Improved Single Channel Speech Enhancement Algorithm Using Adaptive Postfiltering.

Eunwoo Song, Hong-Goo Kang

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

sewplay@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

Abstract

In real environment, background noise exists everywhere and degrades the performance of system. To reduce this distortion, a speech enhancement algorithm can be very useful and variety methods have been proposed. In this paper, we propose a postfilter to improve the performance of optimally modified log-spectral amplitude (OM-LSA) estimator. Proposed algorithm uses the formant postfilter to minimize perceptual distortion caused by background noise. We adjust an emphasizing parameter which is varied by spectral flatness and first reflection coefficient. The performance of the proposed algorithm is evaluated by measuring the log-spectral distance (LSD) and the perceptual evaluation of speech quality (PESQ) score. The test results show the improvement of proposed algorithm compared to conventional OM-LSA.

1. Introduction

In speech communication or broadcasting systems, the needs for speech enhancement algorithm has increased to reduce spectral distortion caused by unwanted noise component. One of the conventional single channel speech enhancement algorithms is a spectral domain speech enhancement (SDSE) such as OM-LSA estimator which minimizes mean squared error based on log-spectra between clean speech and estimated speech [2]. To obtain enhanced signal, spectral gain is estimated in each frequency bin and applied to observed signal with speech presence probability. OM-LSA estimator shows significant improvement of noise reduction, however, it is not optimal in terms of perceptual quality enhancement [6]. Although OM-LSA estimator considers log-spectral amplitude by means of perceptual approach [8], it is limited in individual frequency bin rather than overall spectral structure. To achieve more improvement in perceptual quality, additional module is needed which controls spectral structure. One of the helpful approaches is using an adaptive postfilter because it includes refining a formant structure based on human perception [5].

In this paper, we propose a noise-dependent adaptive postfilter to improve the performance of OM-LSA estimator. We extract features which contain the noise characteristics, spectral flatness and first reflection coefficient. From these features, we find a mapping function that controls the parameters of the postfilter which minimize perceptual distortion.

The organization of this paper is as s follows. In Section 2, as a single channel speech enhancement algorithm, we introduce OM-LSA estimator and extend it with the postfilter in Section 3. Section 4 includes the experimental setup and performance evaluation. Finally, we summarize the proposed algorithm in Section 5.

2. OPTIMALLY MODIFIED LOG-SPECTRAL AMPLITUDE ESTIMATOR

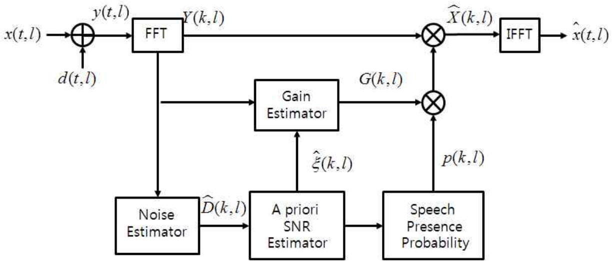

Fig. 1. Block diagram of basic single channel speech enhancement algorithm with OM-LSA estimator.

Fig. 1 represents a basic block diagram of OM-LSA as a SDSE algorithm. Let $x(t)$ and $d(t)$ are speech and uncorrelated additive noise signal, the observed signal $y(t)$ becomes

$x(t)+d(t)$ with independent gaussian random process. In frequency domain, $X(k,l)=A(k,l)e^{j\alpha(k,l)}$ where k is the frequency bin index and l is the frame index. Then the observed signal $Y(k,l)$, is as below

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

Assume that the effect of phase component is neglectable [4][7][8], the estimated clean speech can be represented as

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

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

where $G(k,l)$ represents a spectral gain function shown in Eq.(4).

$$G_{OM-LSA}(k,l)=\{G_{MMSE-LSA}(k,l)\}^{p(l,k)}G_{\min}^{1-p(k,l)} \tag{4}$$

where $p(k,l)$ and $G_{\min}$ denote the speech presence probability and minimum gain, respectively. $G_{MMSE-LSA}(k,l)$ is the gain of minimum mean-square error log-spectral amplitude (MMSE-LSA) [8] estimator and can be obtained by Eq.(5).

