THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.27, No.1E 2008. 3. pp. 1-10

Noise Suppression Using Normalized Time-Frequency Bin Average and Modified Gain Function for Speech Enhancement in Nonstationary Noisy Environments

Soojeong Lee*, Soonhyob Kim*
*Department of Computer Engineering, Kwangwoon University
(Received June 12, 2007; revised October 22, 2007; accepted April 3; 2008)

Abstract

A noise suppression algorithm is proposed for nonstationary noisy environments. The proposed algorithm is different from the conventional approaches such as the the spectral subtraction algorithm and the minimum statistics noise estimation algorithm in that it classifies speech and noise signals in time-frequency bins. It calculates the ratio of the variance of the noisy power spectrum in time-frequency bins to its normalized time-frequency average. If the ratio is greater than an adaptive threshold, speech is considered to be present. Our adaptive algorithm tracks the threshold and controls the trade-off between residual noise and distortion. 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 simplicity and light computational load for estimating the noise. This algorithm reduces the residual noise significantly, and is superior to the conventional methods.

Keywords: *Speech enhancement, Noise reduction, Noise estimator*

I. Introduction

Noise suppression is a crucial factor of many modern speech communications systems. Generally implemented as a preprocessing component, noise suppression improves the performance of communication systems for speech signals corrupted by noise through improving the speech quality or intelligibility. As it is difficult to reduce noise without distorting the speech, the performance of speech enhancement systems is usually a trade-off between speech distortion and noise reduction [1].

Current single microphone speech enhancement methods belong to two categories, namely, time domain methods such as the subspace approach and frequency domain methods such as the spectral subtraction (SS), minimum mean square error (MMSE) estimator [2, 3] and Wiener filter. 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 heavy computational load.

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 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" [4, 5]. The MMSE estimator and Wiener estimator have moderate computational load, but have no mechanism to control trade

Corresponding author: Soojeong Lee (leesoo86@kw.ac.kr)
Kwangwoon University, Wolgye-dong, Nowon-gu, Seoul 447-1, Korea.

off between speech distortion and residual noise [5].

Recently, various studies have examined noise estimation techniques [6, 7, 8, 9, 10]. These methods are designed for unknown nonstationary noise signals using minimum statistics. 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 main drawback of this method is that it takes more time than the duration of the minimum-search window to update the noise spectrum when the noise level increases suddenly.

Cohen proposed a minima-controlled recursive algorithm (MCRA) [9] 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 least twice the window length when the noise spectrum increases suddenly [8].

Moreover, a disadvantage to most of the noise-estimation schemes mentioned above is that residual noise is still present in frames in which speech is absent. In addition, the estimation of the noise spectrum is quite complex.

In this paper, we describe a method to enhance speech by improving its overall quality while minimizing residual noise and distortion. The proposed algorithm is based on calculating the ratio of the variance of the noisy power spectrum in the time-frequency bin to its normalized time-frequency average (NTFA) [11]. The algorithm determines that speech is present only if the ratio is greater than the adaptive threshold. Specifically, our method uses an adaptive scheme for tracking the threshold in a nonstationary noisy environment to control the trade-off between speech distortion and residual noise. The algorithm has the advantages of simplicity and light computational load for estimating the noise. Moreover, this algorithm reduces the residual noise significantly. We compare the new algorithm to the conventional methods [6, 7]. The segmental signal-to-noise ratio (SNR), and the ITU-T (P.835) subjective measure [12] were compared under various noise conditions [13]. We examine the adaptive tracking capability for non-stationary environments. We show that the performance of the proposed algorithm is superior to that of conventional methods.

The structure of the paper is as follows. Section 2 introduces the overall system 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.

II. System model

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

$$x(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ is the clean speech signal and $d(n)$ is the noise signal. Dividing the signal into overlapping frames using a window function and applying the short-time Fourier transform (STFT) to each frame gives the time-frequency representation $X(k,l) =, S(k,l) + D(k,l)$ where k is the frequency bin index and l is the frame index [14]. In a more formal form, the noisy spectrum can be represented as

$$|X(k,l)| = \sum_{n=0}^{N-1} x(n+lL)w(n)e^{-j(\frac{2\pi}{N})nk} \tag{2}$$

where w is the window function, N is the size of the windows and L is the framing step. The power spectrum of the noisy speech $|X(k,l)|^2$ can then be represented as

$$|X(k,l)|^2 = |S(k,l)|^2 + |D(k,l)|^2 \tag{3}$$

where $|S(k,l)|^2$ is the power spectrum of the clean speech signal and $|D(k,l)|^2$ is the power spectrum of the noise signal. The proposed algorithm is summarized in the block diagram shown in Fig. 1. It consists of seven main components: time-frequency analysis

