SPEECH ENHANCEMENT BASED ON MINIMA CONTROLLED RECURSIVE AVERAGING INCORPORATING CONDITIONAL MAXIMUM *A POSTERIORI* CRITERION

Jomg-Mo Kum, Yun-Sik Park and Joon-Hyuk Chang

School of Electronic Engineering
Inha University, Incheon, Korea
E-mail: [jmkum, yspark]@dsp.inha.ac.kr, changjh@inha.ac.kr

ABSTRACT

In this paper, we propose a novel approach to improve the performance of minima controlled recursive averaging (MCRA) based on a conditional maximum *a posteriori* (MAP) criterion. From an investigation of the MCRA scheme, it is discovered that the MCRA method cannot take full consideration of the inter-frame correlation of voice activity since the noise power estimate is adjusted by the speech presence probability depending on an observation of the current frame. To avoid this phenomenon, the proposed MCRA approach incorporates the conditional MAP criterion in which the noise power estimate is obtained using the speech presence probability conditioned on both the current observation and the speech activity decision in the previous frame. Experimental results show that the proposed MCRA technique based on conditional MAP yields better results compared to the conventional MCRA method.

Index Terms— Speech enhancement, Maximum *a posteriori* estimation

1. INTRODUCTION

The robustness of a speech enhancement system is significantly affected by the capability to reliably track noise statistics under adverse environments involving nonstationary noise, weak speech components and low signal-to-noise ratio (SNR) [1]-[9]. One of the successful noise power estimation is the minima controlled recursive averaging (MCRA) technique, which obtains the noise power estimate by averaging past spectral power values using a smoothing parameter adjusted by speech presence probability [10]. The resultant noise power estimate depending on the probability of speech presence is known to be computationally efficient and robust with respect to noise environments. However, this method is sensitive to temporal variation since the presence of speech in each subband is inherently estimated by the ratio of the local energy and its minimum of a given frame. Recently, Shin *et al.* proposed a novel technique to voice activity detection (VAD) considering the inter-frame correlation of voice activity based on the conditional maximum *a posteriori* (MAP) criterion [9]. The conditional MAP scheme was originally motivated by the observation that the speech presence of a given frame is dependent on not only the current observation but also on the voice activity in the previous frame.

In this paper, we propose a MCRA-based noise power estimation incorporating the conditional MAP criterion for speech enhancement. To determine efficiently the speech presence probability of the MCRA method, we consider the inter-frame correlation of speech activity based on the conditional MAP, which depends on both the current observation of a given frame and the speech presence decision in the previous frame. The proposed technique is substantially adopted for a speech enhancement technique and is evaluated with objective and subjective quality test experiments under various noise conditions.

2. REVIEW OF MINIMA CONTROLLED RECURSIVE AVERAGING TECHNIQUE

Let $y(n)$ denote a noisy speech signal that is the sum of a clean speech signal $x(n)$ and an uncorrelated additive noise signal $d(n)$; $y(n) = x(n)+d(n)$. Applying a discrete Fourier transform (DFT), we have in the time-frequency domain

$$Y(k,l) = X(k,l) + D(k,l) \tag{1}$$

where k is the frequency bin and l is the frame index, respectively. Given two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, which respectively indicate speech absence and presence, it is assumed that

$$\begin{aligned} H_0(k,l) : Y(k,l) &= D(k,l) \\ H_1(k,l) : Y(k,l) &= X(k,l) + D(k,l). \end{aligned} \tag{2}$$

Letting $\lambda_d(k,l) = E[|D(k,l)|^2]$ be the variance of noise, the MCRA method obtains the noise power estimate by applying a temporal recursive smoothing to the noise measurement

This work was partly supported by the IT R&D program of MKE/IITA [2008-F-045-01] and the research was financially supported by the Ministry of Knowledge Economy(MKE) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Strategic Technology.

during periods of speech absence as follows:

$$\begin{aligned} H_0'(k,l): \hat{\lambda}_d(k,l+1) &= \alpha_d \hat{\lambda}_d(k,l) + (1-\alpha_d)|Y(k,l)|^2 \\ H_1'(k,l): \hat{\lambda}_d(k,l+1) &= \hat{\lambda}_d(k,l) \end{aligned} \quad (3)$$

where $\alpha_d (0 < \alpha_d < 1)$ is a smoothing parameter. Also, H_0' and H_1' designate hypothetical speech absence and presence, respectively. In [10], a clear distinction is made between the hypotheses in (2), which is used for estimation of clean speech, and the hypotheses in (3), which controls adaptation of the noise statistics. In other words, deciding speech is absent (H_0) when speech is present (H_1) is more destructive when estimating the signal than when estimating noise. Therefore, different decision rules are employed, in that H_1 is chosen to get a higher confidence than H_1' *i.e.*, $P(H_1|Y(t)) \geq P(H_1'|Y(t))$ [10], where $P(H_1'(k,l)|Y(k,l))$ denote the conditional *a posteriori* probability of speech presence. Then, (3) implies

