THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.29, No.2E 2010. 6. pp. 86~99

Performance Analysis of a Class of Single Channel Speech Enhancement Algorithms for Automatic Speech Recognition

Myung-Suk Song*, Chang-Heon Lee*, Seok-Pil Lee**, Hong-Goo Kang*
*Dept. of Electrical and Electronic Engineering, Yonsei University, Seoul
**Broadcasting-Communication Convergence Research Center, Korea Electronics Technology Institute
(Received March 4, 2010; accepted April 14, 2010)

Abstract

This paper analyzes the performance of various single channel speech enhancement algorithms when they are applied to automatic speech recognition (ASR) systems as a preprocessor. The functional modules of speech enhancement systems are first divided into four major modules such as a gain estimator, a noise power spectrum estimator, a priori signal to noise ratio (SNR) estimator, and a speech absence probability (SAP) estimator. We investigate the relationship between speech recognition accuracy and the roles of each module. Simulation results show that the Wiener filter outperforms other gain functions such as minimum mean square error-short time spectral amplitude (MMSE-STSA) and minimum mean square error-log spectral amplitude (MMSE-LSA) estimators when a perfect noise estimator is applied. When the performance of the noise estimator degrades, however, MMSE methods including the decision directed module to estimate a priori SNR and the SAP estimation module helps to improve the performance of the enhancement algorithm for speech recognition systems.

Keywords: *Single channel speech enhancement, Speech recognition, Performance analysis*

I. Introduction

In the past decades, the performance of automatic speech recognition (ASR) systems has increased significantly, but they have not been broadly used for commercial purposes because performance is severely degraded in noisy environments [1]. One of the approaches to overcoming this problem is adopting preprocessing techniques such as speech enhancement or noise reduction modules [2-4]. If noise signals have relatively stationary characteristics compared to speech signals, a single channel speech enhancement technique is very effective in improving ASR performance. However, the effect of preprocessing techniques on ASR performance has not been fully understood as yet [5]. It has not been investigated which functional module among the various techniques is the most important to improving the speech recognition performance and how much or whether it affects performance.

Speech enhancement algorithms consist of four functional modules, namely noise power estimation, gain estimation, a priori SNR estimation and a speech absence or presence probability decision logic with soft-decision [6-9]. The noise power estimation is an essential component which decides the overall performance of the enhancement system, and has been developed based on the assumption

Corresponding author: Myung-Suk Song (earth112@dsp.yonsei.ac.kr)
B601 Department of Electrical and Electronic Eng. School of Engineering, Yonsei University 134 shinchondong seodaemoon-gu, 120-749, Seoul, Korea

of somewhat slowly varying noise environments. A commonly used approach for estimating the noise power spectrum is to average the noisy spectrum over speech absent regions. It generally detects speech absence regions using a hard-decision rule. To reduce the artifacts caused by misdetection, speech pause detection with a soft-decision rule using an a posteriori SNR has also been proposed [8]. However, detection reliability severely deteriorates for non-stationary noise environments. The algorithm for noise estimation based on minimum statistics obtains the noise spectrum by using minima values of the smoothed power spectrum of the noisy signal. Since it is sensitive to outliers, however, estimated noise components are generally biased and their variance can be about twice as large as that of a conventional noise estimator [10][11]. The minima controlled recursive averaging (MCRA) noise estimation algorithm improves the robustness of the minimum tracking with the simplicity of recursive averaging [1][12]. It averages past spectral power values with a smoothing parameter that is adjusted by the speech presence probability in each sub-band, and the speech presence probability is controlled by minima values of a smoothed periodogram.

The gain estimator is a module to determine the reduction level for each frequency bin of a noisy speech signal. The spectral subtraction algorithm, Wiener filtering and the maximum likelihood (ML) envelope estimation are well-known examples [2][9][13]. In the spectral subtraction algorithm, gain is derived from the variance of each signal spectral component using the square root of the maximum likelihood (ML) estimator [2][9]. The gain estimator based on Wiener filtering utilizes the optimal minimum mean-square error (MMSE) method of each signal spectral component [13]. Since spectral subtraction, the Wiener filter and the ML estimator are not optimal spectral amplitude estimators under imperfect noise estimator environments, several criteria combining both a priori and a posteriori SNR have been proposed to derive various short time spectral amplitude (STSA) estimators [2][9][14]. The minimum mean squared error-short time spectral amplitude (MMSE-STSA) estimator is an optimal estimation method as its solution is derived mathematically by minimizing the mean-square error of the cost function based on the Gaussian model under a statistical independence assumption [6]. The minimum mean squared error-log spectral amplitude (MMSE-LSA) estimator uses the criterion of minimizing the mean-square error of the logspectra [7]. The MMSE-LSA gain estimator is verified to be very efficient in reducing musical residual noise phenomena, because a distortion measure in the log spectrum domain is more suitable for speech perception.

A priori SNR estimation is needed to implement the MMSE estimator. The MMSE noise suppressor can be more effective if a nonlinear smoothing procedure is used to obtain more consistent estimates of a priori and a posteriori SNR which are used to control the gain function. And it is well known that the decision directed method is commonly used to decide the a priori SNR, which is also efficient in eliminating musical noise [15][16].

The gain functions also consider the uncertainty of speech presence or absence in real environments. It is well known that the perceptual quality of the enhanced speech signal is improved when the speech absence probability (SAP) is individually calculated in each frequency bin. Several key algorithms that utilize the signal-to-noise ratio (SNR) of each frequency bin have been proposed to estimate the speech absence probability [1][8][17][18][19]. An adaptation of a priori SAP using the SNR information is proposed to improve the performance of MMSE and LSA algorithms [8]. The algorithm recursively averages the index function determined by a hard decision rule based on the a posteriori SNR. Since the hard decision rule could not utilize the SNR information efficiently, however, misclassification of the speech activity could cause undesired artifacts. To further improve the performance, adaptive tracking and a soft decision for the a priori SAP are

needed. The method that combines three parameters, namely the local and global values of the speech absence probability and a soft decision on whether the current frame contains speech or not, for the a priori SAP estimation has been proposed [1] [17].