$$G_{MMSE-LSA}(k,l)=\frac{\xi(k,l)}{1+\xi(k,l)}exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{5}$$

where $v(k,l)$ defined by $v(k,l)=\xi(k,l)\gamma(k,l)/1+\xi(k,l)$. $\xi(k,l)$ and $\gamma(k,l)$ are a priori SNR $\xi(k,l)=A^2(k,l)/\lambda_d(k,l)$ and a posteriori SNR $\gamma(k,l)=R^2(k,l)/\lambda_d(k,l)$, respectively, and $\lambda_d(k,l)$ denotes the noise power spectral density.

3. PROPOSED NOISE-DEPENDENT POSTFILTER

Even if such spectral domain speech enhancement algorithms have been developed, the problem of perceptual distortion still exists. If a more reduction of distortion is avaliable, the perceptual quality of enhanced signal will be improved. One of the these approaches is to have an additional module like a formant postfilter. The formant postfilter, used in speech coding systems [5], reduces quantization noise by emphasizing spectral peak and de-emphasizing spectral valleys. The most commonly used formant postfilter consists of core postfilter $H_f(z)$, tilt correction filter $H_t(z)$, and gain control factor G. Its general form is

$$H(z)=GH_f(z)H_t(z)=G\frac{A(z/\gamma_1)}{A(z/\gamma_2)}(1-\mu z^{-1}) \tag{6}$$

where is $A(z)$ is the linear prediction (LP) filter, γ_i is fixed parameter that controls the degree of spectral emphasis where $0<\gamma_1<\gamma_2<1$, and μ is a tilt compensation parameter.

We propose an algorithm that controls emphasizing parameter γ_i which is not fixed but adaptive, dependent on noise level. Fig.2. is a block diagram of noise dependent postfilter with OM-LSA algorithm. In each frame l, the information about noise characteristic is estimated by spectral flatness and first reflection coefficient. Then each features properly mapped into emphasizing parameters γ_1 and γ_2.

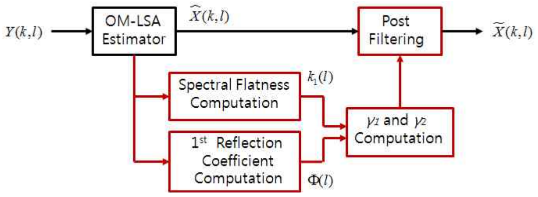

Fig. 2. Block diagram of noise dependent postfilter with single channel speech enhancement algorithm

3.1. Noise Feature

In proposed algorithm, two features are used to determine the noise characteristic. Spectral flatness (SF) is measured by means of determine the degree of noise level and first reflection coefficient (RC) is measured by means of determine unvoiced region from noisy speech.

SF of a spectrum $X(w)$ can be obtained by Eq.(7).

$$\Phi=\frac{\sqrt[N]{\prod_{k=0}^{N-1}|X(k)|^2}}{\frac{1}{N}\sum_{k=0}^{N-1}|X(k)|^2} \tag{7}$$

is the ratio between geometric mean and arithmetic mean where N is the number of frequency bin. Because arithmetic mean is always greater than geometric mean, SF always lies in the range [0,1]. SF equals one only for a perfectly flat spectrum, and decreases when the variation of spectrum increases. That means SF can reveal the noise level [1], because SF in noise region is significantly higher than speech region. For example, the SF calculated in each frame of 10dB noisy signal with white noise (see Fig. 3(a),(b)) is shown in Fig. 3(c). In noise only region ($0\sim0.6$ sec), SF value is peak, on the other hand, SF value is near to zero when speech is superior than noise ($0.8\sim1.0$, $1.3\sim1.4$sec, etc.).