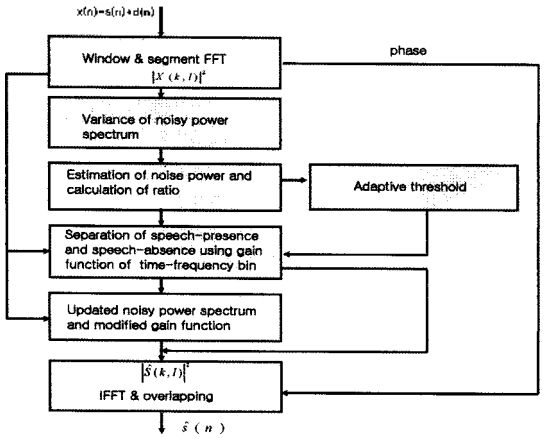


Figure 1. Flow diagram of the proposed speech enhancement algorithm.

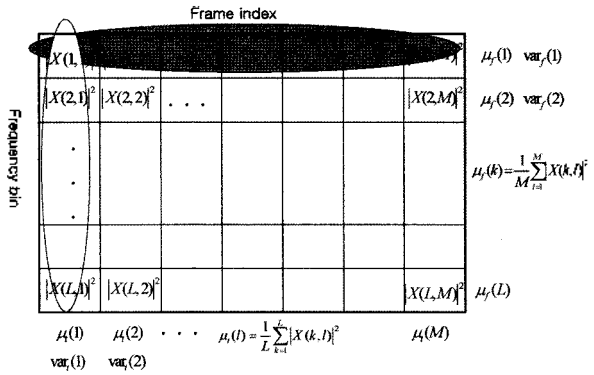


Figure 2. Procedure for estimating noise power using the noisy power spectrum.

and synthesis, noise power estimate, adaptive threshold, separation of speech-presence and absence in time-frequency bin, updated noisy power spectrum, and modified gain function.

III. Proposed noise reduction algorithm

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

$$\mu_t(l)=\frac{1}{K}\sum_{k=1}^{K}|X(k,l)|^2,\ \mu_f(k)=\frac{1}{L}\sum_{l=1}^{L}|X(k,l)|^2 \tag{4}$$

$$var_t(l)=\frac{1}{K}\sum_{k=1}^{K}(|X(k,l)|^2-\mu_t(l))^2 \tag{5}$$

$$var_f(k)=\frac{1}{L}\sum_{l=1}^{L}(|X(k,l)|^2-\mu_f(k))^2 \tag{6}$$

$$\widehat{\sigma_t^2}=\frac{1}{L}\sum_{l=1}^{L}var_t(l),\ \widehat{\sigma_f^2}=\frac{1}{K}\sum_{k=1}^{K}var_f(k) \tag{7}$$

$$\gamma_t(l)=\frac{var_t(l)}{\widehat{\sigma_t^2}},\ \gamma_f(k)=\frac{var_f(k)}{\widehat{\sigma_f^2}}$$

$$1\le k\le K,\ 1\le l\le L \tag{8}$$

where μ_t is the average of noisy power spectrum in the frequency bin, μ_f is the average of noisy power spectrum for the frame index, and $\widehat{\sigma_t^2}$ and $\widehat{\sigma_f^2}$ is the assumed estimate of noise power. Equation 8 gives the ratio of the variance for the noisy power spectrum in the time-frequency bin to its normalized time-frequency average. In the case of a region where a strong speech signal is present, the normalized variance calculated by Equation 8 will be high. This is generally not true for a region without a speech signal. Therefore, we can use the ratio in Equation 8 to classify the speech-presence and absence in the time-frequency bins. Figure 2 shows the model for estimating the noise power using the noisy power spectrum.

3.1. Separation of speech and noise in frames using an adaptive thresholds

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

$$\xi_t(1)=\gamma_t(1)+\delta_t \tag{9}$$

$$\xi_{tL}=\xi_t(1)\cdot\delta_{tL} \tag{10}$$

$$\xi_{tU}=\xi_t(1)\cdot\delta_{tU} \tag{11}$$

$$\alpha_t(t)=\xi_t(l-1)-\gamma_t(l-1) \tag{12}$$

$$IF\ \alpha_t(l)>\delta_t, \tag{13}$$

$$\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_a)+\xi_{tL}\cdot\zeta_a$$

$$else\ IF\ \zeta_a\le\alpha_t(l)\le\delta_t,$$

$$\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_b)+\xi_{tU}\cdot\zeta_b$$

$$else$$

$$\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_a)+\xi_{tU}\cdot\zeta_a$$

where $\xi_t(l)$ is the adaptive threshold set recursively. Equation 9 is the initial value of $\xi_t(l)$, and ξ_{tL}