In this paper, we analyze the effect of each module of the speech enhancement algorithms to the performance of automatic speech recognizer systems. Actually, it is very difficult to independently analyze the independent contribution of each functional module to recognition performance because they are organically coupled with each other, so that it is hard to separate the role of each functional module. Therefore to observe the influence of a particular module to recognition performance, we need to fix the other functional modules. In other words, when we focus on analyzing the effects of the SAP estimator, the noise estimator and a priori SNR estimator modules should be fixed.

We compare performances of four gain functions, namely the Wiener, the MMSE-STSA estimator, the MMSE-LSA estimator, and the optimally modified log-spectral amplitude (OM-LSA) estimator in various SNR noise environments in terms of recognition rates. From the results, we investigate how the gain estimators affect speech recognition accuracy. Effects caused by the performance of the noise estimator are also analyzed. We assume perfect noise estimation first, and use a first-order recursive smoothing to simulate the degradation of the noise estimation accuracy. We investigate the effects of the a priori SNR estimator to the speech recognition accuracy when various noise estimators are applied. To evaluate the effect of speech absence probability to ASR, we select two methods: a method of fixing the SAP values for all frequency bins and an adaptive method of tracking the SAP values for each frequency bin continuously [8].

Experimental results show that the Wiener filter outperforms other gain functions such as MMSE-STSA, MMSE-LSA, and OM-LSA estimators when a perfect noise estimator is applied. When the noise estimator works improperly, however, the decision directed method to estimate a priori SNR and SAP estimation method helps to improve the performance of enhancement algorithms for speech recognition.

The organization of this paper is as follows. Section 2 introduces various single channel speech enhancement algorithms used for performance analysis in this paper. In Section 3, the effects of each module comprising the speech enhancement algorithm on speech recognition performance are investigated by simulations. Section 4 includes the experimental setup and results. Finally, Section 5 summarizes the contributions made in this paper.

2. Single Channel Speech Enhancement Algorithms

This chapter briefly introduces various single channel speech enhancement algorithms used in this paper. Let $X(k,l) = A(k,l)e^{j\alpha(k,l)}$ and D(k,1) denote the k-th coefficient of the discrete Fourier transform of the speech signal x(t) and uncorrelated additive noise signal d(t) at the l-th frame. Then, a coefficient of Fourier transform of the observed signal y(t) can be represented as

$$Y(k,l) = R(k,l)e^{j\theta(k,l)} = X(k,l) + D(k,l), \quad (1)$$

where R(k,1) and $\theta(k,l)$ represent the magnitude and phase of the observed signal y(t). In general, only the estimated magnitude is used for single channel speech enhancement since it is well known that the influence of the phase component is minimal [1] [6] [7]. Thus, the enhanced signal is obtained as follows,

$$\hat{X}(k,l) = \hat{A}(k,l)e^{j\theta(k,l)}. \quad (2)$$

The magnitude estimation $\hat{A}(k,l)$ is given by

$$\hat{A}(k,l) = G(k,l)R(k,l), \quad (3)$$

where G(k,l) denotes a gain function to be multiplied to the noisy signal for enhancement. Fig. 1 represents a basic block diagram of single channel speech enhancement algorithm.

2.1. Wiener Filter

The Wiener Filter corresponds to the criterion of minimizing the mean-square error of the best time domain fit to the speech waveform. Assuming that speech and noise signals obey normal distributions and are not correlated, the Wiener amplitude estimator in the frequency domain derives a gain function as [6] [9] [13] [22]:

$$G_{Wiener}(k,l)=\frac{\lambda_x(k,l)}{\lambda_d(k,l)+\lambda_x(k,l)}=\frac{\xi(k,l)}{1+\xi(k,l)}. \quad (4)$$

In Eq.(4), $\xi(k,l)$ corresponds to the a priori SNR defined by $\xi(k,l)=\lambda_x(k,l)/\lambda_d(k,l)$. $\lambda_x(k,l)$ and $\lambda_d(k,l)$ denote the variances of speech and noise, respectively. Since the performance of the Wiener filtering is related to a priori SNR estimation, the noise power spectral density (PSD) estimation module is the most important.

2.2. MMSE-STSA

Since the Wiener amplitude estimator is derived from the optimal minimum mean-square error signal spectral estimator, it is not an optimal spectral amplitude estimator under the assumed statistical model and criterion. MMSE-STSA estimates the spectral amplitude using the statistical model that utilizes asymptotic statistical properties of the Fourier expansion coefficients. Specifically, we assume

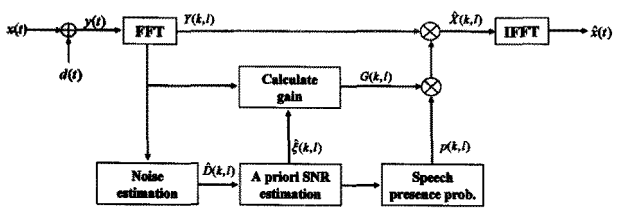


Fig. 1. Basic concept of single channel speech enhancement with four major modules.

that the Fourier expansion coefficients can be modeled as statistically independent zero-mean Gaussian random variables [6].

