A Single Channel Speech Enhancement for Automatic Speech Recognition

*Jinkyu Lee *Hyunson Seo *Hong-Goo Kang

*Dept. of Electrical and Electronic Engineering, Yonsei University, Seoul

*shuya@dsp.yonsei.ac.kr

Abstract

This paper describes a single channel speech enhancement as the pre-processor of automatic speech recognition system. The improvements are based on using optimally modified log-spectra (OM-LSA) gain function with a non-causal a priori signal-to-noise ratio (SNR) estimation. Experimental results show that the proposed method gives better perceptual evaluation of speech quality score (PESQ) and lower log-spectral distance, and also better word accuracy. In the enhancement system, parameters was turned for automatic speech recognition.

1. Introduction

Recently, we can easily watch or download plenty of multi-media clips from the broadcast or Internet. Moreover, thanks to video-sharing website such as Youtube, we are able to access not only domestic data but also can easily download foreign multi-media clips. In several years, automatic captioning or translation applications have been popular for non-native people. In several decades, the performance of automatic speech recognition (ASR) systems has increased significantly, but in order to be used in real applications, there was some problem. It just gives the serviceable performance just in very clean environment. In order to make the system robust in noisy environment, the additional modules are needed. Especially, with relatively stationary noise, a single channel speech enhancement technique is very effective in improving the performance of ASR system. A single channel speech enhancement system can be separated as four different functional modules. The first module is gain estimator, and the second one is noise power spectral density (PSD) estimator. The third one is a priori SNR estimator, and the last one is speech absence probability (SAP) estimator.

The gain estimator is the module to determine the degree of reduction. Plenty of algorithms have been developed for several decades. The Wiener filtering, spectral subtraction [1], maximum likelihood envelope estimation [2], minimum mean-squared error short time spectral amplitude [3], minimum mean-squared error of the log-spctra [4], and optimally modified LSA (OM-LSA) [5] are the method to find gain function. In this paper, OM-LSA method will be analysed, In order to applying the gain function of OM-LSA method, the SAP estimator has to be required to make the system robust. Moreover, with these gain function, also noise power has to be estimated. Firstly decision-directed method was used, and to improve the performance, the non-casual a priori SNR estimator was used, and the improvements will covered in section 4. Finally, improved minima controlled recursive averaging method (IMCRA) [6] was used in noise PSD estimation modules.

2. Automatic Speech Recognition System

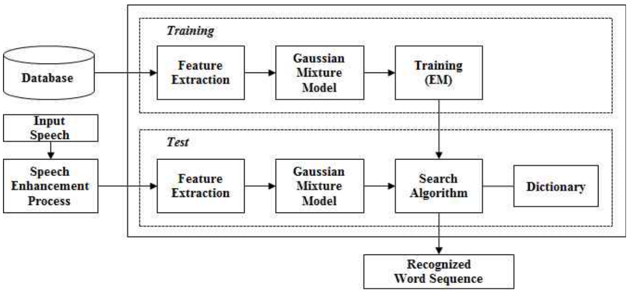

Fig. 1. Block diagram of ASR system using a single channel speech enhancement module as a pre-processor.

Fig. 1. shows the block diagram of proposed ASR system which has a speech enhancement system as a pre-processor. In the experiment, the Aurora-2 [7] database and HTK in its version 3.4 [8] are used to create whole word hidden Markov models (HMMs) for all digits. Each digit model has 16 states having the mixture of 3 Gaussians per feature and state, and one silence model has 3 states having the mixture of 6 Gaussians. Additionally, short pause model for inter word model was created which has single state sharing the mixtures with the silence model. In the document, two training mode have been defined. In this experiment, only the clean utterances was taken as training data, and four different noise signals have been selected to be added at SNRs of -5, 0, 5, 10, 15 and 20 dB as test data.

3. Single Channel Speech Enhancement

A. Gain function

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

The gain function which will be introduced firstly, is Wiener filter. It is used in ASR system widely, due to simplicity and low computational power. Eq.(1) shows that it just needs the variance of speech and noise, and it can be rewrite using just the function of a priori SNR, so that it can be suitable for real-time application. Since the Wiener amplitude estimator is optimal solution of signal spectral amplitude, not optimal spectral amplitude estimator. So, by minimizing the distortion of Eq.(2), the amplitude estimator of A(k,l) can be obtained as Eq.(3).

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

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

However, considering the uncertainty of speech presence probability, it has to be modified as below Eq.(4), and Eq.(5)

$$G_{H_1}=\frac{\xi(k,l)}{1+\xi(k,l)}exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right). \tag{4}$$

$$G_{OM-LSA}=G_{H_1}(k,l)^{p(k,l)}G_{\min}^{1-p(k,l)}. \tag{5}$$

The gain function of OM-LSA method requires estimated a priori speech absence probability (SAP) and minimum threshold of gain for a non-speech component.

