ADAPTIVE NOISE POWER ESTIMATION USING SPECTRAL DIFFERENCE FOR ROBUST SPEECH ENHANCEMENT

Jae-Hun Choi, Sang-Kyun Kim and Joon-Hyuk Chang

School of Electronic Engineering
Hanyang University, Seoul, Korea
E-mail: greatestcjh@naver.com, greenwhity@nate.com, jchang@hanyang.ac.kr (Corresponding author)

ABSTRACT

In this paper, we propose a spectral difference approach for noise power estimation in speech enhancement. The noise power estimate is given by recursively averaging past spectral power values using a smoothing parameter based on the current observation. The smoothing parameter in time and frequency is adjusted by the spectral difference between consecutive frames that can efficiently characterize noise variation. Specifically, we propose an effective technique based on a sigmoid-type function in order to adaptively determine the smoothing parameter based on the spectral difference. Compared to a conventional method, the proposed noise estimate is computationally efficient and able to effectively follow noise changes under various noise conditions.

Index Terms— Noise Estimation, Spectral Difference, Sigmoid function

1. INTRODUCTION

Noise power estimation is a crucial component of speech enhancement. System performance is greatly affected by low signal-to-noise ratio (SNR) conditions and non-stationary noise environments since it is difficult to reliably track rapid variation over a varying noise spectrum [1]-[2]. Soft decision (SD), the well-known noise power estimation technique, has been successfully adopted as a fundamental module of speech enhancement systems [3]. Specifically, the long-term smoothed power spectrum of noise that depends on the probability of speech absence is used. Indeed, the speech absence probability (SAP) is induced from a likelihood ratio test (LRT) based on a statistical model of speech. However, this method needs to be improved since it does not fully consider non-stationary noise. In the case of non-stationary noise, it is difficult to efficiently update the noise power when a fixed long-term smoothing parameter is used. To solve this problem, Cohen [4] proposed a recursive averaging technique controlled by the minima (called MCRA). However, this technique is insensitive to spectral variation because it is inherently based on the local energy of noisy speech and its minimum is derived using an indicator function.

The goal of this paper is to incorporate the spectral difference between two successive frames when updating the noise power. This is due to the clear difference in stationary noise and non-stationary noise in terms of the spectro-temporal properties. This difference allows us to adaptively increase the noise update rate only for non-stationary noise, with the spectrum level changing rapidly in time. In this regard, in contrast to the conventional method with a fixed smoothing factor, the proposed noise update parameter utilizes the sigmoid function for the adaptive factor when the smoothing parameter is determined by the spectral difference between the current observation and the data in the previous frame. In this paper, the proposed method is evaluated in the known speech enhancement algorithm and experimental results of speech quality are given for a data set of real noise samples.

2. REVIEW OF SOFT DECISION BASED-SPEECH ENHANCEMENT

We first briefly review soft decision-based speech enhancement. It is assumed that a noise signal d is added to a speech signal x, with their sum being denoted by a noisy speech signal y. After taking the discrete Fourier transform (DFT) of the noisy signal y, we then have in the time-frequency domain

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

where k is the frequency-bin index ($k = 0, 1, \cdots, M-1$) and t is the frame index. Assuming that speech is degraded by uncorrelated additive noise, two hypotheses, H_0 and H_1, which indicate speech absence and presence, respectively, are given by

$$\begin{aligned} H_0 &: \text{speech absent} : Y(t,k) = D(t,k) \\ H_1 &: \text{speech present} : Y(t,k) = X(t,k) + D(t,k). \end{aligned} \tag{2}$$

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (NRF-2011-0009182). And, this research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-C1090-1121-0007).

With the complex Gaussian probability density functions (pdfs) assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by

$$p(Y(t,k)|H_0) = \frac{1}{\pi\lambda_d(t,k)} \exp\left\{-\frac{|Y(t,k)|^2}{\lambda_d(t,k)}\right\} \quad (3)$$

$$p(Y(t,k)|H_1) = \frac{1}{\pi(\lambda_x(t,k)+\lambda_d(t,k))} \cdot \exp\left\{-\frac{|Y(t,k)|^2}{\lambda_x(t,k)+\lambda_d(t,k)}\right\} \quad (4)$$

where $\lambda_x(t,k)$ and $\lambda_d(t,k)$ denote the variances of X_k and D_k, respectively. If the spectral component in each frequency bin is assumed to be statistically independent, then we have the global speech absence probability (GSAP) conditioned on the current observation of $Y(t)(= \{Y(t,0), Y(t,1), \cdots, Y(t,M-1)\})$ using Bayes'rule as in [3] such that

$$p(H_0|Y(t)) = \frac{1}{1+\frac{p(H_1)}{p(H_0)}\prod_{k=1}^{M}\Lambda(Y(t,k))} \quad (5)$$

where $P(H_0)(= 1 - P(H_1))$ is the *a priori* probability for speech absence and $\Lambda_k(t)$ is the likelihood ratio (LR) in the kth frequency bin [3].

