB car noise for the sp12.wav female speech sample of "The drip of the rain made a pleasant sound" from the NOIZEUS database. Top panel: output power for car noise 5dB using the SSMUL method (solid line), the MSSS method (dotted line), and NTFAS method (heavy line). Bottom panel : enhanced Speech signal using NTFAS.

3.2. Updated noisy power spectrum using classification of speech-presence and absence in frames

The classification rule for determining whether speech is present or absent in a frame is based on the following algorithm:

If $\psi_t(l) > \phi_t$

$$\hat{D}^2_{level}(k,l) = |X(k,l)|^2$$

$$\hat{D}^2_{mean}(k) = \left[\frac{1}{l}\sum_{m=1}^{l}(\frac{1}{K}\sum_{k=1}^{K}\hat{D}^2_{level}(k,l)\right]$$

$$G_{update}(k,l) = G(k,l)\cdot\alpha$$

else

$$\hat{D}^2_{level}(k,l) = \hat{D}^2_{mean}(k)$$

$$G_{update}(k,l) = G(k,l)\cdot(1-\alpha),$$

where decision parameter ϕ_t and parameter α are initially 0.99 and the gain function $G(k,l)$ is 1.0. The threshold $\psi_t(l)$ is compared to the decision parameter ϕ_t. If it is greater than ϕ_t, then speech is determined to be absent in the l^{th} frame; otherwise speech is present. Then, the l^{th} frames of the noisy spectrum $|X(k,l)|^2$ are set to $\hat{D}^2_{level}(k,l)$. We estimate $\hat{D}^2_{level}(k,l)$ frames of the noise power spectrum, and $\hat{D}^2_{mean}(k)$ is calculated by averaging over the frames without speech. The $\hat{D}^2_{mean}(k)$ is the assumed estimate of the residual noise of the frames in the presence of speech. We refer to this value as the "sticky noise" of the speech-presence index. Then we represent $G_{update}(k,l)$, the updated gain function in a frame index using the gain function $G(k,l)$ and the parameter α for the frames in which speech is absent. If the l^{th} frame is considered to be frame in which speech is present, then $\hat{D}^2_{mean}(k)$ is set to $\hat{D}^2_{level}(k,l)$ and $\hat{D}^2_{mean}(k)$ is used to reduce the sticky noise of the frames of in the presence of speech. We can see the sticky noise in the the square region and residual noise in the random peak region in Fig. 5.

As a noted above, $G_{update}(k,l)$ is the updated gain

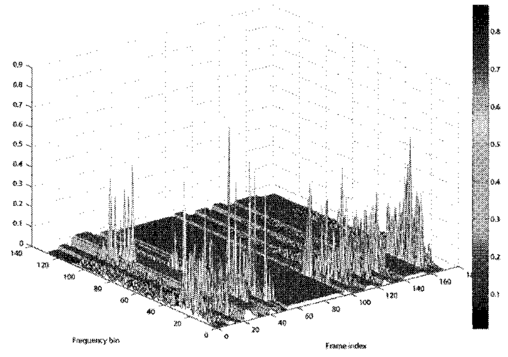


Fig. 5. Estimated noise power spectrum at car noise 10dB sp12.wav of female "The drip of the rain made a pleasant sound" from the NOIZEUS database.

function in a frame index using the gain function $G(k,l)$ and the parameter $(1-\alpha)$ for the frames in which speech is present. Figs. 6 and 7 show the gain function $G(k,l)$ and the updated gain function $G_{update}(k,l)$, respectively:

$$\left|X_{update}(k,l)\right|^2 = \left|X(k,l)\right|^2 - \hat{D}^2_{level}(k,l), \tag{13}$$

$$\left|X_{update}(k,l)\right|^2 = MAX(\left|X_{update}(k,l)\right|^2, \alpha). \tag{14}$$

The updated noisy power spectrum of the frame index $\left|X_{update}(k,l)\right|^2$ is the difference between the noisy power spectrum $\left|X(k,l)\right|^2$ and the frames in which speech is absent. $\hat{D}^2_{level}(k,l)$, as shown in Fig. 8, Fig. 9 and Fig. 5, respectively: (13) reduces the noise of the frames in which speech is absent, and (14) is used to avoid negative values.