$$\begin{aligned} \hat{\lambda}_d(k,l+1) &= \hat{\lambda}_d(k,l) P(H_1'(k,l)|Y(k,l)) \\ &\quad + [\alpha_d \hat{\lambda}_d(k,l) + (1-\alpha_d)|Y(k,l)|^2] \\ &\quad \times (1 - P(H_1'(k,l)|Y(k,l))) \\ &= \hat{\alpha}_d(k,l)\hat{\lambda}_d(k,l) \\ &\quad + [1 - \hat{\alpha}_d(k,l)]|Y(k,l)|^2 \end{aligned} \quad (4)$$

where $\hat{\alpha}_d$ is a time-varying smoothing parameter that is adjusted by the speech presence probability as follows:

$$\hat{\alpha}_d(k,l) = \alpha_d + (1-\alpha_d) P(H_1'(k,l)|Y(k,l)). \quad (5)$$

The conditional speech presence probability, $P(H_1'(k,l)|Y(k,l))$, is calculated

$$P(H_1'(k,l)|Y(k,l)) = \begin{cases} \alpha_p P(H_1'(k,l-1)|Y(k,l-1)) \\ +(1-\alpha_p), \quad \text{if } S_r(k,l) > \delta \\ \alpha_p P(H_1'(k,l-1)|Y(k,l-1)), \\ \qquad \text{otherwise} \end{cases} \quad (6)$$

where $\alpha_p (0 < \alpha_p < 1)$ is a smoothing parameter and δ is a probability threshold of speech signal presence. Also, $S_r(k,l) \triangleq S(k,l)/S_{min}(k,l)$ denotes the ratio between the local energy of noisy speech, $S(k,l)$ and its determined minimum, $S_{min}(k,l)$ for the current frame. Here, $S_r(k,l)$ is originally based on a Bayes minimum-cost decision rule given by

$$\frac{P(S_r(k,l)|H_1(k,l))}{P(S_r(k,l)|H_0(k,l))} \underset{H_0'}{\overset{H_1'}{\gtrless}} \alpha \frac{C_{10} P(H(k,l)=H_0)}{C_{01} P(H(k,l)=H_1)} \quad (7)$$

where $P(H(k,l)=H_0) = (1 - P(H(k,l)=H_1))$ is the *a priori* probability of speech absence and C_{ij} is the cost of deciding H_i' when H_j'. Since (7) is a monotonic function, the decision rule (7) is expressed by $S_r(k,l) \underset{H_0'}{\overset{H_1'}{\gtrless}} \delta$ like (6), in which $\delta \triangleq \alpha \frac{C_{10} P(H(k,l)=H_0)}{C_{01} P(H(k,l)=H_1)}$. To obtain $S(k,l)$ recursive averaging is employed such that

$$S(k,l) = \zeta_s S(k,l-1) + (1-\zeta_s)|Y(k,l)|^2 \quad (8)$$

where $\zeta_s (0 < \zeta_s < 1)$ is a smoothing parameter. In addition, to obtain the minimum value $S_{min}(k,l)$ of the current frame, a samplewise comparison of the local energy and the minimum value of the previous frame is performed [10].

3. ENHANCED MCRA BASED ON CONDITIONAL MAP

In the previous section, we note that the crucial parameter in the MCRA method is $P(H_1'(k,l)|Y(k,l))$, controlling the smoothing parameter for the recursively averaged power spectrum of the noisy signal. However, $P(H_1'(k,l)|Y(k,l))$ does not consider the inter-frame correlation carefully since the conditional probability of speech presence incorporating a Bayes decision rule is simply based on a current observation $Y(k,l)$. Since it is known that speech activities in the adjacent frames have a strong correlation representing $P(H(k,l) = H_1|H(k,l-1) = H_1) > P(H(k,l) = H_1)$ where $H(k,l)$ denotes the correct hypothesis at the kth frequency bin in the lth frame [9], [11], we first consider the decision rule conditioned on both the current observation and the speech presence decision in the previous $(l-1)$ frame as follows:

$$\frac{P(H(k,l)=H_1|S_r, H(k,l-1)=H_i)}{P(H(k,l)=H_0|S_r, H(k,l-1)=H_i)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \alpha, \quad i = 0, 1. \quad (9)$$

Using Bayes' rule, this criterion results in the following likelihood ratio test (LRT) which is analogous to (7).