The MMSE estimator $\hat{A}(k,l)$ of A(k,l) and the gain of the MMSE spectral amplitude estimator which minimizes the mean square error of the A(k,l) are obtained as follows.

$$\hat{A}(k,l)=E\{A(k,l)\,|\,Y(k,l)\}, \quad (5)$$

$$G_{MMSE}(k,l)=\Gamma(1.5)\frac{\sqrt{\upsilon(k,l)}}{\gamma(k,l)}\exp\left(-\frac{\upsilon(k,l)}{2}\right)$$
$$\left[(1+\upsilon(k,l))I_0\left(\frac{\upsilon(k,l)}{2}\right)+\upsilon(k,l)I_1\left(\frac{\upsilon(k,l)}{2}\right)\right], \quad (6)$$

where $\Gamma(\cdot)$ denotes the gamma function, with $\Gamma(1.5)=\sqrt{\pi}/2$ [23]. $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order, respectively. $\upsilon(k,l)$ and the a posteriori SNR $\gamma(k,l)$ are defined by $\upsilon(k,l)=\xi(k,l)\gamma(k,l)/(1+\xi(k,l))$ and $\gamma(k,l)=R(k,l)^2/\lambda_d(k,l)$, respectively. The MMSE estimator is nearly equivalent to the Wiener estimator at high SNR. On the other hand, the MMSE estimator yields significantly lower mean square error and bias under low SNR environments [6].

2.3. MMSE-LSA

While the MMSE estimator is mathematically tractable, it is not the most meaningful one in the perceptual sense. It is well known that a distortion measure based on the mean-square error of the log-spectra is more suitable for speech signal perception [7]. Such a distortion measure is therefore extensively used for speech analysis and recognition.

By minimizing the distortion of Eq.(7), the amplitude estimator of A(k,l) can be obtained as Eq. (8).

$$D_{MSE}=E\{(\log A(k,l)-\log\hat{A}(k,l))^2\} \quad (7)$$

$$\hat{A}(k,l) = \exp\{E[\ln A(k,l) \mid Y(k,l)]\} \quad (8)$$

Utilizing the assumed Gaussian model, the minimum mean-square error log spectral amplitude (MMSE-LSA) estimator is defined by

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right\}. \quad (9)$$

It is interesting to note that the LSA estimator gain function $G_{LSA}(k,l)$ always gives a lower gain than the MMSE gain function $G_{MMSE}(k,l)$. In addition LSA gain functions also converge to the Wiener solution at high SNR [7].

2.4. OM-LSA

In the MMSE-LSA estimator, the gain function should be modified by considering the uncertainty of speech presence in real environments, which requires the computation of the speech absence probability (SAP) [17]. Applying the hypothesis of speech presence (H_1) and absence (H_0) to the MMSE-LSA estimator, the modified amplitude estimator can be obtained by Eq. (10).

$$\begin{aligned}\hat{A}(k,l) &= \exp\{E[\ln A(k,l) | Y(k,l)]\} \\ &= \exp\{E[\ln A(k,l) \mid Y(k,l), H_1]\}^{p(k,l)} \\ &\quad \times \exp\{E[\ln A(k,l) \mid Y(k,l), H_0]\}^{(1-p(k,l))}. \end{aligned} \quad (10)$$

The conditional probability of speech presence p(k,l) and the optimally modified LSA(OM-LSA) gain G_{OM-LSA} can be derived as

$$p(k,l) = P(H_1 | Y(k,l)) = \left\{1 + \frac{(1-q(k,l))}{q(k,l)}(1+\xi(k,l))e^{-v(k,l)}\right\}^{-1}, \quad (11)$$

$$G_{OM-LSA}(k,l) = \{G_{LSA}(k,l)\}^{p(k,l)} \times G_{\min}^{1-p(k,l)}, \quad (12)$$

where q(k,l) and G_{min} represent the a priori speech absence probability (SAP) and minimum threshold of gain for a non-speech component. When speech is absent, the gain is constrained to be larger than the threshold G_{min} determined by a subjective criterion for noise naturalness. Since the gain modification to utilize the uncertainty of speech presence is very efficient in improving the perceptual quality of the MMSE-LSA enhancement system [1] [8] [17], the a priori SAP is a key parameter of the gain modifier to adjust the level of noise suppression. The SAP can be estimated as either a fixed value q for all frequency bins or individually calculated values q(k,l) in each frequency bin [8]. Those SAP estimation algorithms will be explained precisely in Section 3.4. The enhancement system using uncertainty of speech presence additionally needs an estimator for obtaining speech presence probability.

3. performance analysis

The single channel speech enhancement algorithm considered in this paper consists of four major modules, namely a gain estimator, a noise power spectrum estimator, a priori SNR estimator, and an estimator of speech absence or presence probability for soft decision [8] [17]. Unfortunately, there have been no studies analyzing relationships between the major modules and speech recognition performance. This chapter describes effects of each module in a speech enhancement algorithm to the speech recognition performance.

First, Section 3.1 compares effects of gain functions on speech recognition accuracy. The effectiveness of several gain estimators such as the Wiener filter, MMSE estimator, and LSA estimator is compared using a cepstral distance measure. The influence of noise power spectrum estimators on recognition accuracy is analyzed in Section 3.2, based on measurements of cepstal distances of enhanced signals by different gain functions as the performance of noise estimator becomes degraded. In Section 3.3, effects of the decision directed

method that estimates a priori SNR are analyzed. Finally, Section 3.4 includes analysis about the a priori speech absence probability (SAP). Two simple techniques to estimate SAP are introduced to analyze the influence of the SAP estimator on speech recognition performance when different noise estimators are coupled.

16 kHz sampled clean speech signals from the TIMIT database [21] with 630 speakers are used for our experiments. The speech is corrupted by three different noise types, such as white, babble, and destroyer engine noise taken from the Noisex92 database [24], with various SNRs. White noise having uniform spectrum in all frequency regions is commonly used for various experiments in speech signal processing applications. Babble noise, which many people speak simultaneously, can be regarded as more non-stationary and more speech-like than white noise. The destroyer engine noise generated from a defective machine has high energy in several frequency bins and can be considered as a more mechanical noise. The simulation results in tables and figures of this paper are averaged values of the results obtained from 3 different noise environments. The speech enhancement processing is conducted using 512 point FFT with Hanning window and 50 % overlap.