and ξ_{tU} are the lower and upper limits of $\xi_t(l)$, respectively. We define control parameter, δ_t, lower limit constant, δ_{tL}= 0.5 and upper limit constant, δ_{tU} = 2.0, and can obtain $\xi_t(1)$ using the ratio $\gamma_t(1)$ and the control parameter δ_t. In Equation 9, we estimate the $\gamma_t(1)$ for the frame of the initial speech-absence. Equation 10 and 11, the lower and upper limits of $\xi_t(l)$ are defined by the $\xi_t(1)$, δ_{tL} and δ_{tU}. Then, the tracking value $\alpha_t(l)$ calculated as the difference between the adaptive threshold $\xi_t(l-1)$ and the ratio $\gamma_t(l-1)$. Equation 12 is used to track the adaptive thresholds $\xi_t(l)$. This threshold $\xi_t(l)$ is adaptive in the sense that it changes depending on the tracking value $\alpha_t(l)$. The adaptive threshold $\xi_t(l)$ that operate the prior frame is required; the adaptive thresholds not use noise power spectrum estimate from the speech-absence frames such as the SS and MMSE [2, 3]. If $\alpha_t(l)$ is greater than the control parameter δ_t, $\xi_t(l)$ decreases using the lower limit ξ_{tL} and the weighting constant $\zeta_a = 0.8$ for tracking the ratio $\gamma_t(l)$ between the lower ξ_{tL} and upper limits ξ_{tU}. If $\alpha_t(l)$ is between $\zeta_z = 0.0$ and δ_t, $\xi_t(l)$ does slightly increases using the upper limit ξ_{tU} and the weighting constant $\zeta_b = 0.6$. If $\alpha_t(l)$ is less than the tracking constant $\zeta_z = 0.0$, the threshold level $\xi_t(l)$ increases using the upper limit ξ_{tU} and the weighting constant $\zeta_a = 0.8$ for tracking the ratio $\gamma_t(l)$ between the ξ_{tL} and ξ_{tU}. Constants $\delta_{tL} = 0.5$, $\delta_{tU} = 2.0$, $\zeta_a = 0.8$, $\zeta_b = 0.6$, and $\zeta_z = 0.0$ are experimental values we used.

In order to improve the balance between the speech distortion and residual noise, we show that the adaptive threshold allows a trade-off between speech distortion and residual noise by controlling δ_t in Figure 3. Figure 3 shows the effect of δ_t on SNR gains. The output SNR is calculated in a manner similar to the input SNR. The noise power is calculated as the power of the speech signal obtained by subtracting the filtered speech signal from the clean speech signal. 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. The δ_t parameter is set to a constant of 0.1 based on initial experiments, but a fixed δ_t will clearly not be optimal over a wide range of SNRs. For example, setting δ_t to 0.25 yields high SNR gain at a low input SNR of 5 dB; however, it also degrades the input speech signal at a high SNR of 15 dB.

Distortion of the original speech signal is extremely undesirable in real practical environments. Second, Figure 4 shows the effect of δ_t on signal distortion (SIG) scores. Simulation results show that the increase in δ_t is beneficial for noisy signals with low SNRs about 5 dB; however, it also degrades the input speech signal at a high SNR of 15 dB. Consequently, we can control the trade-off between speech distortion and residual noise in the frame index using δ_t.

As explained in Section 1, the noise power spectrum of the MS algorithm estimate is obtained by tracking

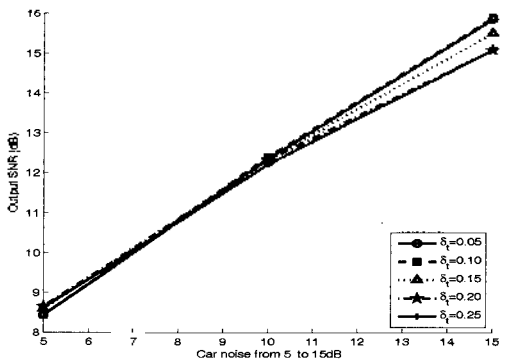


Figure 3. Effect of various δ_t values on SNR gains.