3.3. Classification of speech-presence and absence in frequency bins using an adaptive sigmoid function based on *a posteriori* SNR

In a manner parallel to that described bins in the previous subsection, our method uses an adaptive algorithm with a sigmoid function to track the threshold in a frequency bins:

$$\psi_f(k) = \left[\frac{1}{1+\exp\left(10\cdot(\gamma_f(k)-\delta_f)\right)}\right], \tag{15}$$

where $\psi_f(k)$ is the adaptive threshold using the sigmoid function in the frequency bins. We define a control parameter δ_f. The threshold $\psi_f(k)$ is adaptive in the sense that it changes depending on the control parameter δ_f. The control parameter of the frequency bin δ_f is derived from the linear function using the *a posteriori* signal to noise ratio (SNR) in frequency bins.

$$\delta_f = \delta_{fs}\cdot SNR(k) + \delta_{fo}, \tag{16}$$

$$SNR(k) = 10\cdot\log\left[\frac{\left(\left|X(k,l)\right|^2\right)}{\left(\left|\hat{D}_{level}(k)\right|^2\right)}\right], \tag{17}$$

where $\left|\hat{D}_{level}(k)\right|^2$ is the estimate of the noise power spectrum in frequency bins.

$$\delta_{fo} = \delta_{f\max} - \delta_{fs}\cdot SNR_{\min}, \tag{18}$$

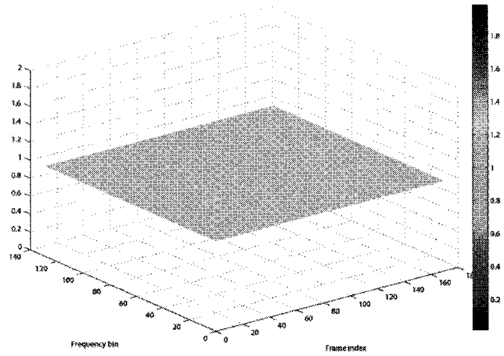


Fig. 6. Gain function.

$$\delta_{fs} = \frac{\delta_{f\min} - \delta_{f\max}}{SNR_{\max} - SNR_{\min}}, \tag{19}$$

where δ_{fs} is the slope of the δ_f, δ_{fo} is the offset of the δ_f. The constants $\delta_{\min} = 0.1$, $\delta_{\max} = 0.5$, $SNR_{\min} = -5\,dB$ and $SNR_{\max} = 20\,dB$ are the experimental values we used. Simulation results indicate that the control parameter δ_f will be optimal over a wide range of SNRs. Fig. 3 shows that the adaptive threshold ψ_f accounts for the frequency bin index by controlling δ_f. Consequently, we can control the trade-off between speech distortion and residual noise in the frequency bins using δ_f in Fig. 10.

3.4. Noise reduction using a modified gain functioand updated noisy power

The classification algorithm for determining whether speech is present or absent in a frequency bin is

$$
\begin{aligned}
&\mathit{If}\ \psi_f(k) > \phi_f\\
&\qquad G_{\mathrm{mod}\,i}(k,l) = G_{update}(k,l)\cdot\alpha\\
&\mathit{else}\\
&\qquad G_{\mathrm{mod}\,i}(k,l) = G_{update}(k,l)\cdot(1-\alpha).
\end{aligned}
$$

In the same manner as for the time index, where decision parameter ϕ_f is initially 0.95, this threshold $\psi_f(k)$ is compared to the decision parameter ϕ_f. If it is greater than ϕ_f, then speech is determined to be absent in the frequency bin k^{th}; otherwise speech is present. The $G_{\mathrm{mod}\,i}(k,l)$ represents the modified gain function for the time and frequency bins using the gain function $G_{update}(k,l)$, the parameter α, and $(1-\alpha)$.

$$\left|\hat{S}(k,l)\right|^2 = G_{\mathrm{mod}\,i}(k,l)\cdot\left|X_{update}(k,l)\right|^2 \tag{20}$$

Finally, the estimated clean speech power spectrum $\left|\hat{S}(k,l)\right|^2$ can be represented as a product of the modified gain function for the time-frequency bins and the updated noisy power spectrum of the time-frequency bins. The estimated clean speech signal can then be transformed back to the time domain using the inverse short-time Fourier transform (STFT) and synthesis with the overlap-add method. We can see the modified gain function and the estimated clean speech power spectrum in Figs. 11 and 12, respectively.

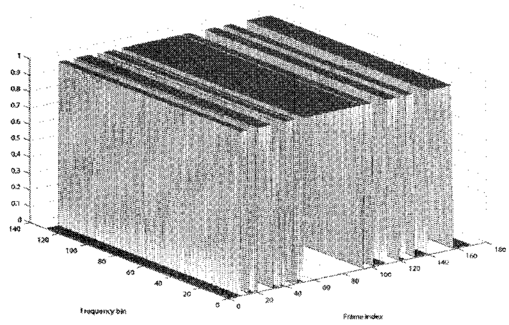


Fig. 7. Updated gain function.