3.1. Gain Estimator

Comparisons of the performance of all gain functions at once is not possible because parameters needed for calculating each gain function are different in each case. For example, both the MMSE and the LSA gain functions need estimated noise spectrum and estimated a priori SNR, while the Wiener filter needs only the estimated noise spectrum (See Fig. 1). In addition, since the estimation accuracy of parameters for gain functions are deeply related to other functional modules, the other modules need to be fixed when changes caused by the gain function is to be observed. Thus, in our experiments, it is assumed that the noise estimation is perfectly performed and the decision-directed method for a priori SNR estimation is not considered.

In order to quantify contributions of each module to speech recognition performance, the mel frequency cepstral distance (MFCD) between the enhanced speech and the clean speech is measured. For the l-th frame, the MFCD is computed as

$$MFCD(l)=\frac{1}{K}\sqrt{\sum_{i=1}^{K}[C_C(i,l)-C_E(i,l)]^2}, \tag{13}$$

where $C_C(i,l)$, $C_E(i,l)$ and K are the i-th mel frequency cepstral coefficients of the clean and the enhanced speech signal at the l-th frame, and the number of cepstral coefficients, respectively. The distortion of the mel frequency cepstral coefficient (MFCC) is a common and useful parameter in determining the performance of ASR systems, and thus a good measure to represent how well noisy speech is enhanced for ASR systems.

The MFCDs obtained from each algorithm under various noise types and input SNR environments are shown in Table 1. The MFCDs are calculated under 3types of noise environments, such as white, babble, and destroyer engine noise. The averaged values are shown in Table 1. Since perfect noise estimation is assumed and the decision directed method for estimating a priori SNR is not employed, the estimated a priori SNR $\hat{\xi}(k,l)$ is replaced by the instantaneous SNR $(\gamma(k,l)-1)$, where $\gamma(k,l)$ denotes a posteriori SNR. Actually, the instantaneous SNR $(\gamma(k,l)-1)$ represents the perfect a priori SNR $\xi(k,l)$ in this case.

Table 1. Averaged MFCDs of signals enhanced by various enhancement algorithms with perfect noise estimation in various noise environments.

Algorithm		0 dB	5 dB	10 dB	20 dB
Wiener		1.852	1.690	1.476	1.029
MMSE		1.917	1.772	1.599	1.201
LSA		1.8779	1.723	1.542	1.132
OM-LSA	Fixed	1.996	1.798	1.579	1.148
	Malar	1.955	1.811	1.594	1.156

As shown in Table 1, the Wiener filter with good noise estimation shows the best performance. These results seem reasonable, since the Wiener filter is an optimal solution when it has perfect knowledge of the noise components. Besides, as the Wiener filter utilizes only the instantaneous SNR instead of a priori SNR estimated by the decision-directed method that is smoothing the input spectrum, it minimizes the distortion which can be caused by spectrum smoothing when the noise spectrum is known. Hence, the Wiener filter with good noise estimation can be regarded as an upper bound for all surveyed enhancement algorithms adapted as a preprocessor for the ASR system.

The results show that the MFCD of the MMSE estimator is bigger than those of LSA. Indeed, the performance of LSA is always superior to the MMSE for all following experiment categories. This is caused by the fact that both the LSA and the MFCD operate on the log domain. Since the LSA's criterion is more closely related to the cepstrum, it is possible to directly minimize the error in the cepstrum domain and get even better results.

It is notable that the performance in the babble noise environment is worse than that of the white noise environment. This is because babble noise introduces distortion in the frequency regions where most speech components exist. MFCDs from the destroyer engine noise are bigger than others, since large energy components of noisy signals are not completely suppressed and still remain even after enhancement processing. The order in which each algorithm performs is still preserved, however, and the Wiener filter shows the best performance, and LSA outperforms MMSE irrespective of noise types.

3.2. Noise Spectrum Estimator

The performance of single channel speech enhancement algorithms mainly depends on the efficiency of noise power spectrum estimator. This subsection focuses on analyzing effects caused by the performance of the noise estimator. Actually, the true noise power spectral density (PSD) is unattainable even we are able to use any reliable noise estimator. The difference between the true noise PSD and the PSD estimated by the noise estimator needs to be controlled. To control the amount of the difference, the noise estimator is artificially generated for simulations as

$$\hat{\lambda}_d(k,l) = \alpha_N \hat{\lambda}_d(k,l-1) + (1-\alpha_N)\lambda_d(k,l), \tag{14}$$

where $\hat{\lambda}_d(k,l)$, $\lambda_d(k,l)$ and α_N are the estimated noise power, the true noise power of the k-th frequency bin in the l-th frame, and a smoothing factor, respectively. The noise estimator with $\alpha_N = 0$ means perfect estimation. As the smoothing factor α_N increases close to 1, the performance of the noise estimator becomes degraded. Since the worst case of α_N is experimentally impossible to realize, we replace it with the exceptional estimator, which uses averaged noisy spectrum during non-speech period detected by a perfect VAD (Voice Activity Detector).