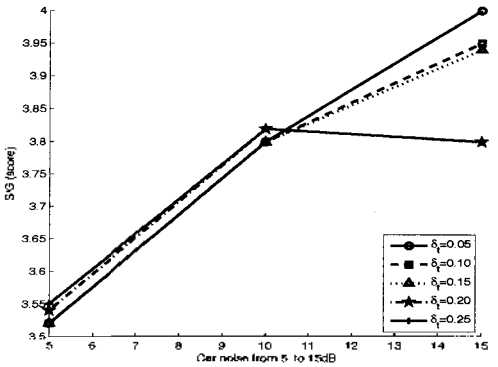


Figure 4. Effect of various δ_t values on SIG.

the minimum of the noisy speech power spectrum over a specified window of finite frames. This fact is illustrated in Figure 5 where babble noise at 10 dB is followed by speech with white Gaussian noise at 0 dB [8]. There is a definite delay of a finite number of frames when the SNR changes. We show that the adaptive thresholds of proposed method immediately track the variation of the ratio as the noise power suddenly increases in Figure 6. If the proposed method generate speech distortion, we can solve this problem to decrease the control parameter δ_t. The proposed method could arise with residual noise, and vice versa. Figure 6 shows the improved tracking capability of the proposed algorithm. Specially, the speech signal is degraded by highly nonstationary car noise at 5 dB and white Gaussian noise 0 dB segmental SNR. The NTFA algorithm is essentially much faster than that of the MS [6] algorithm in Figure 5. The separation rule for determining whether speech is present or absent in a frame is based on the following algorithm:

$$If\ \gamma_t(l) \le \xi_t(l),\quad 1 < l \le L \tag{14}$$

$$\widehat{D^2_{level}}(k,l) = |X(k,l)|^2,\ 1 \le k \le K,\ 1 \le l \le L \tag{15}$$

$$\widehat{D^2_{mean}}(k,1) = \frac{1}{l}\sum_{m=1}^{l}\left(\frac{1}{K}\sum_{k=1}^{k}\widehat{D^2_{level}}(k,l)\right) \tag{16}$$

$$G_{update}(k,l) = G(k,l) \cdot \alpha \tag{17}$$

$else$

$$\widehat{D^2_{level}}(k,l) = \widehat{D^2_{mean}}(k,1) \tag{18}$$

$$G_{update}(k,l) = G(k,l) \cdot (1-\alpha) \tag{19}$$

where gain constant α is set to 0.001, and the gain function $G(k,l)$ is initially 1.0. This ratio $\gamma_t(l)$ is then compared to the adaptive threshold $\xi_t(l)$. If it is greater than $\xi_t(l)$, then speech is declared to be present in l-th frame; otherwise speech is absent. If the l-the frame is considered to be a speech-absence frame, then $\widehat{D^2_{level}}(k,l)$ is set to l-th frame of the noisy spectrum $|X(k,l)|^2$. We estimate $\widehat{D^2_{level}}(k,l)$ frames of the noise power spectrum, and $\widehat{D^2_{mean}}(k,1)$ is calculated by averaging over the frames without speech. The $\widehat{D^2_{mean}}(k,1)$ is the assumed estimate of the residual noise of the speech-frames where speech is present. Then we represent $G_{update}(k,l)$, the updated gain function in a frame index using the gain function $G(k,l)$ and the constant α for the frames in which speech is absent. If l-th frame are considered to be frames in which speech is present, then $\widehat{D^2_{mean}}(k,1)$ is set to $\widehat{D^2_{level}}(k,l)$, and $\widehat{D^2_{mean}}(k,1)$ is used to reduce the residual noise of the frames of in the presence of speech. Then, we can represents $G_{update}(k,l)$, the updated gain function in a frame index, using the gain function $G(k,l)$ and the constant α for the frame in which speech is absent. As mentioned above, $G_{update}(k,l)$ denotes the updated gain function in a frame index using the gain function $G(k,l)$ and the constant $1-\alpha$ for the frames with speech. $G_{update}(k,l)$ is used to

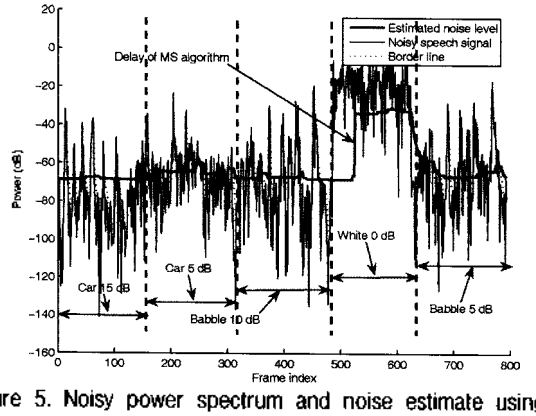


Figure 5. Noisy power spectrum and noise estimate using for car noise (15 dB and 5 dB), babble noise (10 dB), white Gaussian noise 0 dB and babble noise 5 dB in a nonstationary at f=625 Hz.