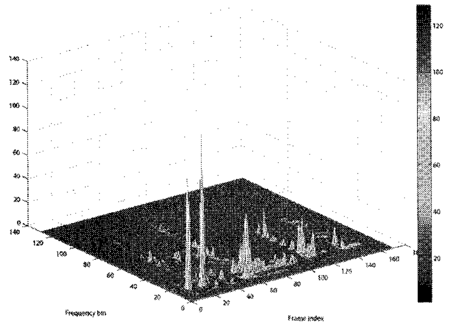


Fig. 8 . Updated noisy power spectrum with 10dB car noise for the female sp12.wav speech sample “The drip of the rain made a pleasant sound”from the NOIZEUS database.

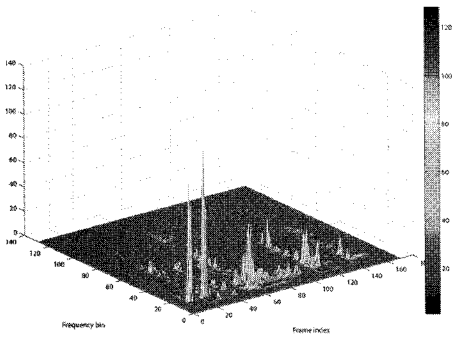


Fig. 9. Noisy power spectrum with 10dB car noise for the female sp12.wav speech sample “The drip of the rain made a pleasant sound” from the NOIZEUS database.

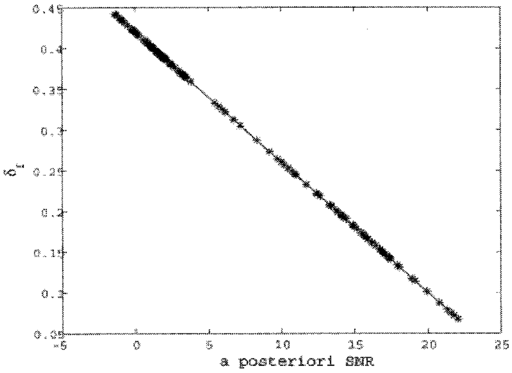


Fig. 10. The linear function using the a posteriori SNR for the contorl parameter δ_f.

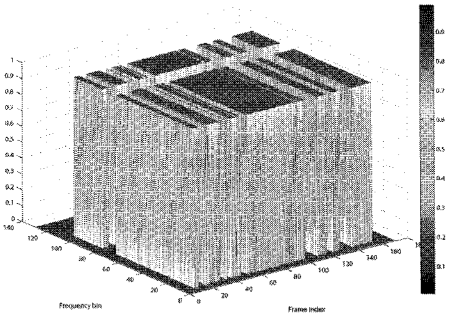


Fig. 11. Modified gain function.

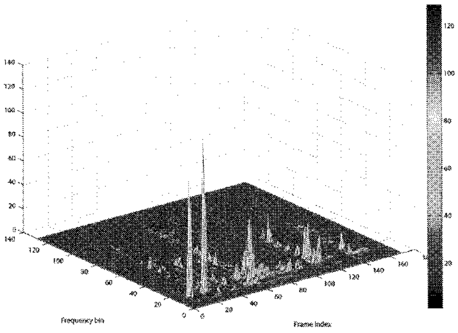


Fig. 12. Estimated clean speech power spectrum with 10dB car noise for the female sp12.wav speech sample “The drip of the rain made a pleasant sound” from the NOIZEUS database.

4. EXPERIMENTAL RESULTS AND DISCUSSION

For our evaluation, we selected three male and three female noisy speech samples from the NOIZEUS database [10]. The signal was sampled at 8 kHz and transformed by the STFT using 50%

overlapping Hamming windows of 256 samples. Evaluating of the new algorithm and a comparing it to the multi band spectral subtraction (MULSS) and MS with spectral subtraction (MSSS) methods [6,13] consisted of two parts. First, we tested the segment SNR. This provides a much better quality measure than the classical SNR since it indicates an average error over time and frequency for the enhanced speech signal. Thus, a higher segment SNR value indicates better intelligibility. Second, we used ITU-T P.835 as a subjective measure of quality [11]. This standard is designed to include the effects of both the signal and background distortion in ratings of overall quality [10].