Fig. 2 shows normalized root mean squared error (RMSE) values of the estimated noise power spectrum as the smoothing factor α_N varies. The RMSE values of the estimated noise spectrum are obtained by

$$RMSE_N = \frac{1}{LK}\sum_{l=0}^{L-1}\sum_{k=0}^{K-1}\sqrt{(\hat{\lambda}_d(k,l)-\lambda_d(k,l))^2}, \tag{15}$$

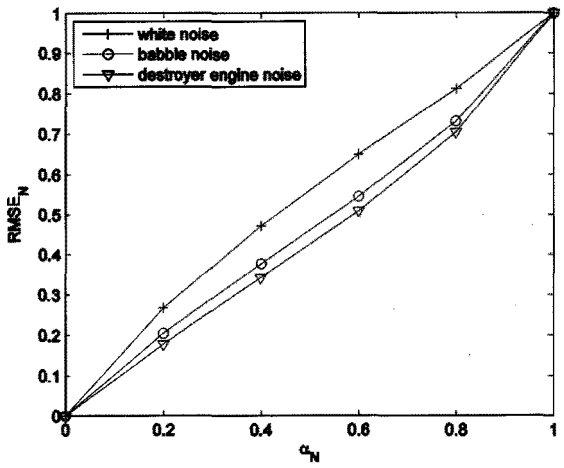


Fig. 2. The normalized $RMSE_N$ of estimated noise spectrum.

where L and K denote the number of frames and number of frequency bins, respectively. The normalized RMSE values in Fig. 2 are calculated by normalizing the RMSE value in Eq.(15) by its maximum value (when $\alpha_N=1$). Fig. 2 shows that the noise estimator used in the simulations, including the exceptional estimator, represents linear relationship to the smoothing factor α_N in the RMSE sense.

Table 2 shows the MFCDs of enhanced speech signals by each algorithm as the smoothing factor α_N varies from zero to one. The experiments are conducted in 0 dB white, babble, and destroyer engine noise environments, and the averaged MFCD results are shown in Table 2. The decision directed method is not employed.

Results show that the MFCDs of all surveyed algorithms are monotonically increasing as the performance of the noise estimator becomes worse. The consistent increase of all MFCDs confirms that noise estimation significantly affects the performance of enhancement systems for speech recognition. The MFCDs of the enhanced signal by the Wiener filter especially increase faster than others. The Wiener filter works best with a perfect noise estimator and does worst with the noise estimator using VAD among the surveyed algorithms. The MMSE and LSA gain functions show better performance than the Wiener filter when noise estimator operates coarsely, because of their differences on gain curves [6]. Table 2 also confirms that LSA always outperforms MMSE, because the LSA's criterion is more closely related to the cepstrum, it directly minimizes the error in the cepstrum domain.

Table 2. Averaged MFCDs of signals enhanced by various enhancement algorithms varying the smoothing factor of noise estimator α_N in various noise environments.

Algorithm		0.0	0.2	0.4	0.6	0.8	VAD
Wiener		1.852	2.674	3.004	3.242	3.447	3.612
MMSE		1.917	2.975	3.164	3.29	3.415	3.537
LSA		1.878	2.941	3.150	3.272	3.383	3.524
OM-LSA	Fixed	1.996	2.914	3.092	3.213	3.372	3.524
	Malar	1.955	2.927	3.074	3.181	3.353	3.518

3.3. A priori SNR Estimator

The estimation of a priori SNR is also an essential part for noise suppression such as in the MMSE and the LSA estimator. The decision directed method is one of the most commonly used estimators for a priori SNR and shown to be essential in eliminating musical noise phenomena [15]. We survey effects of the decision directed method on speech recognition performance. The decision directed rule can be obtained by recursively averaging a priori SNR of the previous frame and the instantaneous SNR of the current frame [6].

$$\hat{\xi}(k,l)=\alpha_d G^2(k,l-1)\gamma(k,l-1)+(1-\alpha_d)\max[\{\gamma(k,l)-1\},0], \tag{16}$$

where $\gamma(k,l)$, α_d, and $G^2(k,l-1)\gamma(k,l-1)$ represent a posteriori SNR at the k-th frequency bin in the l-th frame, a smoothing factor, and the a priori SNR resulting from the processing of the previous frame, respectively. $\alpha_d=0$ means that a priori SNR $\hat{\xi}(k,l)$ is estimated as instantaneous SNR $\gamma(k,l)-1$. As the smoothing factor α_d increases near to 1, estimated a priori SNR $\hat{\xi}(k,l)$ approaches the estimated a priori SNR of the previous frame $\hat{\xi}(k,l-1)$.

Fig. 3 shows MFCDs of signals enhanced by using different noise estimators as the smoothing factor α_d

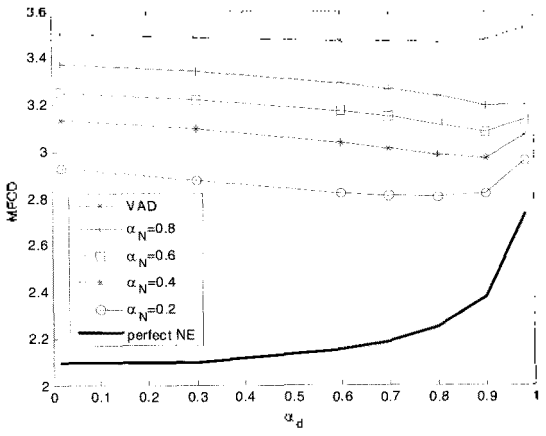


Fig. 3. Averaged MFCDs of enhanced signals using LSA gain function according to α_d when different noise estimators are used under various noise environments.

varies. The LSA algorithm is applied to the 0 dB white, babble, and destroyer engine noise environments. The averaged values are shown in Fig. 3. Only the MFCD of the enhanced signal under perfect noise estimation monotonically increases as the smoothing factor α_d varies. That is, if the noise power spectrum is known precisely, a priori SNR $\hat{\xi}(k,l)$ can be estimated as instantaneous SNR $\gamma(k,l)-1$. Since the decision directed method estimates a priori SNR by averaging the a priori SNR of the previous frame and the a posteriori SNR of the current frame recursively, it results in degraded time resolution so that phones of short duration are not recognized well. However, the MFCDs under worse noise estimation environments decrease as the smoothing factor α_d increases. Those achieve minimum values around $\alpha_d = 0.9$. That is, when the noise estimator works improperly, as is general in real applications, enhancement algorithms for ASR systems need to be supported by the decision directed method. Even though the decision directed method has been developed to improve perceptual quality, it helps to improve speech recognition performance when the noise power spectrum is improperly estimated.