$$\frac{P(S_r|H(k,l)=H_1, H(k,l-1)=H_i)}{P(S_r|H(k,l)=H_0, H(k,l-1)=H_i)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \alpha_i', \quad i = 0, 1 \quad (10)$$

where $\alpha_i' = \alpha \frac{P(H(k,l)=H_0|H(k,l-1)=H_i)}{P(H(k,l)=H_1|H(k,l-1)=H_i)}$. It is assumed that speech presence of the current frame dominantly determines the distribution of the observed signal in the current frame [9]. Therefore (10) can be simplified as follows:

$$\frac{P(S_r|H(k,l)=H_1)}{P(S_r|H(k,l)=H_0)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \alpha_i', \quad i = 0, 1. \quad (11)$$

From this, we know that the threshold $\alpha_0' (= \alpha \frac{P(H(k,l)=H_0|H(k,l-1)=H_0)}{P(H(k,l)=H_1|H(k,l-1)=H_0)})$ is used if absence of the speech signal is detected in the previous frame (*i.e.*, $P(H(k,l-1)=H_0|Y(k,l-1)) \approx 1$) while the threshold $\alpha_1' (= \alpha \frac{P(H(k,l)=H_0|H(k,l-1)=H_1)}{P(H(k,l)=H_1|H(k,l-1)=H_1)})$ is adopted otherwise. It should be noted that multiple thresholds can provide more accurate speech presence probabilities rendering the improved

noise power estimator. Finally, the proposed method combines the above double decision rules using the soft decision scheme as follows:

$$S_r \underset{H_0'}{\overset{H_1'}{\gtrless}} \eta \tag{12}$$

where $\eta = P(H'(k,l-1) = H_1|Y(k,l-1))\alpha_0' + (1 - P(H'(k,l-1) = H_1|Y(k,l-1))\alpha_1'$. From (12), we can see that α_0' replaces η in the case of speech presence in the previous frame and α_1' becomes more dominant as speech presence probability of the previous frame decreases. This method can be considered desirable because of taking full considerations of the soft decision scheme. In the proposed approach speech presence probability is more accurate than the conventional method, and this can account for the improved noise power estimation in the MCRA method.

As a main application of the proposed technique, we adopt a speech enhancement algorithm based on a minimum mean-square error (MMSE) as following [1], [8]:

$$\hat{X}(k,l) = G(\xi(k,l), \gamma(k,l))Y(k,l) \tag{13}$$

where $\hat{X}(k,l)$ is the estimated clean speech spectrum and $G(\cdot)$ indicates the noise suppression gain. Also, $\gamma(k,l)$ is the *a posteriori* SNR and $\xi(k,l)$ is the *a priori* SNR defined by

$$\gamma(k,l) \equiv \frac{|Y(k,l)|^2}{\lambda_d(k,l)}, \tag{14}$$

$$\xi \equiv \frac{\lambda_x(k,l)}{\lambda_d(k,l)}. \tag{15}$$

Here, the MMSE noise suppression gain is given by

$$G(\hat{\xi}(k,l), \hat{\gamma}(k,l)) = \frac{\sqrt{\pi\nu(k,l)}}{2\hat{\gamma}(k,l)} exp(-\frac{\nu(k,l)}{2}) \times[(1+\nu(k,l))I_0(\frac{\nu(k,l)}{2}) + \nu(k,l)I_1(\frac{\nu(k,l)}{2})] \tag{16}$$

in which I_0 and I_1 are the modified Bessel functions of zero and first order. Also, $\nu(k,l)$ is defined based on (4), (14) and (15) as given by

$$\nu(k,l) = \frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)}\hat{\gamma}(k,l) \tag{17}$$

where $\hat{\xi}(k,l)$ and $\hat{\gamma}(k,l)$ is inherently estimated using (14) and (15).

4. EXPERIMENTAL RESULTS

The proposed approach was adopted for the MCRA-based speech enhancement method as given in [1], [8], [10] and was evaluated with a quantitative comparison and subjective quality test experiment under various noise environments. One hundred test phrases, spoken by four male and four female speakers, were used as experimental data. Each phrase consisted of two different meaningful sentences and lasted 8 s. The MCRA-based noise power estimation was performed for each frame of 10 ms duration with a sampling frequency of 8 kHz. Three types of noise sources such as white, babble and F16 noises from the NOISEX-92 database were added to the clean speech waveform at SNRs of 5, 10 and 15 dB. In all cases, speech enhancement was conducted with the experimentally optimized parameter values; $\alpha_d = 0.95$, $\alpha_p = 0.2$, $\zeta_s = 0.45$. To determine the parameters α_0' and α_1', we investigated the conditional probabilities $P(H(l) = H_0|H(l-1) = H_0)(= 39.1\%)$, $P(H(l) = H_0|H(l-1) = H_1)(= 1.2\%)$, $P(H(l) = H_1|H(l-1) = H_0)(= 1.2\%)$ and $P(H(l) = H_1|H(l-1) = H_1)(= 58.5\%)$ based on the speech material which length was 456 s. In this regards, α_0' and α_1' used in the experiment were $\alpha_0' = 9$ and $\alpha_1' = 3.5$.