4.1. Segment SNR and speech signal

We measured the segment SNR over short frames and obtained the final result by averaging the value of each frame over all the segments. Table 1 shows the segment SNR improvement for each speech enhancement algorithm. For the input SNR in the range 5-15dB for white Gaussian noise, car noise, and babble noise, we noted that the segment SNR after processing was clearly better for the proposed algorithm than for the MULSS and the MSSS methods [6,13]. The proposed algorithm yields a bigger improvement in the segment SNR with lower residual noise than the conventional methods. The NTFAS algorithm in particular produces good results for white Gaussian noise in the range 5 to 15dB. Figs. 13 and 14 show the NTFAS algorithm's clear superiority in the 10dB car noise environment.

For nonstationary noisy environments, the conventional methods worked well for high input SNR values of 10 and 15dB; however, the output they produced could not be easily understood for low SNR values of car noise (5dB) and white noise (0dB), and they produced residual noise and distortion as shown in Fig. 15. This outcome is also confirmed by time-frequency domain results of speech enhancement methods illustrated in Figs. 15 and 16. A different result is clear in Fig. 15(a) and (b) for the waveforms of the clean and noisy speech signals, respectively, (c) the waveforms of speech enhancement using the MULSS method, (d) the MSSS method, and (e) the proposed NTFAS method. Fig. 15(c) and (d) show that the presence of residual noise at $t > 7.8s$ is due

Table 1. Segmental SNR at white, car and babble noise 5 through 15dB.

	Noise (dB)	white	babble	car
MULSS	5	4.96	5.89	7.08
	10	8.13	9.28	8.05
	15	10.05	9.89	10.35
MSSS	5	6.83	5.41	6.71
	10	11.20	9.65	10.96
	15	15.23	14.11	14.92
NTFAS	5	**9.98**	**6.44**	**7.58**
	10	**11.93**	**10.68**	**11.87**
	15	**16.53**	**14.49**	**15.70**

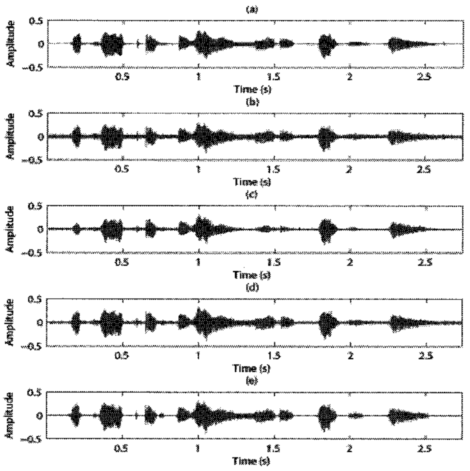


Fig. 13. Example of noise reduction with 10dB car noise with female sp12.wav speech sample "The drip of the rain made a pleasant sound" from the NOIZEUS database for the three enhancement algorithms. (a) original signal, (b) noisy signal, (c) signal enhanced using the MULSS method, (d) signal enhanced using the MSSS method, and (e) signal enhanced using the NTFAS method.

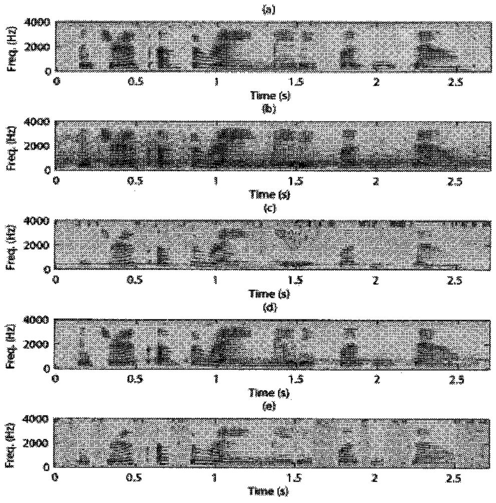


Fig. 14. Example of noise reduction with 10dB car noise with female sp12.wav speech sample "The drip of the rain made a pleasant sound" from the NOIZEUS database for the three enhancement algorithms. (a) original spectrogram, (b) noisy spectrogram, (c) spectrogram using the MULSS method, (d) spectrogram using the MSSSmethod, and (e) spectrogram using the NTFAS method.