3.4. SAP Estimator

The OM-LSA gain function additionally needs the estimated speech absence probability SAP compared to the original LSA algorithm (See Fig. 1). Several SAP estimation algorithms have been developed [1] [8] [17] [18]. However, this paper focuses on two most commonly used methods such as constant SAP, q, and the method proposed by Malah [8].

The simplest idea for SAP estimation is to apply a fixed SAP value, $q=0.2$, to all frequency bins [6]. However, speech signals can be considered as having quasi-harmonic and non-stationary characteristics. Furthermore, speech energy may not be present in all frequency bins. Thus, it seems more suitable to apply a different value to each frequency bin for each frame, instead of assigning the same value of q to all frequency bins.

One of the estimation methods to obtain distinct values of q for each frequency bin in each frame is proposed by Malah in [8]. This method uses a recursive averaging of index function $\mathbf{I(k,l)}$, which is a hard decision rule of speech presence based on the a posteriori SNR, $\gamma(k,l)$, and represents speech absence likelihood in each frequency bin. (i.e. $\mathbf{I(k,l)}=0$, if $\gamma(k,l) \geq \gamma_{th}$, and $\mathbf{I(k,l)}=1$, otherwise). The estimated a priori SAP $\hat{q}(k,l)$ is as follows

$$\hat{q}(k,l) = \alpha_q \hat{q}(k,l-1) + (1-\alpha_q) I(k,l) \tag{17}$$

This method, estimating distinct values of SAP for different frequency bins which are tracked in time, copes reasonably well in non-stationary speech regions and frequency bins where speech does not exist.

Table 1 shows the MFCDs of signals enhanced by the OM-LSAs with a perfect noise estimator. It is worthy to note that the MFCDs of the OM-LSAs are higher than those of other algorithms. This could be due to artifacts caused by the stricter than necessary assumption, such as assuming perfect knowledge of the noise spectrum and the a priori SNR. However, it reveals that the SAP estimator developed as a modification to improve perceptual quality for human listeners could cause adverse effects in the performance of an ASR system, especially if the noise estimator works very well.

At the bottom of Table 2, the MFCDs of signals enhanced by the OM-LSAs with different noise estimators are shown. The OM-LSA gain estimators are more robust than other gain functions under unreliable noise estimation environments, such as $\alpha_N \geq 0.6$. It is because the OM-LSA estimator takes into account the possibilities of speech absence for the non-speech region and weak speech components while the MMSE and the LSA assume that the speech signal always exists for every spectral component. The SAP estimator helps to improve performance of enhancement by making up for

mistakes of the coarsely working noise estimator.

4. Experimental results and discussion

This Section provides experimental results and discussions. The mean opinion score (MOS) test and the speech recognition test are conducted. The MOS test is used to show the relationship between perceptual quality and speech recognition performance under good noise estimation environments. The recognition performance is analyzed under several conditions, such as when the noise estimator works perfectly or its performance gets degraded, and whether the effect of the decision directed method is considered or not.

4.1. Experimental Environments

The MOS test is used as a perceptual quality measurement to estimate the performance of various enhancement algorithms. A total of 20 listeners are asked to score 1–5 points for each enhanced speech sample. The hidden Markov toolkit (HTK) is used for speech recognition simulations [20]. The HTK recognizer is trained by using the clean speech database and tested by using both noisy and enhanced speech samples. Noisy speech is generated using two different noise types, the white and babble noise taken from the Noisex92 database, with different SNRs. The TIMIT speech database [21] with 630 speakers is used for simulations. For phone recognition, the 61 TIMIT phones are mapped to a reduced set of 39 phones in training and testing procedures [25], and results are reported on the reduced set. The analysis of recognition results is performed for only 18 vowels headed by /a/, /e/, /i/, /o/, and /u/. The recognition results of other phones are excluded from the analysis in order to handle the data that shows more analyzable results, because the weak energy phones such as fricatives are minimally enhanced by any single channel enhancement algorithm, especially in low SNR environments. The recognition rates of most of consonants, especially weak-energy consonant phonemes headed by /b/, /d/, /g/, /p/, /t/, and /k/, are much lower than the analyzable level in low SNR. Moreover, since there are little differences between the recognition rates for low energy components, the comparison of results for vowels is enough to evaluate the performance. The recognition rate is calculated by summing the correctly recognized percentage of the vowel phonemes.

4.2. Experimental Results and Discussion

Table 3 shows the averaged MOS test scores of enhanced signals obtained from each enhancement algorithm with a reliable noise estimator ($\alpha_N = 0.2$) under various noise environments such as white, babble, and destroyer engine noise. The OM-LSA algorithm with the SAP estimator used in Malah's method shows the best performance in terms of the perceptual quality. The LSA estimator shows better performance than the MMSE estimator, since human hearing characteristics is better incorporated using a logarithmical magnitude. The Wiener filter shows the worst perceptual quality among all tested algorithms due to the musical noise, though it is theoretically the optimal solution for a mean square error (MSE) criterion. The results show that considering human acoustic characteristics and speech presence uncertainty may lead to enhanced perceptual quality.

Table 4 shows the speech recognition rates of enhanced signals when perfect noise estimation is used without the decision-directed method as Eq. (16) in various noise environments. The Wiener filter shows the best performance when it has a

Table 3. Averaged MOS test scores in various noise environments.

