Performance Analysis of Various Single Channel Speech Enhancement Algorithms for Automatic Speech Recognition

Myung-Suk Song, Chang-Heon Lee, and Hong-Goo Kang

DSP Lab, Yonsei University, Korea

[earth112, leech, and hgkang]@dsp.yonsei.ac.kr

Abstract

This paper analyzes the performance of various single channel speech enhancement systems when they are applied to automatic speech recognition (ASR) systems as a preprocessor. Until now the researches on speech enhancement algorithms have focused on improving the perceptual quality of speech signal. However, it has not been verified yet whether the improvements of the perceptual quality also increase the speech recognition rate. By investigating several enhancement modules designed for improving the perceptual quality, we analyze the relationship between a speech recognizer and speech enhancement systems. Simulation results show that the decision-directed method and speech absence probability (SAP) estimation proposed for improving perceptual quality influence adverse effects to the speech recognition performance.

Index Terms: speech enhancement, speech recognition, HTK

1. Introduction

For the past decades the performance of automatic speech recognition (ASR) systems increases significantly when it operates in high signal-to-noise ratio (SNR) conditions. However, it has not been commonly used for commercial purpose yet because the performance is severely degraded in noisy environments [1]. One of the good approaches to overcome the problem is adopting pre-processing techniques such as speech enhancement or noise reduction. If the noise is varying relatively stationary compared to speech signals, a single channel speech enhancement technique is very effective to improve the ASR performance. However, the impact of the pre-processing techniques to ASR has not been deeply understood particularly the relative effectiveness of each module. In other words it has never been studied which functional module is the most important to improve the speech recognition performance and how much or whether it affects the performance.

Speech enhancement algorithms generally consist of three functional modules such as noise power estimation, gain estimation and determination of speech absence or presence probability for soft-decision [2][3]. The noise power estimation that is an essential component to decide the overall performance of the enhancement system has been developed based on the assumption of a slowly varying noise environment. A commonly used approach for estimating the noise power spectrum is to average the noisy signal over speech absent regions. For non-stationary noise environments an algorithm based on minimum statistics and minima controlled recursive averaging (MCRA) approach have been proposed[1][4].

The gain estimator is a module to determine the reduction level to each frequency bins of noisy speech signal. In addition to Wiener approach, several criteria with combining both *A priori* and *A posteriori* SNR have been proposed to derive various gain estimators. The minimum mean squared error-short time spectral amplitude (MMSE-STSA) estimator and the minimum mean squared error-log spectral amplitude (MMSE-LSA) estimator are typical examples [2][5]. The gain functions need to be modified to consider the uncertainty of speech presence or absence in real environment. It is well known that the perceptual quality of enhanced speech signal is improved when speech absence probability (SAP) is individually calculated in each frequency bins [3]. Several key algorithms that utilizes signal-to-noise ratio (SNR) of each frequency bin have been proposed to estimate the speech absence probability [1][3][6]. Algorithms such as gain estimator and SAP estimator have been developed to improve the perceptual quality, especially for speech coding. The effectiveness of the functional modules in a view of perceptual quality can be verified by a mean opinion score (MOS) test. The test provides a starting point of our paper to analyze the relationship between the perceptual quality and speech recognition.

In this paper we analyze the effect of various speech enhancement algorithms and functional modules, which has been focused on improving the perceptual quality, to the performance of automatic speech recognizer systems. We use the hidden markov toolkit (HTK) [7] with the TIMIT database [8]. Actually, it is very difficult to analyze the role of each functional module independently because they are organically coupled each other, thus hard to separate the role of each functional module. Therefore to observe the influence of particular module to recognition performance, we need to fix the other functional modules. In other words when we focus on the effects of SAP estimator the noise estimator and A priori SNR estimator should be fixed. Since the recognition performance is highly related to the accuracy of noise estimator, we assume perfect noise estimation first, and use a first-order recursive smoothing to degrade the accuracy of the noise estimation.

We compare the recognition performance of four gain functions such as Wiener, MMSE-STSA estimator, MMSE-LSA estimator, and the optimally modified LSA (OM-LSA) estimator [6]. From the results, we investigate how the gain estimators relating to other functional modules affect the speech recognition capability. To evaluate the impact of speech absence probability to ASR performance we select two methods: fixed for all frequency bins and an adaptive method: a method which tracks the SAP values for each frequency bin continuously [3].

2. Single channel speech enhancement algorithm

In this section, we briefly summarize the single channel speech enhancement algorithm used in this paper.

Let $X(k,l) = A(k,l)e^{j\alpha(k,l)}, D(k,l)$ and $Y(k,l) = R(k,l)d^{j\theta(k,l)}$ denote the k-th coefficient of discrete Fourier transform of speech signal $x(t)$, uncorrelated additive noise signal $d(t)$, and observed signal $y(t)$ in frequency domain. Where l represents the frame index.

