AN IMPROVED ESTIMATION OF A PRIORI SPEECH ABSENCE PROBABILITY FOR SPEECH ENHANCEMENT : IN PERSPECTIVE OF SPEECH PERCEPTION

Min-Seok Choi and Hong-Goo Kang

MCSP Lab., Yonsei University, Korea
[zzugie, hgkang]@mcsp.yonsei.ac.kr

ABSTRACT

The purpose of this paper is to improve the perceptual quality of a single channel speech enhancement algorithm using MMSE LSA estimator. The proposed algorithm uses a non-linear decision rule and an adaptive recursive averaging factor for tracking *a priori* speech absence probability (SAP) fast. We also introduce one-third of approximated critical bandwidth to efficiently smooth the *a priori* SAP and final gain term, which successfully eliminates the musical noise without much distortion of signal. The performance of the proposed algorithm is evaluated by performing subjective A/B listening tests and measuring spectral distance. Simulation results verify the effectiveness of the proposed algorithm compared to conventional algorithms.

1. INTRODUCTION

In speech communication systems, single channel noise reduction techniques are used to improve the perceptual quality of noise-corrupted speech signals. The techniques are also applicable for a preprocessor of speech signal processing systems such as recognition, coding, and so on [1].

The minimum mean-square error log-spectral amplitude (MMSE LSA) estimator is proposed to minimize the mean square error between the logarithmic spectrum of enhanced speech and clean speech. The algorithm is very efficient for noise reduction if it correctly estimates required parameters such as noise power spectral density (PSD) and the *a priori* SAP [2][3].

The *a priori* SAP represents the probability of speech absence in the input signal. Using the *a priori* SAP, the conditional probability of speech presence for given observed spectral component is calculated, which uses as a gain modifier of the spectral component. Many approaches to tracking the *a priori* SAP are developed, but we cannot say that which one is better because of difficulties in determining the criterion of quality measure. Recently, an adaptation of *a priori* SAP using the signal-to-noise ratio(SNR) information is proposed to improve the performance of MMSE LSA algorithm [4]. In the algorithm, a hard-decision approach is used for tracking the value [4]. Therefore, *a priori* SAP could not utilize the SNR information efficiently and misclassification of the speech activity causes undesired artifacts. To further improve the performance, adaptive tracking and soft decision of the *a priori* SAP is needed. Cohen proposed method by combining three parameters for the *a priori* SAP estimation [3]. In his work, local and global values of speech absence probability, and a soft decision on whether the current frame contains speech or not are used as the parameters of the estimator [5]. Though, this algorithm improves the performance of enhancement systems in SNR aspects, it is still not optimal in terms of perceptual quality aspect. In fact, it is well known that SNR improvement does not match exactly with perceptual quality. Therefore, another approach is needed to build a perceptually enhanced system. The reduction of musical noise is one of the most important parts to improve the perceptual quality of speech enhancement system. Smoothing of some parameters such as *a priori* SNR, *a posteriori* SNR, or gain of system can be helpful to reduce the musical noise [6].

In this paper, we propose an algorithm to efficiently estimate the *a priori* SAP for the MMSE LSA system. The non-linear soft decision rule and an adaptive tracking factor are proposed to achieve the goal. In addition, the *a priori* SAP and the gain is also smoothed in the frequency domain with respect to the bark band rate, which is helpful for eliminating musical noise.

2. MINIMUM MEAN-SQUARE ERROR LOG-SPECTRAL ESTIMATOR

Let $x(t)$ and $d(t)$ are speech and the additive noise signal, then the observed signal $y(t)$ becomes $x(t) + d(t)$ with the additive noise assumption. In frequency domain, $X(k,l) = A(k,l)e^{j\alpha(k,l)}$, $D(k,l)$ and $Y(k,l) = R(k,l)e^{j\theta(k,l)}$ denote the kth coefficient of discrete Fourier transform of $x(t)$, $d(t)$, and $y(t)$. Where l represents the frame index.

Assuming that the speech and the noise are independent gaussian random process, the spectral amplitude estimator $\hat{A}(k,l)$ which minimizes the mean square error with loga-

rithm of the $A(k,l)$ equals

$$\begin{aligned}\hat{A}(k,l) &= \exp\{E[\ln A(k,l)|Y(k,l)]\} \\ &= \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left\{\int_{v(k,l)}^{\infty} \frac{e^{-t}}{2\,t}\, dt\right\} R(k,l) \\ v(k,l) &= \frac{\xi(k,l)}{1+\xi(k,l)}\,\gamma(k,l), \end{aligned} \quad (1)$$

where $\xi(k,l)$ and $\gamma(k,l)$ denote the *a priori* SNR and *a posteriori* SNR, respectively [2][3].