B. SAP Estimator

Different from Winer filtering and MMSE-LSA estimatior, the OM-LSA gain function needs the estimated speech absence probability SAP. There are two method to estimate SAP. The first one is to consider it as a constant value. However, this method does not gives us good performance, because the characteristic of speech signal is not stationary. So that, instead of constant value, it can be applied that different values in each frame and frequency bin.

C. A Priori SNR Estimator

The estimation of a priori SNR is also one of the most important modules in single channel speech enhancement system. Firstly, the decision-directed method can be derived by recursively averaging a priori SNR of the previous frame and the instantaneous SNR of the current frame.

$$\xi_{DD}(k,l)=\max\left\{\alpha\frac{\widehat{A^2}(k,l-1)}{\lambda_D}+(1-\alpha)[\gamma(k,l)-1],\xi_{\min}\right\}. \tag{6}$$

Eq. (6) shows how a priori SNR can be estimated using decision-directed method. The next step, we extend this concept to non-causal estimation [9], which is useful in applications hat can tolerate a delay of at least 100 ms in the estimated signal.

$$\begin{aligned}\hat{\lambda}_{X|l+L}(k,l)=\max\{&\mu\hat{A}^2(k,l-1)+(1-\mu),\\ &\times[\mu'\sum_{i=-w}^{w}b(i)\lambda_{X|l+L-1}(k-i,l-1)\\ &+(1-\mu')\lambda'_{X|[l,l+L]}(k,l),\ \lambda_{\min}\}.\end{aligned} \tag{7}$$

$$\lambda'_{X|[l,l+L]}(k,l)=\max\left\{\frac{\sum_{(n,i)\in\Gamma}b(i)\,|\,Y(k-i,l+n)\,|^2}{\sum_{(n,i)\in\Gamma}b(i)}-\beta\lambda_D,0\right\}. \tag{8}$$

$$\xi_{NC}(k,l)=\frac{\hat{\lambda}_{X|l+L}(k,l)}{\lambda_D(k,l)}. \tag{9}$$

Eq. (7), (8), (9) shows a priori SNR can be estimated using non-causal method. The first part of Eq. (7) is exactly same as the decision-directed method. However, instead of using the instantaneous SNR of the current frame, it uses recursively averaging smoothed variance of future frames, and also uses recursively averaging its previous value.

4. Experimental Results

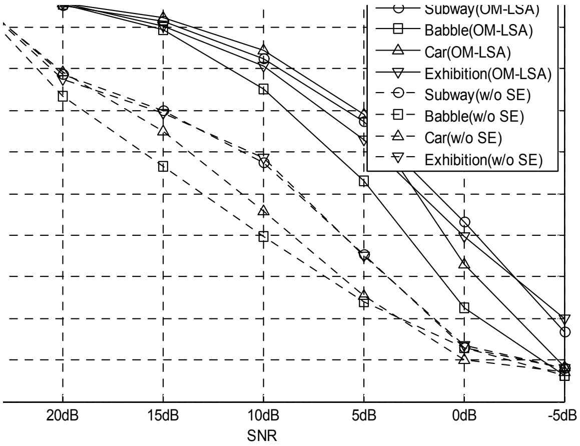

Fig. 2. Recognition rates of original signals and enhanced signals under four different noisy environments.

Table 1. Word accuracy as percentage for without pre-processor

SNR/dB	w/o pre-processor				
	Subway	Babble	Car	Exhibition	Average
CLEAN	99.17	98.85	98.75	99.23	99.00
20	78.66	73.31	79.27	77.48	77.18
15	69.97	56.56	65.05	69.21	65.20
10	57.41	39.69	45.87	58.59	50.39
5	35.31	24.12	25.47	35.08	30.00
0	12.99	12.97	10.05	13.39	12.35
-5	7.92	6.35	7.25	7.93	7.36

Table 2. Word accuracy as percentage for with OM-LSA pre-processor

SNR/dB	with OM-LSA & IMCRA				
	Subway	Babble	Car	Exhibition	Average
CLEAN	99.05	98.85	98.81	99.17	98.97
20	93.03	95.71	96.69	93.09	94.63
15	87.2	89.99	93.89	85.75	89.21
10	77.37	75.39	85.03	75.69	78.37
5	60.82	50.57	66.6	59.64	59.41
0	38.99	20.04	32.51	36.87	32.10
-5	16.64	6.86	7.72	19.28	12.63

Table 3. Word accuracy as percentage for with non-causal OM-LSA pre-processor

SNR/dB	with Non-Causal OM-LSA & IMCRA				
	Subway	Babble	Car	Exhibition	Average
CLEAN	99.26	99.03	99.14	99.29	99.18
20	95.06	95.59	95.65	95.74	95.51
15	91.31	89.33	92.16	90.13	90.73
10	82.28	75.27	84.40	80.78	80.68
5	67.18	52.96	69.04	63.01	63.05
0	43.11	22.67	32.96	39.56	34.58
-5	16.73	6.20	8.23	19.99	12.79