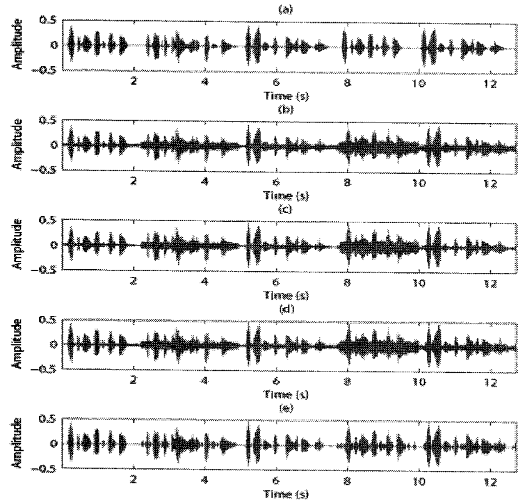


Fig. 15. Time domain results of speech enhancement for 15dB car noise, 5dB car noise, 10dB babble noise, 0dB white noise, and 5dB SNR babble noise in a nonstationary environment. The noisy signal comprises five concatenated sentences from the NOIZEUS database. The speech signal were two male and one female sentences from the AURORA 2 corpus. (a) original speech, (b) noisy speech, (c) speech enhanced using MULSS method,; (d) speech enhanced using the MSSS method, (e) speech enhanced using the NTFAS method.

partly to the inability of the speech enhancement algorithm to track the sudden appearance of a low SNR. In contrast, panel (e) shows that the residual noise is clearly reduced with the proposed NTFAS algorithm.

4.2. The ITU-T P.835 standard

Noise reduction algorithms typically degrade the speech component in the signal while suppressing the background noise, particularly under low-SNR conditions. This situation complicates the subjective evaluation of algorithms as it is not clear whether listeners base their overall quality judgments on the distortion of the speech or the presence of noise. The overall effect of speech and noise together was rated using the scale of the Mean Opinion Score (MOS), scale of background intrusiveness (BAK), and the SIG

Table 2. The overall effect (OVL) using the Mean Opinion Score (MOS), 5= excellent, 4= good, 3= fair, 2= poor, 1= bad.

	Noise (dB)	white	babble	car
MULSS	5	1.84	2.47	**2.78**
	10	3.14	2.96	**3.05**
	15	3.57	3.49	3.90
MSSS	5	2.98	**2.66**	2.74
	10	4.41	**3.19**	3.04
	15	4.43	**5.00**	3.30
NTFAS	5	**3.55**	2.55	2.31
	10	**4.62**	2.67	2.87
	15	**4.73**	4.56	**4.40**

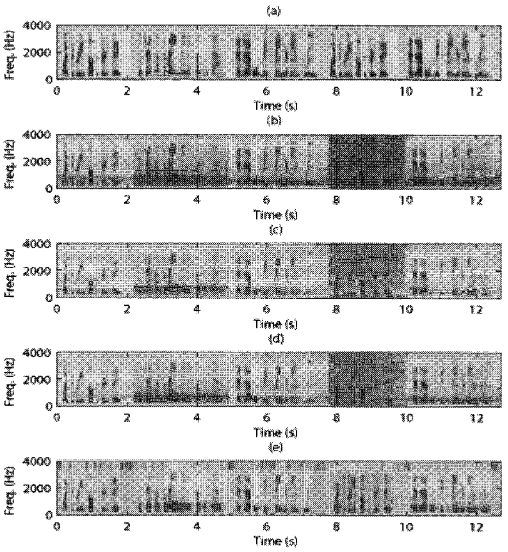


Fig. 16. Frequency domain results of speech enhancement for 15dB car noise, 5dB car noise, 10dB babble noise, 0dB white noise, and 5dB SNR babble noise in a nonstationary environment. The noisy signal comprises five concatenated sentences from the NOIZEUS database. The speech signal were two male and one female sentences from the AURORA 2 corpus. (a) original spectrogram, (b) noisy spectrogram, (c) spectrogram using the MULSS method, (d) spectrogram using the MSSS method, (e) spectrogram using the NTFAS method.

Table 3. Scale of Background Intrusiveness (BAK), 5= not noticeable, 4= somewhat noticeable, 3= noticeable but notintrusive, 2= fairly conspicuous, somewhat intrusive, 1= very intrusive.

	Noise (dB)	white	babble	car
MULSS	5	**3.58**	2.21	**2.83**
	10	3.31	2.37	3.01
	15	**5.00**	3.01	1.79
MSSS	5	3.38	1.63	2.18
	10	**4.11**	2.46	2.69
	15	3.54	3.00	2.60
NTFAS	5	3.25	**2.54**	2.17
	10	3.63	**2.85**	**3.09**
	15	4.58	**5.00**	**5.00**

Table 4. Scale of Signal Distortion (SIG), 5=no degradation, 4= little degradation, 3= somewhat degraded, 2= fairly degraded, 1= very degraded.