2.1. Wiener Filter

The Wiener filter corresponds to the criterion of minimizing the mean-square error of best time domain fit to the speech waveform. A gain of Wiener amplitude estimator is derived

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

A priori SNR $\xi(k,l)$ is defined by $\xi(k,l) = \lambda_x(k,l)/\lambda_d(k,l)$. Where $\lambda_x(k,l)$ and $\lambda_d(k,l)$ denote the variances of speech and noise, respectively.

2.2. MMSE-STSA

The Wiener filtering is not an optimal spectral amplitude estimator under the assumed statistical model and criterion. Assuming that speech and noise are independent Gaussian random process, the gain of MMSE spectral amplitude estimator which minimizes the mean square error of the $A(k,l)$ equals [2]

$$G_{MMSE}(k,l) = \Gamma(1.5)\frac{\sqrt{\upsilon(k,l)}}{\gamma(k,l)} M(-0.5;1;-\upsilon(k,l)), \quad (2)$$

where $\Gamma(\bullet)$ denotes the gamma function, with $\Gamma(1.5) = \sqrt{\pi}/2$, $M(a;c;x)$ is the confluent hyper-geometric function. $\upsilon(k,l)$ and *A posteriori SNR* $\gamma(k,l)$ are defined by $\upsilon(k,l) = \xi(k,l)\gamma(k,l)/(1+\xi(k,l))$, $\gamma(k,l) = R(k,l)^2/\lambda_d(k,l)$, respectively.

2.3. MMSE-LSA

Since it is well known that a distortion measure which is based on the mean-square error of the log-spectra is more suitable for speech processing [5], minimum mean-square error log spectral amplitude (MMSE-LSA) estimator defined by

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty}\frac{e^{-t}}{t}dt\right\}. \quad (3)$$

2.4. OM-LSA

In an MMSE-LSA estimator, the gain function should be modified by considering the uncertainty of speech presence in real environment, which requires the computation of speech absence probability (SAP) [6]. The conditional probability of speech presence $p(k,l) = P(H_1 \mid Y(k,l))$ can be derived

$$p(k,l) = \left\{1 + \frac{(1-q(k,l))}{q(k,l)}(1+\xi(k,l))\exp(-\upsilon(k,l))\right\}^{-1} \quad (4)$$

where $q(k,l)$ represents the *A priori* speech absence probability (SAP). Finally, the optimally modified LSA (OM-LSA) gain $G_{OM-LSA}(k,l)$ is determined by the conditional probability of speech presence with minimum threshold $G_{\min}$:

$$G_{OM-LSA}(k,l) = \{G_{LSA}(k,l)\}^{p(k,l)} \cdot G_{\min}^{1-p(k,l)}. \quad (5)$$

Since the gain modification to utilize the uncertainty of speech presence is very efficient to improve the perceptual quality of MMSE-LSA enhancement system [1][3][6], the *A priori* SAP is a key parameter of the gain modifier to adjust the level of noise suppression.

3. A priori speech absence probability estimation algorithm

In this section we describe two algorithms to estimate *A priori* SNR $q(k,l)$ introduced in this paper.

3.1. Fixed SAP

The simplest idea for SAP estimation is using a fixed value for all frequency bins [2]. A constant probability, q ,is assumed for all frequency components and all the analyzed input frames.

However, voiced speech can be considered quasi-harmonic and non-stationary. Furthermore speech energy may not be present in every spectral component. Thus we should allow for a different value in each frequency bin of frame, instead of assigning the same value of q to all frequency bins.

3.2. Varying SAP in Time

One of the estimation methods to obtain distinct values of q for each frequency bin in each frame is proposed by Malah in [3]. This method uses a recursive averaging of index function $I(k,l)$, which is a hard decision rule based on *A posteriori* SNR, $\gamma(k,l)$, and represents speech absence likelihood in each frequency bin. (i.e. $I(k,l) = 0$, if $\gamma(k,l) \geq \gamma_{th}$, and $I(k,l) = 1$, otherwise).

The estimated *A priori* SAP $\tilde{q}(k,l)$ is as follows

$$\tilde{q}(k,l) = \alpha_d \tilde{q}(k,l-1) + (1-\alpha_d)I(k,l). \quad (6)$$

This method, estimating distinct values of SAP for different bins which are tracked in time, handles reasonably well when speech is non-stationary and may not be present in every frequency bin when voiced.