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) [7]. Let's assume that H_1 represents the state of speech presence and H_0 is the state of speech absence, the conditional probability of speech presence can be derived

$$p(H_1|Y(k,l)) = \frac{\Lambda(k,l)}{1+\Lambda(k,l)} \quad (2)$$

$$\begin{aligned}\Lambda(k,l) &= \frac{(1-q(k,l))}{q(k,l)} \frac{p\left(Y(k,l)\,|H_1\right)}{p\left(Y(k,l)|H_0\right)} \\ &= \frac{(1-q(k,l))}{q(k,l)} \frac{\exp(v(k,l))}{1+\xi(k,l)}, \end{aligned} \quad (3)$$

where $q(k,l)$ represents the *a priori* probability of speech absence. In fact, *a priori* SAP can be a constant value in each frame or a function of frequency index k.

Finally, the gain which is defined as $G(k,l)$ is modified by $p(H_1|Y(k,l))$.

$$\begin{aligned} G(k,l) &= \hat{A}(k,l)/R(k,l) \\ &= \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left\{\int_{v(k,l)}^{\infty} \frac{e^{-t}}{2\,t}\, dt\right\} \\ G_D(k,l) &= p(H_1|Y(k,l))\, G(k,l). \end{aligned} \quad (4)$$

The gain modification to utilize the uncertainty of speech presence is very efficient to improve the performance of MMSE LSA system [2][3]. In the next section, we describe the issue on tracking the $q(k,l)$ [4][5].

3. TRACKING THE A PRIORI SPEECH ABSENCE PROBABILITY (SAP)

To improve the perceptual quality of enhanced speech, the amount of residual noise and musical noise effect should be minimized. As we already mentioned in the previous section, the *a priori* SAP is a key parameter of the gain modifier and it can adjust the level of noise suppression. Therefore, the fast tracking of *a priori* SAP is helpful to minimize the residual noise [7].

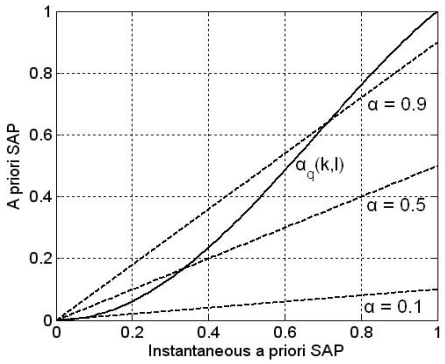

Fig. 1. Speed of recursive averaging with respect to the averaging factor

Musical noise is occurred by impulsive spectral components whose frequency is random in consecutive time domain [6]. The musical noise can be reduced by spreading out the impulsive components using the smoothing of SNR, gain or gain modifier. However, the frequency domain smoothing method can also degrade the clearness and crispness of speech quality, thus the smoothing method should be carefully designed.

In this work, a non-linear soft decision rule and an adaptive recursive averaging factor is used for fast tracking of *a priori* SAP in time domain. In addition, we take the smoothing of the *a priori* SAP in frequency domain to reduce the musical noise. The procedure of estimating a priori SAP is as follows.

A priori SAP is derived using an adaptive recursive averaging with an instantaneous SAP, $\tilde{q}(k,l)$.

$$\begin{aligned} q(k,l) &= (1-|\alpha_q(k,l)|)\; q(k,l-1) + |\alpha_q(k,l)|\; \tilde{q}(k,l) \\ \alpha_q(k,l) &= \sin\left\{\tfrac{\pi}{2}\left(\tilde{q}(k,l) - q(k,l-1)\right)\right\}, \end{aligned} \quad (5)$$

where $\alpha_q(k,l)$ is the adaptive recursive averaging factor which is derived using the difference between the *a priori* SAP of the previous frame and $\tilde{q}(k,l)$. The sine function is used for fast tracking of *a priori* SAP when the difference with the SAP in the previous frame, $q(k,l-1)$ is large. In the conventional method of recursive averaging, the $\alpha_q(k,l)$ is set to have a constant value, thus the tracking speed of the averaging is fixed. Since the characteristic of speech signal is varied dynamically, the fixed adaptation of $q(k,l)$ is not efficient. Fig. 1 represents the result of recursive averaging when we assume that the *a priori* SAP of previous frame is '0'. The tracking speed of proposed method changes nonlinearly depending on the distance between $q(k,l-1)$ and $\tilde{q}(k,l)$.