Table 4. PESQ scores and log spectral distance of decision-directed method and non-causal method

	Noise Type	Method	Input SegSNR(dB)					
			20	15	10	5	0	-5
PESQ Scores	subway	DD	2.86	2.62	2.51	1.98	1.51	0.92
		NC	2.92	2.66	2.55	2.04	1.50	0.91
	babble	DD	3.01	2.78	2.46	2.28	1.95	0.89
		NC	3.07	2.81	2.50	2.28	1.97	0.95
Log Spectral Distance	subway	DD	1.29	1.52	1.63	1.81	2.13	2.60
		NC	1.26	1.50	1.60	1.77	2.10	2.54
	babble	DD	0.89	1.11	1.29	1.47	1.70	2.04
		NC	0.87	1.12	1.27	1.47	1.69	2.02

Fig. 2. depicts the recognition rates of signals with and without single channel speech enhancement as a pre-processor. The spectral gain used in our evaluation is the gain of OM-LSA. The noise type 'Car' shows significant improvement, because it has relatively stationary characteristic among them. However, in low SNR cases, since not only background noise but also fricative sounds was suppressed, there is almost no improvements.

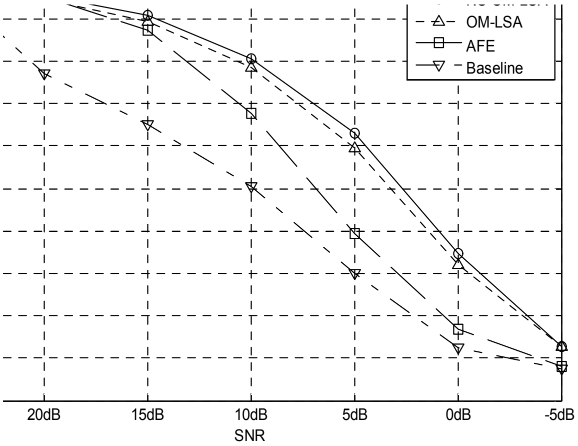

Fig. 3. Average recognition rates of original signals and enhanced signals.

Table 3. shows the result of single channel enhancement with OM-LSA, the non-causal a priori SNR estimator and IMCRA [6]. In the non-causal module, a priori SNR was estimated by using parameters $\mu=\mu'=0.8, b=[0.25\,0.5\,0.25]$, $L=3, \beta=2$, $\xi_{\min}=-20dB$. In high SNR cases, the number of future frame used in estimating as 7 shows the best performance. However, in low SNR cases, it causes over-estimation, so that the size 3 is used for stability. Furthermore, Table 4. presents the results of the perceptual evaluation of speech quality (PESQ) score and the log spectral distance (LSD) achieved by using the decision-directed and non-causal a priori SNR estimator. The non-causal estimator yields a higher improvement lower LSD, and higher PESQ scores.

Fig. 3. shows the average recognition rate among those four types of noise in each applied algorithm. When the input SNR is high, AFE algorithm using Winer filter gives us the best performance, but other cases, lower than 20dB, proposed method gives us the best performance.

5. Conclusion

Throughout this paper, we have investigated the effect of the single channel enhancement modules as a pre-processor of automatic speech recognition system, and compared the effects of several different modules. The results of several measurement, such as MOS test and LSD, were used in order to compare the relationship between perceptual quality and speech recognition performance. However, these measurements do not give the information of recognition rate, and those are not highly depends on the performance of ASR system. So that, HMM modules had to be re-trained in every times the enhancement algorithm was modified. Since we assumed the off-line environment, which means non-causal system can be applied, the computational power or the complexity of algorithms are not our interest. Therefore, several highly complex single channel enhancement algorithms was applied such as OM-LSA and IMCRA. Those methods gives us better recognition rate rather than when AFE [6] front-end module is applied. Furthermore, non-causal a priori SNR estimator increase the performance slightly, but in stationary noisy environment, its performance was better than using the decision-directed method.

Reference

[1] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 27, no. 2, Apr. 1979

[2] R. J. Mcaulay, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Transactions on Acoustic, Speech, and Signal Processing,* vol. ASSP-28, no. 2, Apr. 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mena-square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustic, Speech, and Signal Processing,* vol. ASSP-32, pp. 1109-1121, Dec. 1984.

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

[5] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator", *IEEE Signal Processing Letters, vol. 9, no. 4,* Apr. 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging", *IEEE Transactions on Speech and Audio Processing, vol. 11, no.5,* Sep. 2003.

[7] H. G. Hirsch, D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *ISCA ITRW ASR 2000,* Sep. 2000.

[8] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, V. Valtchev, P. Woodland, "The HTK Book", copyright 1995-1999 Microsoft Corporation, copyright 2001-2002 Cambridge University Engineering Department.

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