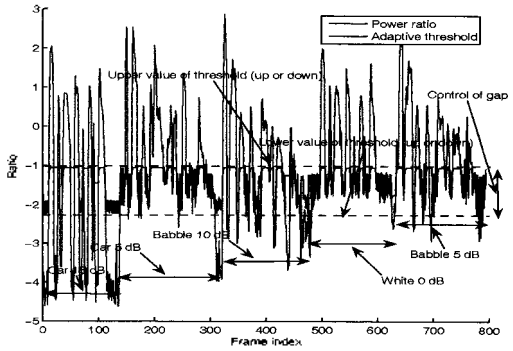


Figure 6. Ratio for adaptive thresholds determined on the time index for car noise (15 dB and 5 dB), babble noise 10 dB, white noise 0 dB and babble noise 5 dB in a nonstationary noisy environments.

remove residual noise in the frame index.

The updated noisy power spectrum $|X_{update}(k,l)|^2$ of the frame index is the difference between the noisy power spectrum $|X(k,l)|^2$ and the frames $\widehat{D^2_{level}}(k,l)$ in which speech is absent, as shown in Equation 20:

$$|X_{update}(k,l)|^2 = |X(k,l)|^2 - \widehat{D^2_{level}}(k,l) \quad (20)$$

$$|X_{update}(k,l)|^2 = MAX(|X_{update}(k,l)|^2,\ \alpha) \quad (21)$$

$$|X(k,l)|^2 = |X_{update}(k,l)|^2,\ 1 \le k \le K,\ 1 \le l \le L \quad (22)$$

Equation 20 reduces the noise of the frames in which speech is absent, and Equation 21 is used to avoid negative values [13].

3.2. Separation of speech and noise in frequency bins using adaptive thresholds

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

$$\xi_f(1) = \gamma_f(1) + \delta_f \quad (23)$$

$$\xi_{fL} = \xi_f(1) \cdot \delta_{fL} \quad (24)$$

$$\xi_{fU} = \xi_f(1) \cdot \delta_{fU} \quad (25)$$

$$\alpha_f(k) = \xi_f(k-1) - \gamma_f(k-1),\quad 1 \le k \le K \quad (26)$$

$$\begin{aligned} &If\ \alpha_f(k) \ge \delta_f,\quad 1 \le k \le K \quad (27)\\ &\quad \xi_f = \xi_f(k-1) \cdot (1-\eta_a) + \xi_{fL} \cdot \eta_a\\ &else\ If\ \eta_z \le \alpha_f(k) \ge \delta_f\\ &\quad \xi_f = \xi_f(k-1) \cdot (1-\eta_a) + \xi_{fU} \cdot \eta_b\\ &else\\ &\quad \xi_f = \xi_f(k-1) \cdot (1-\eta_a) + \xi_{fU} \cdot \eta_a \end{aligned}$$

where $\xi_f(k)$ is the adaptive threshold set recursively. Equation 23 is the initial value of $\xi_f(k)$, and ξ_{fL} and ξ_{fU} are the lower and upper limits of $\xi_f(k)$, respectively. We define control parameter, δ_f, lower limit constant, $\delta_{fL} = 0.5$ and upper limit constant, $\delta_{fU} = 2.0$, and can obtain $\xi_f(k)$ using the ratio $\gamma_f(k)$ and the control parameter δ_f. In Equation 23, we estimate the $\gamma_f(1)$ the frequency bin of the initial speech-absence.

Equation 24 and 25, the lower and upper limits of $\xi_f(k)$ are defined by the $\xi_f(1)$, δ_{fL} and δ_{fU}. Then, the tracking value $\alpha_f(k)$ calculated as the difference between the adaptive threshold $\xi_f(k-1)$ and the ratio $\gamma_f(k-1)$. Equation 26 is used to track the adaptive thresholds $\xi_f(k)$. This threshold $\xi_f(k)$ is adaptive in the sense that it changes depending on the tracking value $\alpha_f(k)$. The adaptive threshold $\xi_f(k)$ that operate the prior frame are required; the adaptive thresholds not use noise power spectrum estimate from the speech-absence frequency bin. If $\alpha_f(k)$ is greater than the control parameter δ_f, $\xi_f(k)$ decreases using the lower limit ξ_{fL} and the weighting constant $\eta_a = 0.8$ for tracking the ratio $\gamma_f(k)$ between the lower ξ_{fL} and upper limits ξ_{fU}. If $\alpha_f(k)$ is between $\eta_z = 0.0$ and δ_f, $\xi_f(k)$ does slightly increases using the upper limit ξ_{fU} and the weighting constant $\eta_b = 0.6$. If $\alpha_f(k)$ is less than the tracking constant $\eta_z = 0.0$, the threshold level $\xi_f(k)$ increases using the upper limit ξ_{fU} and the weighting constant $\eta_a = 0.8$ for tracking the ratio $\gamma_f(k)$ between the ξ_{fL} and ξ_{fU}. Figure 7 shows that the adaptive threshold accounts for the frequency bin index by controlling δ_f.