In the proposed algorithm, the instantaneous SAP is derived by merging two parameters.

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

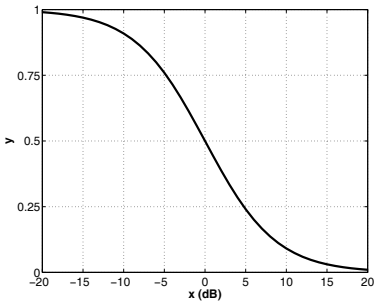

Fig. 2. curve of $y = 1/(1+x)$

where $B(l)$ represents a soft-decision factor describing the speech absence probability of the current frame, and $I(k,l)$ is a factor related to kth spectral component. If $B(l)$ is close to '1', $\tilde{q}(k,l)$ would be almost same as $B(l)$, so it becomes constant in the consecutive frame. However, for the small value of $B(l)$, $\tilde{q}(k,l)$ is affected by the characteristics of each frequency component.

$B(l)$ and $I(k,l)$ of equation (6) is defined as the normalized distance between the threshold and the smoothed *a posteriori* SNR.

$$B(l) = \frac{\bar{\gamma}_{th}}{\bar{\gamma}_{th} + \gamma_{med}(l)} = \frac{1}{1 + \frac{\gamma_{med}(l)}{\bar{\gamma}_{th}}} \tag{7}$$

$$I(k,l) = \frac{\gamma_{th}}{\gamma_{th} + \zeta(k,l)} = \frac{1}{1 + \frac{\zeta(k,l)}{\gamma_{th}}}, \tag{8}$$

where the $\gamma_{med}(l)$ and $\zeta(k,l)$ are the median value of the *a posteriori* SNR and the smoothed *a posteriori* SNR, e.g.

$$\gamma_{med}(l) = \underset{k=\langle M \rangle}{\text{med}} \left(\gamma\left(k,l\right)\right) \tag{9}$$

$$\zeta(k,l) = \alpha\, \gamma(k,l) + (1-\alpha) \underset{k=\langle M(k') \rangle}{\text{med}} \left(\gamma\left(k,l\right)\right), \tag{10}$$

where '$\text{med}(f(k))$' represents the median of $f(k)$. The median function is used instead of normal averaging to avoid the mis-decision of the average value due to absolutely big or small value in the specified averaging region. The efficiency of median for noise reduction is proved by simulation results. Note that the $\gamma_{med}(l)$ and $\zeta(k,l)$ is used only for *a priori* SAP smoothing.

In the equation (9) and (10), M is the number of total frequency bins and $M(k')$ is the number of frequency bins (approximated critical band rate) at k'th critical band. In other words, k' represents the index of approximated critical band and the median in equation (10) is calculated for each critical band rate. Critical band rate is used to avoid the degradation of the speech clearness (see Appendix A). The effect of musical noise is greater at high frequency region because it includes less speech components but wider masking effects than low frequency region. Therefore, the wide region of smoothing at the high frequency region is efficient for musical noise reduction. The relationship between $\gamma_{med}(l)$ and $B(l)$ or $\zeta(k,l)$ and $I(k,l)$ is plotted in Fig. 2 by assuming that the threshold is '1'. As you see in Fig. 2, the equation like (7) or (8) is good for fast tracking of *a priori* SAP.

We can further reduce the amount of noise by applying an ad-hoc processing to noise only frame. If the average of $q(k,l)$ is greater than some constant,

$$\frac{1}{M} \sum_{k=0}^{M-1} \{q(k,l)\} > q_{th}, \tag{11}$$

then the frame l is assumed as a noise only frame and its *a priori* SAP is set to pre-defined maximum. In this work, we set q_{th} to 0.85 and maximum of *a priori* SAP to 0.99.

4. SMOOTHING OF GAIN FUNCTION

In section 3, we mentioned that the smoothing operation reduced the musical noise. Simulation results show that the usage of both *a priori* SAP and gain smoothing is more helpful. However, the gain smoothing directly affects to the clearness/crispness of speech, thus the bandwidth for smoothing should be carefully designed. We use the one defined in 'Appendix A' to avoid the perceptual degradation of speech.

$$\tilde{G}(k,l) = \beta\, G(k,l) + (1-\beta) \underset{k=\langle M(k') \rangle}{\text{med}} \left(G\left(k,l\right)\right) \tag{12}$$

The similar function to eq. (10) is used for gain smoothing. By controlling the smoothing factor β, we can adjust whether we focus on reducing the level of musical noise or improving the clearness of speech. In our experiments, the good result can be obtained if we set β to 0.2.