Algorithm		0 dB	5 dB	10 dB	20 dB
Wiener		1.91	2.32	2.74	3.93
MMSE		2.01	3.09	2.89	3.87
LSA		2.37	2.73	3.27	4.19
OM-LSA	Fixed	2.55	3.03	3.52	4.36
	Malar	2.73	3.21	3.57	4.40

perfect knowledge of noise components, as it utilizes only the instantaneous SNR so that it minimizes the distortion which can be caused by spectrum smoothing when the noise spectrum is known, while the others use a priori SNR estimated by the decision-directed method that is smoothing the input spectrum. Although the Wiener filter is rarely used for enhancement applications due to its perceptual quality, it can work well for a recognition system with a good noise estimator. It is verified that the MMSE-STSA and MMSE-LSA show similar performances to the Wiener filter in high SNR environments [6][7]. The recognition rate of MMSE degrades somewhat more rapidly than that of LSA in low SNR environments. It is related to the observation that the MMSE gain function suppresses weak speech or non-speech components less aggressively than the LSA estimator. The results also show the effects of the SAP estimation on the ASR performance. As shown in the results, the perceptual improvement does not always match the recognition performance. The OM-LSA algorithms using a dynamic SAP estimator show worse recognition performance. To consider the speech uncertainty improving perceptual quality can adversely affect speech recognition performance due to unnecessary speech distortions under good noise estimation environments.

Fig. 4 represents the recognition rate of the algorithms as the averaging factor of the noise estimator varies. We may assume that the performance of the estimator degrades as the smoothing factor approaches one. Experiments are performed without using the decision-directed method in white, babble and destroyer engine noise (0 dB SNR) environments and the averaged values are shown in Fig. 4. The results show that the recognition rates of all surveyed algorithms monotonically decrease as the performance of the noise estimator becomes worse. Especially, the recognition accuracy of enhanced signals by the Wiener filter degrades faster than the others, while it outperforms other algorithms in environments where the noise estimator works reliably, such as in the cases where $\alpha_N \leq 0.3$. Results also show that the algorithms using SAP estimators such as Malah's lead to improvement of the recognition performance under unreliable noise estimation environments using $\alpha_N \geq 0.4$, while they yield worse performance in other regions. In other words, the Wiener filter with very good noise estimation is far better than any other algorithm for speech recognition, and the algorithms using SAP estimation can improve the recognition performances under unreliable noise estimation environments.

Fig. 5 depicts the recognition rates of signals enhanced using different noise estimators, varying the smoothing factor α_d of the decision directed method. Experiments are performed with the LSA gain estimator in white, babble, and destroyer engine noise (0 dB) environments and the averaged values are shown in Fig. 5. Results show that recognition rates of the signal enhanced by a perfect noise

Table 4. Averaged recognition rate (%) with a perfect noise estimator in various noise environments.

SNR		0 dB	5 dB	10 dB	20 dB
clean		68.14			
Noisy		33.88	39.45	44.72	56.78
Wiener		58.02	61.53	63.91	66.42
MMSE		55.59	59.36	62.64	66.6
LSA		56.43	60.12	63.10	66.62
OM-LSA	Fixed	55.23	59.59	62.75	65.97
	Malar	55.75	59.69	62.92	66.28

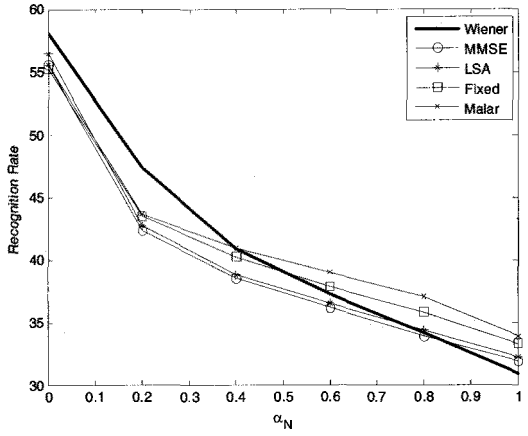


Fig. 4. Averaged recognition rates of enhanced signals under various noise environments.

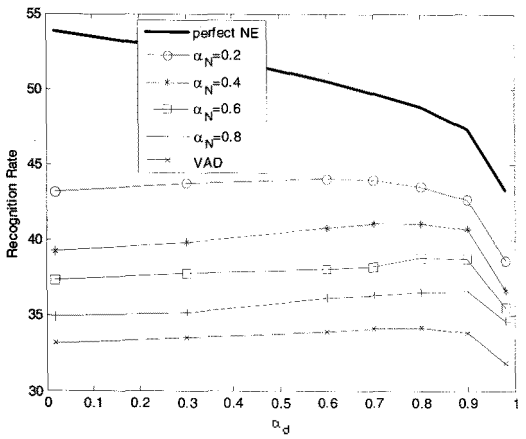


Fig. 5. Averaged recognition rates of enhanced signals with varying the α_d under various noise environments.

estimator is monotonically decreasing as α_d increases and it has a maximum value when α_d=0. In other words, when noise components are perfectly known, any smoothing in the time domain leads to adverse effects in speech recognition.

On the other hand, the recognition rates under worse noise estimation environments increase as the smoothing factor α_d increases. The recognition rates show maximum values around α_d=0.8. That is, when the noise estimator works improperly, the decision directed method can improve the enhancement algorithm in terms of the speech recognition rate.

5. Conclusion

Throughout this paper, most of our attention was focused on describing the relationship between a speech recognizer and speech enhancement systems by investigating several enhancement modules. We reviewed various single channel speech enhancement algorithms. The Wiener filter, MMSE, LSA, and OM-LSA gain estimators were introduced. The noise estimator, a priori SNR estimator, and SAP estimator of single channel speech enhancement systems were described.

The effects of each module on speech recognition performance were investigated by simulations. We compared effects of several gain functions to speech recognition performance by using a cepstral distance measure. The effects of the performance of the noise estimator were analyzed under environments where the performance of noise estimator was degraded. We also investigated effects of the decision directed method in estimating the a priori SNR. The influence of the SAP estimator on speech recognition was analyzed when different noise estimators were coupled.