RC is measured by time domain signal $x(t)$ and can be obtained by

$$k_1 = \frac{R_{xx}(1)}{R_{xx}(0)} = \frac{\sum_{n=0}^{L-2} x(n)x(n+1)}{\sum_{n=0}^{L-1} x^2(n)}, \qquad (8)$$

where $R_{xx}(\tau)$ is autocorrelation function, n is time index, and L is window length. Because $R_{xx}(0)$ is always more than $R_{xx}(1)$, RC is always in the range [-1,1]. RC is closed to ±1 when the signal has high correlation like speech. Fig. 3(d) is first reflection coefficient of noisy signal. In noise only region, RC is almost zero compared to speech region which has higher RC. An interesting fact is that RC has additional information of unvoiced signal. In unvoiced region such as /s/ ($1.5 \sim 1.6$, $2.2 \sim 2.3$sec, etc.), RC also has higher value but its sign is negative. We utilize this characteristic to determine whether each frame is noise-like region or unvoiced-speech-like region. If the frame has determined to unvoiced-speech-like region, the formant postfilter will de-emphasize the formant frequencies rather than emphasizing in common case.

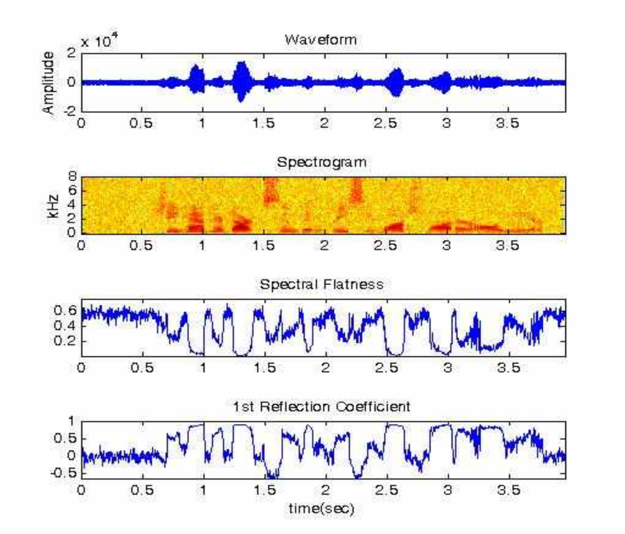

Fig. 3: (a) and (b) is 10dB Speech signal for the utterance "She had your dark suit in greasy wash water all year." with white noise. (c) Spectral flatness. (d) First reflection coefficient.

3.2. Noise Dependent Postfilter.

We performed simulations with several noise type and different SNR. A simple mapping function from SF and RC to emphasizing parameters γ_1 and γ_2 is obtained as

$$\gamma_1(l) = \begin{cases} \gamma_{1,\min} & , SF(l) < SF_{1,\min} \\ \alpha_1 SF(l) + \alpha_2 & , SF_{1,\min} \le SF(l) < SF_{1,\max} \\ \gamma_{1,\max} & , SF(l) \ge SF_{1,\max} \end{cases} \qquad (9)$$

$$\gamma_2 = \gamma_1 + \beta(l) RC(l),$$

$$\beta(l) = \begin{cases} \beta_{\min} & , SF(l) < SF_{2,\min} \\ \alpha_3 SF(l) + \alpha_4 & , SF_{2,\min} \le SF(l) < SF_{2,\max} \\ \beta_{\max} & , SF(l) \ge SF_{2,\max} \end{cases} \qquad (10)$$

where α_i and β are a weighting factor, $\gamma_{1,\min}$, $\gamma_{1,\max}$, $\beta_{\min}$, $\beta_{\max}$, $SF_{i,\min}$, and $SF_{i,\max}$ threshold parameters shown in Table 1. γ_2 has more dynamic range than γ_1, because the experiential result show that γ_2 is more sensitive to spectral distortion more than γ_1. The mapping function represents that more regressive postfiltering is done when the frame determined more noise-like, higher value of SF. On the other hand, when the frame determined more speech-like, lower value of SF, more robust filtering is achieved by controlling the parameters γ_1 and γ_2.

Table 1. Values of parameters used for the control of the emphasizing parameter γ_1 and γ_2.