4. Experimental results and discussion

4.1. Experimental Environments

The mean opinion score (MOS) test is used as perceptual quality measurement to estimate the performance of various enhancement algorithms. Total 20 listeners are asked to score 1~5 point for each enhanced speech sample. The hidden Markov toolkit (HTK) is used for our speech recognition simulations. The HTK recognizer is trained by clean training

data and tested by both noisy data and enhanced noisy data. The noisy speech is degraded by two different noise types, such as white and babble noise taken from Noisex92 database, with different SNR.

The TIMIT database [8] with 630 speakers is used for our simulation. For a phoneme recognition, the 61 TIMIT phones are mapped to a reduced set of 39 phones in training and testing procedures [9], and results are reported on this reduced set. The analysis of the recognition results is performed for only vowels (/a/, /e/, /i/, /o/, /u/). The recognition results of other phones are excluded from an analysis because the weak energy phones such as fricatives is hardly enhanced by any single channel enhancement algorithms, especially in low SNR environments. Since there are little differences between the recognition rates for low energy components, comparison of results for vowels is satisfactory to evaluate the performance.

The noise estimator used in our simulation is

$$\hat{\lambda}_d(k,l) = \alpha_d \hat{\lambda}_d(k,l-1) + (1-\alpha_d)\lambda_d(k,l) \,, \quad (7)$$

where $\hat{\lambda}_d$, λ_d , and α_d are an estimated noise power, true noise power of k-th frequency bin in l-th frame, and a smoothing factor, respectively. $\alpha_d = 0$ means the perfect noise estimator. As a smoothing factor, α_d , increases near to 1, the performance of noise estimator gets degraded.

4.2. Experimental Results

Table 1 and 2 show the MOS test scores of enhanced signals by each enhancement algorithms with a reliable noise estimator ($\alpha_d = 0.2$) in various noise environments. The OM-LSA algorithm with the SAP estimator used in Malah's method shows the best performance in aspects of the perceptual quality. LSA estimator shows better performance than MMSE estimator, since the human acoustic characteristic could be considered as a log scale. Wiener filter shows the worst perceptual quality due to the musical noise among all tested algorithms, though it is theoretically an optimal solution for MSE criterion. The results show that statistical models of derivation criterion and speech presence uncertainty enhance the perceptual quality.

Table 3 and 4 show the speech recognition rates of enhanced signals when the perfect noise estimator is used with and without the decision-directed method [2] in white noise environments. The Wiener filter with a good noise estimator shows the best performance, because the Wiener filter is an optimal solution when it has a perfect knowledge of noise components. Although Wiener filter is rarely used for enhancement applications due to its perceptual quality, it works well for the recognition system with a good noise estimator.

It is expected that the MMSE-STSA and MMSE-LSA show similar performance to the Wiener filter in high SNR environments [2][5]. The recognition rate of MMSE degrades more rapidly than that of LSA in low SNR. This is due to difference of gain functions suppressing noise for weak speech components or non-speech regions. It is related that the gain function of LSA estimator always gives a lower gain than the MMSE gain function.

The simulation results also show that the decision-directed method causes adverse effects on speech recognition performance when a good estimator is coupled. The decision-directed method is known as one of the best estimator for *A priori* SNR to improve the perceptual quality by reducing residual noise components. Since it estimates *A priori SNR* by recursively a smoothing gain term of previous frame and *A posteriori* SNR of current frame, however, it causes bad time resolution so that short duration phones are not recognized well.

The results also show the effects of the SAP estimator on the ASR. As shown in the results, the perceptual improvement does not match up to the recognition performance. The Malah's method using a dynamic SAP estimator shows worse recognition performance. A speech uncertainty concept to improve the perceptual quality can give a worse effect to the recognition due to unnecessary speech distortions under good noise estimation environments.

Table 3 and Table 4 show that the recognition performance of enhancement algorithms using the perfect noise estimator in the stationary noise environments is degraded by methods such as a decision-directed method and SAP estimator. However, since a decision-directed method and SAP estimator are developed to improve perceptual quality in non-stationary noise environments, we also conduct same experiments in non-stationary noise environments.

Algorithm		00dB	05dB	10dB	20dB
Wiener		1.89	2.34	2.75	3.91
MMSE		2.18	2.49	2.84	3.81
LSA		2.50	2.84	3.30	4.11
OM-LSA	Fixed	2.84	3.15	3.61	4.33
	Malah's	3.06	3.36	3.64	4.36

Table 1. MOS test scores (babble noise)

Algorithm		00dB	05dB	10dB	20dB
Wiener		1.92	2.29	2.73	3.94
MMSE		1.84	3.68	2.94	3.93
LSA		2.24	2.62	3.23	4.26
OM-LSA	Fixed	2.25	2.91	3.42	4.38
	Malah's	2.39	3.05	3.50	4.43

Table 2. MOS test scores (white noise)