The SD method adopts the long-term smoothed noise power estimate $\hat{\lambda}_d(t,k)$, which is given by [3]:

$$\hat{\lambda}_d(t+1,k) = \zeta_d\hat{\lambda}_d(t,k) + (1-\zeta_d)E[|D(t,k)|^2|Y(t,k)] \quad (6)$$

where ζ_d is the smoothing parameter, which is set to 0.99 in [3].

3. PROPOSED APPROACH BASED ON SOFT DECISION EMPLOYING SPECTRAL DIFFERENCE

In the previous section, we reviewed the noise estimation method based on soft decisions where the noise power spectrum is estimated by recursively averaging past spectral power values using the fixed long-term smoothing parameter given by ($\zeta_d = 0.99$) under a general stationarity assumption of noise power spectra [3]. However, the detection reliability severely deteriorates for non-stationary noise environments since the fixed smoothing parameter restricts the robust tracking capability of the noise estimator. For example, a lower value in ζ_d for non-stationary noise results in a faster response to noise adaptation, but may increase fluctuations between non-speech and speech segments. In contrast, a higher value ($\cong 1$) of ζ_d to handle the stationary noise for stability results in slow adaptation in the case of non-stationary noise. Thus, we need a smoothing parameter that is adjusted by the amount of background noise that is non-stationary for robust noise adaptation. For this, we first define the spectral difference between adjacent noise frames in order to identify the degree of the noise non-stationarity. Specifically, the spectral difference $\Delta(t,k)$ at the kth frequency bin in the tth frame is defined by the normalized difference between $|Y(t,k)|^2$ and $|Y(t-1,k)|^2$ such that

$$\Delta(t,k) = \frac{||Y(t,k)|^2 - |Y(t-1,k)|^2|}{\overline{Y}(t)} \quad (7)$$

where $\overline{Y}(t) = \frac{1}{M}\sum_{k=0}^{M-1}|Y(t,k)|^2$ denote the average noise power of the current frame. We then obtain the geometric mean of $\Delta(t,k)$ for the individual frequency bins, given by:

$$\Delta(t) = \frac{1}{M}\sum_{k=0}^{M-1}\Delta(t,k). \quad (8)$$

Subsequently, the long-term smoothing is performed such that

$$\overline{\Delta}(t) = \alpha_d\overline{\Delta}(t-1) + (1-\alpha_d)\Delta(t) \quad (9)$$

where $\alpha_d (0 < \alpha < 1)$ is a parameter for smoothing the spectral difference. Also, it should be noted that the update routine for $\overline{\Delta}(t)$ should be given during the speech absence since we are interested in the spectral difference in noise. For this, we derive an updated routine of $\overline{\Delta}(t)$ by utilizing the soft decision scheme and (5) as follows:

$$\overline{\Delta}(t) = \left\{\alpha_d\overline{\Delta}(t-1) + (1-\alpha_d)\Delta(t)\right\}p(H_0|Y(t)) + \overline{\Delta}(t-1)(1-p(H_0|Y(t))) \quad (10)$$

where (12) becomes (11) when $p(H_0|Y(t)) = 1$ while $\overline{\Delta}(t)$ is not updated (i.e., $\overline{\Delta}(t) \cong \overline{\Delta}(t-1)$) and sustained in the case of $p(H_0|Y(t)) = 0$.

Note that an estimator from $\overline{\Delta}(t)$ could be a relevant measure to take into account the degree of noise non-stationarity since $\overline{\Delta}(t)$ becomes high when the noise characteristics vary quickly in terms of power, as in the case of the nonstationary noise. In contrast, $\overline{\Delta}(t)$ yields a lower value in the case of the stationary noise, the characteristics of which change slowly. The more nonstationary the noise source (babble compared to white), the larger $\overline{\Delta}(t)$ gets, as we expect (in Fig. 1). Also, Fig. 2 shows a representative example of $\overline{\Delta}(t)$, showing the histogram of $\overline{\Delta}(t)$ under various noise types. From these figures, it is evident that $\overline{\Delta}(t)$ varies depending on noise fluctuation. $\overline{\Delta}(t)$ can be considered a control factor to adjust the noise spectrum depending on the change in noise power level.