	Noise (dB)	white	babble	car
MULSS	5	1.79	2.81	2.87
	10	2.69	3.26	3.74
	15	3.15	3.37	3.75
MSSS	5	1.93	3.25	**3.92**
	10	2.96	**3.63**	**3.92**
	15	4.53	**3.87**	**4.01**
NTFAS	5	**2.69**	**3.28**	3.60
	10	**4.06**	3.30	3.63
	15	**4.72**	3.73	3.80

[10]. The proposed method resulted in a great reduction in noise, while providing enhanced speech with lower residual noise and somewhat higher MOS, BAK, and SIG scores than the conventional methods. It also degraded the input speech signal in highly nonstationary noisy environments. This is confirmed by an enhancement signal and ITU-T P.835 test [11]. The results of the evaluation are shown in Tables 2, 3, and 4. The best result for each speech enhancement algorithms is shown in bolds.

5. CONCLUSIONS

In this paper, we proposed a new approach to the enhancement of speech signals that have been corrupted by stationary and nonstationary noise. This approach is not a conventional spectral algorithm, but uses a method that separates the speech-presence and speech-absence contributions in time-frequency bins. We call this technique the NTFAS speech enhancement algorithm. The propose method used an auto control parameter for an adaptive threshold to work well in highly nonstationary noisy environments. The auto control parameter was affected by a linear function by application *a posteriori* signal to noise ratio (SNR) according to the increase or the decrease of the noise level. The proposed method resulted in a great reduction in noise while providing enhanced speech with lower residual noise and somewhat MOS, BAK and SIG scores than the conventional methods. In the future, we plan to evaluate its possible application in preprocessing for new communication systems, human-robotics interactions, and hearing aid systems.

REFERENCES

[1] M. Bhatnagar, *A Modified Spectral Subtraction Method Combined with Perceptual Weighting for Speech Enhancement*, Master's Thesis, University of Texas at Dallas, 2003.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

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

[4] Y. Hu, *Subspace and Multitaper Methods for Speech Enhancement*, Ph.D. Dissertation. University of Texas at Dallas, 2003.

[5] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech Audio Processing*, vol. 2, no. 2, pp. 346-349, 1994.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech Audio Processing*, vol. 9, no. 5, pp. 504-512, 2001.

[7] R. Sundarrajan and C. L. Philipos, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, pp. 220-231, 2006.

[8] I. Cohen, "Noise spectrum in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech Audio Processing*, vol. 11, no. 5, pp 466-475, 2003.

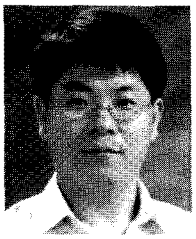
[9] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 725-728, 2004.

[10] C. L. Philipos, *Speech Enhancement (Theory and Practice)*, 1st edition, CRC Press, Boca Raton, FL, 2007.

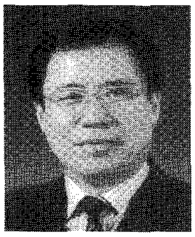
[11] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Recommendation,* p. 835, 2003.

[12] S. J. Lee and S. H. Kim, "Speech enhancement using gain function of noisy power estimates and linear regression," *Proc. of IEEE/FBIT Int. Conf. Frontiers in the Convergence of Bioscience and Information Technologies*, pp. 613-616, October 2007.

[13] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. of International Conference on Acoustics*, Speech and Signal Processing, pp. 4164-4167, 2002.



Soo-Jeong Lee received the B.S. degree in Computer Science from Korea National Open University in 1997, and the M.S. and Ph.D. degrees in Computer Engineering from Kwangwoon University, Seoul, Korea, in 2000 and 2008, respectively. He is currently a Post-Doc. Fellow, Sungkyunkwan University (BK 21 Program). His research interests include speech enhancement, adaptive signal processing, and noise reduction.



Soon-Hyob Kim received the B.S. degree in Electronics Engineering from Ulsan Unversity, Korea in 1974, and the M.S. and Ph.D. degrees in Electronics Engineering from Yonsei University, Korea, in 1976 and 1983, respectively. He is currently a Professor, Dept. of Computer Engineering, Kwangwoon University. His area of interest are speech recognition, signal processing, and human-computer interaction.