Table 1. Relative Estimation Error Obtained from the MCRA and Proposed Method

Noise	Method	SNR (dB)		
		5	10	15
White	MCRA	0.379	0.475	0.578
	Proposed	0.370	0.384	0.421
Babble	MCRA	0.779	0.786	1.486
	Proposed	0.680	0.702	0.730
F16	MCRA	0.470	0.620	1.444
	Proposed	0.379	0.367	0.432

First, the performance of noise power estimation was evaluated frame by frame based on the normalized relative estimation error ε_n defined by [10].

$$\varepsilon_n = \frac{1}{N}\sum_{l=0}^{N-1}\frac{\sum_k[\hat{\lambda}_d(k,l) - \lambda_d(k,l)]^2}{\sum_k {\lambda_d}^2(k,l)} \tag{18}$$

where $\lambda_d(k,l)$ is the actual noise power estimate directly obtained by the noise signal [10], $\hat{\lambda}_d(k,l)$ is the noise power estimated by the conventional MCRA method and the proposed method for the number of frames N. Table I shows the results of the relative estimation error for the given noise power estimation methods under the various noise conditions. From the results, it can be seen that the proposed scheme gave us an improvement over the previous MCRA approach [10].

For the purpose of evaluating the performance of the presented method in terms of speech quality, we first measured a perceptual evaluation of speech quality (PESQ) based on the ITU-T P.862 tests [12]. Table II presenting the results of the PESQ shows that the proposed conditional MAP-based MCRA approach outperformed the original MCRA-based scheme under the given noise conditions, where we can see that incorporation of the conditional MAP definitely has a positive effective in terms of the quantitative comparison.

Secondly, to evaluate the subjective quality of the proposed scheme, we carried out a set of informal tests under the

Table 2. PESQ scores of the MCRA and Proposed Method.

Noise	Method	SNR (dB)		
		5	10	15
White	MCRA	1.947	2.295	2.636
	Proposed	2.034	2.373	2.701
Babble	MCRA	2.316	2.646	2.927
	Proposed	2.339	2.668	2.956
F16	MCRA	2.030	2.442	2.757
	Proposed	2.123	2.515	2.820

Table 3. MOS of the MCRA and Proposed Method.

Noise	Method	SNR (dB)		
		5	10	15
White	MCRA	2.17	2.56	3.03
	Proposed	2.23	2.65	3.07
Babble	MCRA	2.93	3.32	3.69
	Proposed	2.95	3.42	3.76
F16	MCRA	2.23	2.77	3.20
	Proposed	2.33	2.90	3.46

same noise conditions. Subjective opinions were decided by a group of 20 listeners; each listener gave a score for each test sentence: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad). All listener scores were then averaged to yield a mean opinion score (MOS). The MOS test results are summarized in Table III, in which a higher value indicates preference. Table III demonstrates that the proposed approach improved over the conventional MCRA method under the given conditions. It is noted that performance improvement was found to be greater for the F16 noise case at all SNRs. These results confirm that the proposed conditional MAP scheme substantially improves the MCRA method in speech quality enhancement.

5. CONCLUSIONS

In this paper, we have proposed a novel approach to incorporate conditional MAP into the conventional MCRA-based noise power estimation scheme. The noise power estimate was given by the recursive smoothing parameter incorporating speech absence probability conditioned on both the current observation and the voice activity decision in the previous frame. Compared to the conventional MCRA method, it was demonstrated that the proposed technique provides better noise power estimates for the speech enhancement systems.

6. REFERENCES

[1] 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, no. 6, pp. 1109-1121, Dec. 1984.

[2] 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-32, no. 2, pp. 443-445, Apr. 1985.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1979.

[4] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO*, Edinburgh, U.K., pp. 1182-1185, Sept. 1994.

[5] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. EUROSPEECH*, Madrid, Spain, pp. 1513-1516, Sept. 1995.

[6] J. Meyer, K. U. Simmer and K. D. Kammeyer, "Comparison of one-and two-channel noise-estimation techniques," in *Proc. IWAENC*, London, U.K., pp. 137-145, Sept. 1997.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, Nov. 2001.

[8] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[9] J. W. Shin, H. J. Kwon, S. H. Jin and N. S. Kim, "Voice activity detection based on conditional MAP criterion," *IEEE Signal Processing Letters*, vol. 15, pp. 257-260, Feb. 2008.

[10] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-15, Jan. 2002.

[11] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no 1, pp. 1-3, Jan. 1999.

[12] ITU-T P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.