Figures 8 and 9 show the effect of δ_f on SNR gains and scale of the SIG. Simulation results indicate that the optimal value of δ_f is 0.01 for noisy signals with SNR 5 through 15 dB. Figure 9 shows that the adaptive threshold accounts for the frequency bin index by controlling δ_f.

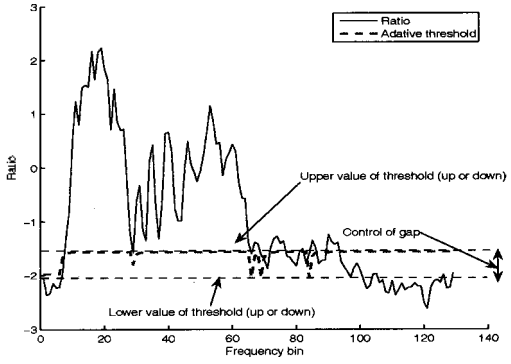


Figure 7. Ratio for adaptive thresholds determined on the frequency bin index in a nonstationary noisy environment.

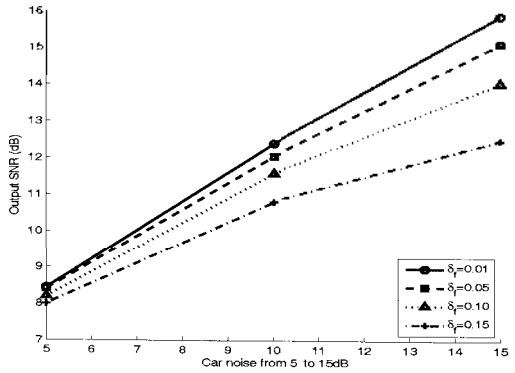


Figure 8. Effect of various algorithm δ_f values on SNR gains.

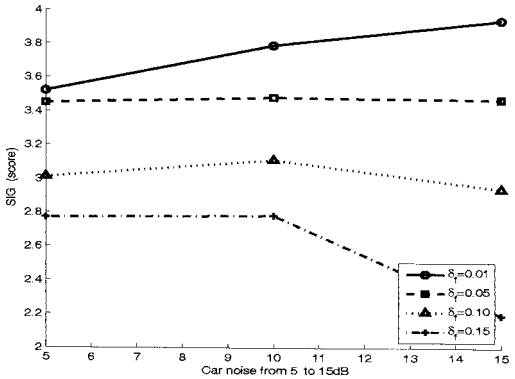


Figure 9. Effect of various algorithm δ_f values on SIG.

Then, the decision algorithm for determining if speech is present or absent in a frequency bin is

$$If\ \gamma_f(k) \le \xi_f(k),\quad 1 < k \le K \tag{28}$$

$$G_{modi}(k,l) = G_{update}(k,l) \cdot \alpha \tag{29}$$

$$G(k,l) = G_{modi}(k,l) \cdot \alpha \tag{30}$$

$$else$$

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

$$G(k,l) = G_{modi}(k,l) \tag{31}$$

$$1 \le k \le K,\quad 1 \le l \le L \tag{32}$$

In the same manner as for the time index, if the ratio $\gamma_f(k)$ is greater than the adaptive threshold $\xi_f(k)$, then the bin is declared to contain speech; otherwise speech is absent. $G_{modi}(k,l)$ represents the modified gain function for the time and frequency bin.

$$|\hat{S}(k,l)|^2 = G(k,l) \cdot |X(k,l)|^2 \tag{33}$$

Finally, the estimated clean speech power spectrum $|\hat{S}(k,l)|^2$ can be represented as a product of the modified gain function $G(k,l)$ and the updated noisy power spectrum $|X(k,l)|^2$.

The estimated clean speech signal can then be obtained by the overlap-add method, where the estimated power spectrum is transformed to the time domain by the inverse STFT.

IV. Experimental results and discussion

The noisy signals used in our evaluation are taken from the NOIZEUS database [13]. The speech signals were spoken by two female and two male speakers. The analyzed signal was sampled at 8 kHz, and transformed into the STFT using 50% (128 samples) overlapping Hamming windows of 256 samples.

4.1. Segmental SNR

Segmental SNR is measured over short frames and final result is obtained by averaging the value of each frame over all the segments. Table 1 shows the segmental SNR improvement for each speech enhancement algorithm. For the input SNR range 5-15 dB for white Gaussian noise, car noise, and babble noise, we noted that the segmental SNR after processing is clearly better for the proposed algorithm than that for the conventional methods [6, 7] under white Gaussian and babble noise environmental conditions. The NTFA algorithm in particular produces good results for white Gaussian noise in the range 5 to 15 dB.