5. PERFORMANCE EVALUATION AND DISCUSSION

The proposed *a priori* SAP estimation algorithm is implemented into MMSE LSA systems. The gain modifier which has been reported at [2] are used for the system. The noise power spectral density(PSD) is derived using the minimum statistic algorithm by Martin [8]. The noisy speech degraded by three types of noise such as white, babble, and factory noises with 10dB and 0dB SNR is used to evaluate the performance of the systems. The noise sources are taken from the NOISEX-92 database.

The reference system is designed using the algorithms described in [4] and [5]. Total 18 persons are asked to choose the one that has the best sound among two test signals (A/B test). The result is summarized in Table 1 and 2. The performance of the proposed algorithm is superior to the conventional ones in high SNR environments and it

Table 1. The number of person who chose the algorithm (SNR = 10dB)

algorithm	white noise	factory noise	babble noise
[3] with [4]	3 (16.7%)	3 (16.7%)	5 (27.8%)
proposed	15 (83.3%)	14 (77.8%)	7 (38.9%)
Don't know	0 (0.0%)	1 (5.6%)	6 (33.3%)
[3] with [5]	1 (5.6%)	1 (5.6%)	1 (5.6%)
proposed	17 (94.4%)	15 (83.3%)	12 (66.7%)
Don't know	0 (0.0%)	2 (11.1%)	5 (27.8%)

Table 2. The number of person who chose the algorithm (SNR = 0dB)

algorithm	white noise	factory noise	babble noise
[3] with [4]	8 (44.4%)	8 (44.4%)	5 (27.8%)
proposed	9 (50.0%)	8 (44.4%)	4 (22.2%)
Don't know	1 (5.6%)	2 (11.1%)	9 (50.0%)
[3] with [5]	3 (16.7%)	2 (11.1%)	8 (44.4%)
proposed	14 (77.8%)	9 (50.0%)	7 (38.9%)
Don't know	1 (5.6%)	7 (38.9%)	3 (16.7%)

also shows better or similar performance in low SNR environments. Table 3 and 4 show the log-spectral distance of enhanced speech. From the results, we also observe that the distance of the proposed algorithm is much smaller than other methods, which indicates that the proposed algorithm is also good in terms of the objective measure.

Appendix A. Critical band rate

The critical band concept is important for describing hearing sensations. Bandwidth of critical band or critical band rate is determined based on the spectral partitioning of our hearing system. Therefore, the smoothing of spectral components in the critical band region has less effect to the perception of speech.

In this paper, approximately one-third of critical band rate ($0.2f$) is used for defining the bandwidth of k' in (10)

Table 3. Log-spectral distance of enhanced signal(dB) (SNR = 10dB)

algorithm	white noise	factory noise	babble noise
unprocessed	18.4300	13.6339	10.3735
[3] with [4]	5.2342	3.8959	4.3005
[3] with [5]	4.7358	4.1044	4.8611
proposed	2.1235	1.6053	2.1709

Table 4. Log-spectral distance of enhanced signal(dB) (SNR = 0dB)

algorithm	white noise	factory noise	babble noise
unprocessed	26.8649	21.5299	17.3906
[3] with [4]	8.8344	6.6434	9.3802
[3] with [5]	7.8652	6.8687	10.5408
proposed	3.4307	3.1743	7.5530

[9]. Table 5 shows the approximated number of frequency bins according to the critical band rate. It assumes that the bandwidth of input signal is $4kHz$ and the window size of discrete Fourier transform is 256 points. Each number represents the number of frequency bins in k'th bark band.

Table 5. Approximated one-third critical band rate

$\mathbf{k'}$	0-26	27-33	34-38	39-42
the num. of bins	1	2	3	4
$\mathbf{k'}$	43-45	46,47	48,49	50,51
the num. of bins	5	6	7	8

6. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-27, pp. 113-120, 1979.

[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] 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.

[4] 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.

[5] 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.

[6] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE trans. Speech and Audio processing*, vol. 2, pp. 345-349, Apr. 1994.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE trans. Acoust., Speech, Signal processing*, vol. ASSP-28, pp. 137-145, Apr. 1980

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

[9] E. Zwicker and H. Fastl, *Psycho-acoustics fact and models*, Springer, 1999.