$SF_{1,\min} = 0.16$	$SF_{2,\min} = 0.1$	$\gamma_{1,\min} = 0.4$
$SF_{1,\max} = 0.2$	$SF_{2,\max} = 0.23$	$\gamma_{1,\max} = 0.5$
$\alpha_1 = 2.288$	$\alpha_3 = 2.336$	$\beta_{\min} = 0.1$
$\alpha_2 = 0.041$	$\alpha_4 = -0.127$	$\beta_{\max} = 0.4$

4. PERFORMANCE EVALUATION

The proposed postfiltering algorithm is implemented into OM-LSA algorithm with IMCRA noise estimator [3]. Noisy speech is generated using three types such as white, pink, and factory noises taken from the Noisex92 database, with 10dB and 0dB SNR.

In each items, log-spectral distance is measured and its result is summarized in Table 2 and Table 3. The distance of proposed algorithm is smaller than OM-LSA. It means that the spectral distortion is reduced by applying the postfilter to conventional OM-LSA algorithm.

Table 2. Log-spectral distance of enhanced signal(dB) (10dB SNR)

Algorithm	White		Pink		Factory	
	Female	Male	Female	Male	Female	Male
Unprocessed	8.0848	7.3713	7.0307	6.2358	5.2146	4.4868
OM-LSA	4.7080	4.3038	3.9847	3.6266	3.5108	3.3967
Proposed	4.6053	4.2012	3.9264	3.6020	3.6021	3.5860

Table 3. Log-spectral distance of enhanced signal(dB) (0dB SNR)

Algorithm	White		Pink		Factory	
	Female	Male	Female	Male	Female	Male
Unprocessed	10.753	10.008	9.6119	8.7898	7.5401	6.734
OM-LSA	6.3516	5.8568	5.4314	4.9015	4.4540	4.2489
Proposed	6.2721	5.7629	5.2475	4.7408	4.4124	4.2473

To verify the perceptual quality of proposed algorithm, PESQ score is measured. From the result in Table 4 and Table 5, the objective perceptual quality is also good at proposed algorithm compared to OM-LSA algorithm.

Table 4. PESQ score of enhanced signal (10dB SNR)

Algorithm	White		Pink		Factory	
	Female	Male	Female	Male	Female	Male
Unprocessed	1.933	2.018	1.944	2.096	2.026	2.203
OM-LSA	2.590	2.546	2.438	2.692	2.707	2.790
Proposed	2.615	2.575	2.455	2.743	2.731	2.836

Table 5. PESQ score of enhanced signal (0dB SNR)

Algorithm	White		Pink		Factory	
	Female	Male	Female	Male	Female	Male
Unprocessed	1.059	1.14	1.095	1.194	1.196	1.349
OM-LSA	1.431	1.682	1.455	1.698	1.716	1.768
Proposed	1.512	1.699	1.456	1.616	1.732	1.799

5. Conclusion

We present in this paper an algorithm for single channel speech enhancement with adaptive fomant postfiltering. The basic approach taken here is to minimize perceptual distortion caused by uncorrelated additive noise. We proposed a mapping function from noise characteristics to postfilter parameters. Noise characteristics are determined by two parameters, spectral flatness and first reflection coefficient. By controlling the parameters γ_1 and γ_2, we can achieve more reduction of noise after the core speech enhancement module.

6. Reference

[1] B. Yegnanarayana, C. A. Avendano, H. Hermansky and P. S. Murthy, "Processing linear prediction residual for speech enhancement," in *Proc. EUROSOEECH'97,* 1997.

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

[3] 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.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *IEEE Signal Processing,* vol. 81, no. 11, pp. 2403-2418, Oct. 2001.

[5] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 3, no.1, pp. 59-71, Jan, 1995.

[6] M. S. Choi, H. G. Kang, "An improved estimation of a priori speech absence probability for speech enhancement : in perspective of speech perception," *Int. Conf. Acoustics, Speech and Signal Processing 2005,* 2005

[7] 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.

[8] 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.