Table 1. Segmental SNR at white, babble and car.

	Noise (dB)	white	babble	car
SSMS	5	7.29	6.33	5.66
	10	11.62	10.68	10.99
	15	15.66	15.24	15.10
WIENERWT	5	10.30	8.90	6.14
	10	14.48	12.25	9.90
	15	17.85	16.24	15.19
NTFAS	5	10.83	8.43	6.39
	10	14.75	12.40	11.65
	15	18.28	15.90	15.74

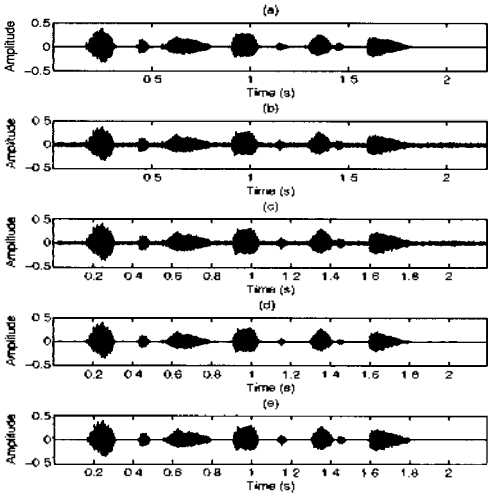


Figure 10. Example of noise reduction at car noise 5 dB sp 05. wav of male "Wipe the grease off his dirty face" from the NOIZEUS database using three enhancement algorithm. (a) Original speech; (b) Noisy speech; (c) Enhanced speech using SSMS method; (d) Enhanced speech using WIENERWT method; (e) Enhanced speech using NTFA method.

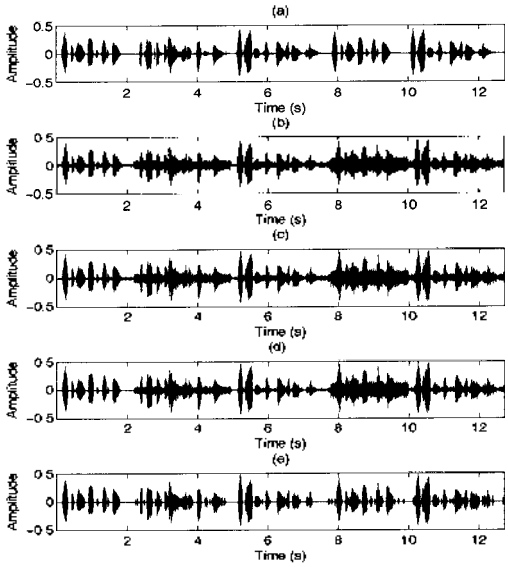


Figure 11. Time domain results of speech enhancement for car noise at 15 dB, car noise at 5 dB, babble noise at 10 dB, white noise 0 dB, and babble noise 5 dB SNR in a nonstationary environment. (a) Original speech; (b) Noisy speech; (c) Enhancement speech using SSMS; (d)Enhancement speech using WIENERWT; (e)Enhancement speech using NTFA.

Figure 10 shows the results in the 5 dB car noise environment. The best result for each speech enhancement algorithms is shown in bolds.

For nonstationary noisy environments, the conventional methods worked well for high input SNR values of 10 and 15 dB; however, their output showed poor intelligibility for low SNR values of car noise (5 dB) and white noise (0 dB), and they produced residual noise and distortion as shown in Figure 11. Figure 11 shows the proposed algorithm's clear superiority in a highly nonstationary noisy environment. This outcome are also confirmed by the time and frequency domain results of speech enhancement by the three methods shown in Figure 11. In a Figure 11 (c), different outcome is observed in the waveforms of speech enhancement by the SSMS method, (d) the WIENERWT method, and (e) the proposed speech enhancement the NTFA method. Panels (c) and (d) in Figure 11 from t>2.2s to t<4.5s and at t > 7.8 s show the presence of residual noise partly due to the inability of the speech enhancement algorithm to track the sudden appearance of low SNR noise. In contrast, as shown in panel (e), the residual noise is clearly reduced with the proposed speech enhancement algorithm. The noisy signal comprises five concatenated sentences from the NOIZEUS database. The speech signal were spoken by two male and one female sentences extracted from the AURORA 2 corpus.