The MOS test was used to show the relationship between perceptual quality and speech recognition performance. The recognition performance was analyzed under several assumptions, such as when the noise estimator was assumed to work perfectly or when performance was degraded, and whether the effect of the decision directed method is considered or not. In simulation results, the Wiener filter shows better performance than other gain functions such as MMSE and LSA estimators under a perfect noise estimator environments. When the performance of the noise estimator becomes degraded, however, the decision directed method to estimate a priori SNR and SAP estimation method helped to improve performance of enhancement algorithms for speech recognition.

References

1. I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403-2418, Oct. 2001.
2. Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, Apr. 1979.
3. H. K. Kim, R. C. Rose, and H. G. Kang, "Acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments" , *EUROSPEECH-proceeding 2001*, vol. 1, pp. 421-424, Sep. 2001.
4. N. W. D. Evans, J. S. D Mason, W. M. Liu and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction" , *Proc. Int. Conf. Acoustics, Speech, Signal Processing 2006*, pp. 145-148, 2006.
5. M. S. Song, C. H. Lee, and H. G. Kang, "Performance analysis of various single channel speech enhancement algorithms for automatic speech recognition" , *Interspeech 2006 ICSLP*, pp. 1451-1454, Sep. 2006.

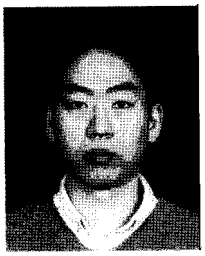
6. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp.1109-1121, Dec. 1984.
7. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
8. D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. Int Conf. Acoustics, Speech, Signal Processing 1999*, pp.789-792, 1999.
9. R. J. Mcaulay, "Speech enhancement using a soft-decision noise suppression filter" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, Apr. 1980.
10. R. Martin, "Spectral subtraction based on minimum statistics" , *Proceedings of the Seventh European Signal Processing Conference, EUSIPCO 94, Edinburgh, Scotland*, 13-16, pp. 1182-1185, Sep. 1994.
11. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 9, pp.504-512, Jul. 2001.
12. I. Cohen, "Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 11, no. 5, Sep. 2003.
13. H. L. Van Trees, *Detection, Estimation and Modulation Theory, part I*., New York: Wiley, 1968, pp. 54-56, 198-206, 205-207.
14. C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 13, no. 4, Jul. 2005.
15. O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor" ,*IEEE transaction on Speech and Audio processing*, vol. 2, no. 2, Apr. 1994.
16. I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation" , *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 13, no.5, Sep. 2005.
17. I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal processing letters*, vol. 9, no 4, pp.113-116, Apr. 2002.
18. M. S. Choi and H. G. Kang, "An improved estimation of a priori speech absence probability for speech enhancement : In perspective of speech perception" , *Proc. Int. Conf. Acoustics, Speech, Signal Processing 2005*, pp. 1117-1120, 2005.
19. Y. Hu and P.C.Loizou, "Subjective comparison of speech enhancement algorithms" , *Proc. Int. Conf. Acoustics, Speech, Signal Processing 2006*, pp. 153-156, 2006.
20. S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, V. Valtchev, P. Woodland, "The HTK Book," copyright 1995-1999 Microsolt Corporation, copyright 2001-2002 Cambridge University Engineering Department.
21. J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: and acoustic phonetic continuous speech database, National Institute of Standards and technology (NIST), Gaithersburg, Maryland, (prototype as of December 1988).
22. J. Chen, J. Benesy, Y. Huang, and S. Doclo, " New insights into the noise reduction wiener "Iter" , *IEEE transaction on Audio, Speech, and Language processing*, vol. 14, no. 4, Jul. 2006
23. D. Middelton, Introduction to Statistical Communication Theory, New York: McGraw-Hill, 1960, ch.7, appendix 1.
24. A. Varga, H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems" , *Speech Commun.* vol. 12, no. 3, pp.247-251, Jul. 1993.
25. K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, Nov. 1989.

[Profile]

• **Myung-Suk Song**

1998-2005: Dept. of Electrical Engineering, Yonsei University (B.S.)
2005-2007: Dept. of Electrical Engineering, Yonsei University (M.S.)
2007-Present: Dept. of Electrical Engineering, Yonsei University (Ph.D.)
Research: Digital Signal Processing, Speech / Audio Signal Processing, Speech Enhancement, Speech Recognition, 3-D Audio

• **Chang-Heon Lee**

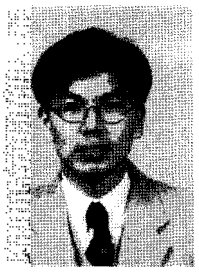


1996-2003: Dept. of Electrical Engineering, Yonsei University (B.S.)
2003-2005: Dept. of Electrical Engineering, Yonsei University (M.S.)
2005-Present: Dept. of Electrical Engineering, Yonsei University (Ph.D.)
Research: Digital Signal Processing, Speech Signal Processing, Speech Coding, Speech Enhancement

• **Seok-Pil Lee**

1986-1990: Dept. of Electrical Engineering, Yonsei University (B.S.)
1990-1992: Dept. of Electrical Engineering, Yonsei University (M.S.)
1992-1997: Dept. of Electrical Engineering, Yonsei University (Ph.D.)
1997-2002: Senior Research Engineer, Daewoo Electronics Corp.,Korea
2002-Present: Director, Digital Medua Research Center, Korea Electronics Technology Institute (KETI), Korea
Research: Digital TV, Personalized Service, IPTV

• Hong-Goo Kang



1985-1989: Dept. of Electrical Engineering, Yonsei University (B.S.)
1989-1991: Dept. of Electrical Engineering, Yonsei University (M.S.)
1991-1995: Dept. of Electrical Engineering, Yonsei University (Ph.D.)
1996-2002: Senior Technical Staff Member, AT&T Labs-Research
2002-2005: Assistant Professor, Yonsei University
2005-Present: Associated Professor, Yonsei University
Research: Speech Signal Processing, Adaptive Digital Signal Processing and Real-time Implementation