Algorithm		00dB	05dB	10dB	20dB
Clean		68.14			
Noisy		35.21	43.40	46.07	55.39
Wiener		60.78	64.55	66.25	67.68
MMSE		50.20	54.16	58.51	64.70
LSA		53.29	58.54	62.26	65.81
OM-LSA	Fixed	52.52	57.63	61.74	65.25
	Malah's	52.97	57.88	61.88	65.73

Table 3. Recognition rate (%) with decision-directed method (white noise)

Algorithm		00dB	05dB	10dB	20dB
Clean		68.14			
Noisy		35.21	43.40	46.07	55.39
Wiener		60.78	64.55	66.25	67.68
MMSE		58.00	61.77	64.84	67.38
LSA		59.21	62.91	65.48	67.59
OM-LSA	Fixed	57.94	62.68	65.30	67.34
	Malah's	58.26	62.36	65.09	67.32

Table 4. Recognition rate (%) without decision-directed method (white noise)

Algorithm		00dB	05dB	10dB	20dB
Clean		68.14			
Noisy		33.21	37.48	44.04	57.48
Wiener		59.88	62.10	63.74	65.67
MMSE		48.18	52.92	57.62	64.60
LSA		50.79	56.46	60.71	65.29
OM-LSA	Fixed	49.15	55.15	60.22	64.65
	Malah's	49.59	55.67	60.13	64.96

Table 5. Recognition rate (%) with decision-directed method (babble noise)

Algorithm		00dB	05dB	10dB	20dB
Clean		68.14			
Noisy		33.21	37.48	44.04	57.48
Wiener		59.88	62.10	63.74	65.67
MMSE		56.22	60.01	62.82	66.52
LSA		56.77	60.44	62.98	66.35
OM-LSA	Fixed	56.03	59.91	62.44	65.27
	Malah's	56.56	59.86	62.82	65.60

Table 6. Recognition rate (%) without decision-directed method (babble noise)

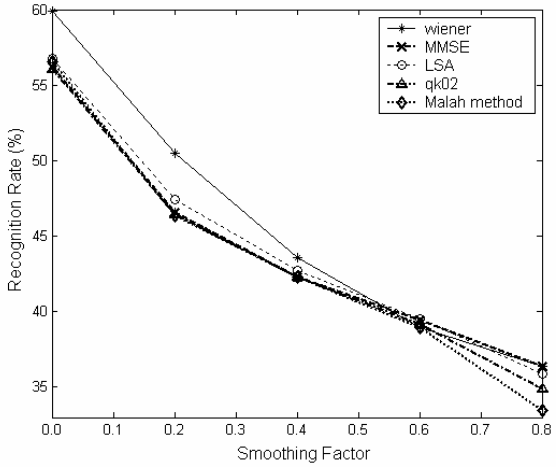

Fig.1. Recognition rate of each algorithm according to the smoothing factor

The recognition rates of enhanced signals with perfect noise estimator in babble noise environment are shown in Table 5 and 6. The babble noise is considered as having more non-stationary characteristics. The results show similar patterns to those of white noise environment, while overall recognition rates are a little lower than those of white noise environments.

When the perfect noise estimator is attended, a decision-directed method and SAP estimator degrade the recognition performance in both stationary noise and non-stationary noise environments. Thus, we are required to run simulation with noise estimator having various performances.

Fig. 1 plots the recognition rate of each algorithm by varying the averaging factor α_d of noise estimator. We may assume that the performance of the estimator degrades as the smoothing factor approaches to one. Experiments are performed without using the decision-directed method in babble noise (0dB SNR) environment. This result shows that the recognition rates of all surveyed algorithms are monotonically decreasing as the performance of noise estimator gets worse. Especially, the recognition accuracy of enhanced signals by the Wiener filter is degraded faster than others. The result also shows that algorithms using SAP estimators such as Malah's cause more degradation of the recognition performance under unreliable noise estimation environments using $\alpha_d = 0.6 \sim 0.8$.

5. Conclusions

In this paper we attempted to understand how much each functional module of the speech enhancement algorithm affects to the automatic speech recognition performance. The results of speech recognition tests showed that enhancement modules such as SAP estimator and decision-directed method which were developed to improve the perceptual quality causeed adverse effects to the performance of the ASR system.

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

[3] 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*, p.789-792, 1999.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE trans. Acoust., Speech, Signal processing*, vol.9, pp.504-512, Jul.2001

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[6] 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.

[7] Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Valtcho Valtchev, Phil Woodland, "The HTK Book," copyright 1995-1999 Microsoft Corporation, copyright 2001-2002 Cambridge University Engineering Department.

[8] 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)

[9] Kai-Fu Lee and Hsiao-Wuen Hon. "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 11, Nov. 1989