4.2. The ITU-T P.835 Standard

Speech enhancement algorithms typically degrade the speech signal component while suppressing the background noise, particularly in low-SNR conditions. This situation complicates the subjective evaluation of speech enhancement algorithm as it is not clear as to whether listeners base their overall quality judgments on the signal distortion and noise distortion. The methodology proposed in [12] reduces the listener's uncertainty by requiring them to successively attend to and rate the waveform on the speech alone, the noise alone, and the overall effect of speech and noise on quality. The speech signal alone using a five-point scale of signal distortion (SIG) [13]. The proposed method resulted in a great reduction in noise. It also degraded the speech signal in highly nonstationary noisy environments. Degradation of the speech signal is extremely undesirable in real

Table 2. 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
SSMS	5	1.65	2.69	3.22
	10	2.28	3.75	3.96
	15	2.96	3.90	3.13
WIENERWT	5	2.43	2.47	2.45
	10	3.33	3.79	3.28
	15	3.83	3.94	3.63
NTFA	5	3.07	3.52	3.13
	10	4.42	3.79	3.62
	15	4.64	3.94	3.66

speech recognition systems. Consequently, an automatic noise estimation and separation algorithm is required. The results of the evaluation are shown in Table 2.

V. Conclusions

In this paper, we have presented a new speech enhancement algorithm that is well suited to three different noise environments. The new algorithm is based on calculating the ratio of the variance of the noisy power spectrum in the time-frequency bin to its normalized time-frequency average. We showed that appropriate choices of δ_t and δ_f produced enhanced speech signal. The proposed method resulted in significant reduction in noise while providing enhanced speech with lower residual noise and higher SNR and SIG scores than the conventional methods. It also degraded the input speech signal in highly nonstationary noisy environments. In particular, it produces an output with improved segment SNR and lower residual noise for white Gaussian noise. In the future, we plan to test the robustness of the proposed algorithm under diverse conditions including live signal environments.

Acknowledgement

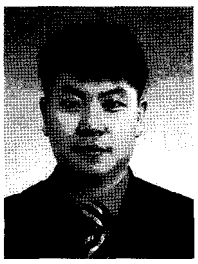
The present research has been conducted by the Research Grant of Kwangwoon University in 2007.

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. Acoust. Speech Signal Processing, 27 (2), 113-120, 1979.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square errror short-time spectral amplitude estimator", IEEE Trans. Acoust. Speech Signal Processing, 32(6), 1109-1121, 1984.
4. O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", IEEE Trans. Speech Audio Processing, 2(2), 346-349, 1994.
5. Y. Hu, "Subspace and multitaper methods for speech enhancement", Ph.D. dissertation, University of Texas at Dallas, 2003.
6. R.Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", IEEE Trans. Speech Audio Processing, 9(5), 504-512, 2001.
7. Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum", IEEE Trans. Speech Audio Processing, 12(1), 59-67.
8. R. Sundarrajan and C.L. Philipos, "A noise-estimation algorithm for highly nonstationary environments", Speech Communication, 48, 220-231, 2006.
9. I. Cohen, "Noise spectrum in adverse environments: improved minima controlled recursive averaging", IEEE Trans. SpeechAudio Processing, 11(5), 466-475, 2003.
10. I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator", IEEE Signal Processing Letters, 11(9), 725-728, 2004.
11. C. L. Philipos, "Speech Enhancement (Theory and Practice", 1st edition, CRC Press, Boca Raton, FL,2007.
12. J. L. Soo, "Speech enhancement based on Variance of Spectral Power in Time-Frequency for Speech Recognition in Non-stationary Noisy Environments", Ph.D. Thesis, Kwangwoon University, Seoul, Korea, December.
13. ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm", ITU-T Recommendation pp.835.
14. L. Lin, W. H. Holmes and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", Electronic Letters, 39(9), 754-755, 2003.
15. A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing", 2nd edition, Prentice Hall, Upper Saddle River, NJ, 1999.

[Profile]

• Soojeong Lee



1992.3~1997.2 : B.S. degree in Computer science, Korea National Open University
1997.3~2000.2 : M.S. degree in Computer Engineering, Kwangwoon University
2004.3~2008.2 : Ph.D. degree in Computer Engineering, Kwangwoon University
2007. 12~현재 : (주)에이팻 기술연구소 신기술 개발팀장

• Soonhyob Kim

B.S. degree in electronics engineering from the Ulsan University, Korea, 1970.3~1974.2
M.S. degree in electronics engineering from the Yonsei University, Korea, 1974.3~1976.2
Ph.D. degree in electronics engineering from the Yonsei University, Korea, 1976.3~1983.2
1979. 3~present Professor, Dept. of Computer Engineering, Kwangwoon University
1986. 8~1987. 8 Visiting professor, Dept. of Electrical & Computer Eng. Univ. of Texas at Austin
1998. 1~2000. 12 President, the Acoustical Society of Korea
2000. 10~2004.10 Chairman of the invitation committee ICSLP 2004
2001. 1~present President Emeritus, the Acoustical Society of Korea