To cope with this idea, we propose the adaptive weighting factor incorporating a sigmoid type function. Specifically, an adaptive value based on the sigmoid type function according to $\overline{\Delta}(t)$ is applied to the long-term smoothing parameter in the noise update as shown below

$$\zeta_d^{SD}(t) = \frac{\delta\exp[-\beta(s(t)-s_0)]}{1+\exp[-\beta(s(t)-s_0)]} + \sigma \quad (11)$$

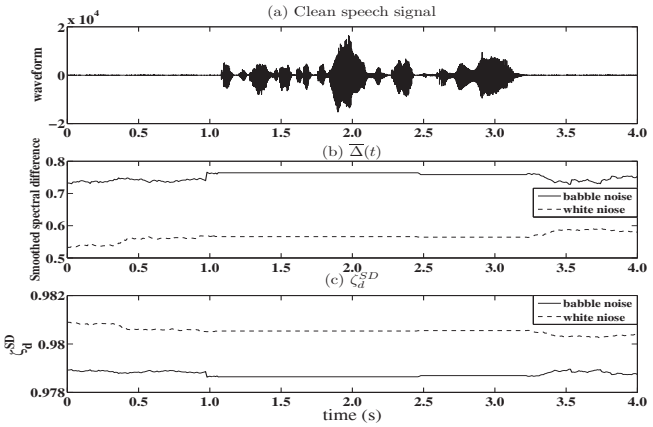

Fig. 1. Comparison of $\overline{\Delta}(t)$ for noisy speech corrupted by the babble and white noise (SNR = 5dB).

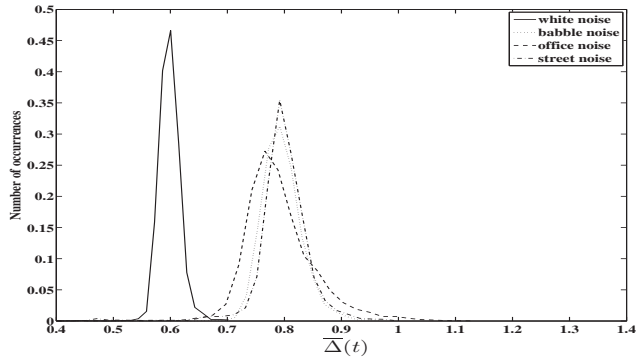

Fig. 2. Normalized distribution of spectral difference for the various noisy speech signal.

with

$$s(t) = \log(1/\overline{\Delta}(t)) \tag{12}$$

where $\beta(>0)$ is the slope parameter and s_0 denotes an offset. Through extensive speech enhancement experiments, $\zeta_d^{SD}(t)$ is obtained using $\beta = 0.75$, $s_0 = -9$, the constant $\delta = -19$, and $\sigma = 0.999$. Notice that the sigmoid type function of (13) makes the long-term smoothing parameter, ζ_d^{SD}, inversely proportional to $\overline{\Delta}(t)$ while limiting the value to the interval, $(\zeta_{d,\min}, \zeta_{d,\max})$. The high transition of noise spectral power is characterized by the increase of $\Delta(t)$. Increasing $\overline{\Delta}(t)$ results in decreasing $\log(1/\overline{\Delta}(t))$, and thus the proposed weighting factor, ζ_d^{SD}, approaches $\zeta_{d,\min}$, which makes it possible to update the noise power more quickly. On the other hand, decreasing $\overline{\Delta}(t)$ results in increasing $\log(1/\overline{\Delta}(t))$ on each frame, and thus a high value of the weighting factor is applied to update noise on a more robust value. As a result, the proposed estimation for the noise power is given as

$$\begin{aligned}\hat{\lambda}_d^{SD}(t+1,k) &= \zeta_d^{SD}(t)\hat{\lambda}_d^{SD}(t,k) \\ &+ (1-\zeta_d^{SD}(t))E[|D(t,k)|^2|Y(t,k)].\end{aligned} \tag{13}$$

The present scheme can efficiently update the noise power in the case of nonstationary noise in which the spectrum vary rapidly. This implies that the proposed noise power estimate is more accurate than the previous noise power estimate, which has a fixed long-term smoothing parameter, and could improve the performance of speech enhancement. As an application of the presented technique, we adopt a speech enhancement algorithm based on a minimum mean-square error (MMSE) as follows:

$$\hat{X}(t,k) = G(\hat{\xi}(t,k), \hat{\gamma}(t,k))Y(t,k) \tag{14}$$

where $\hat{X}(t,k)$ is the estimated clean speech spectrum and $G(\cdot,\cdot)$ denotes the noise suppression gain. The noise suppression gain is given by

$$\begin{aligned}&G(\hat{\xi}(t,k), \hat{\gamma}(t,k)) \\ &= \frac{\sqrt{\pi\nu(t,k)}}{2\hat{\gamma}(t,k)} \exp\left(-\frac{\nu(t,k)}{2}\right) \\ &\cdot \left[(1+\nu(t,k)I_0\left(\frac{\nu(t,k)}{2}\right) + \nu(t,k)I_1\left(\frac{\nu(t,k)}{2}\right)\right]\end{aligned} \tag{15}$$

in which I_0 and I_1 are the modified Bessel function of the zero and first orders. $\nu(t,k)$ is defined using $\hat{\xi}(t,k)$ and $\hat{\gamma}(t,k)$ as

$$\nu(t,k) = \frac{\hat{\xi}(t,k)}{1+\hat{\xi}(t,k)}\hat{\gamma}(t,k) \tag{16}$$

where

$$\hat{\xi}(t,k) = \frac{\hat{\lambda}_x(t,k)}{\hat{\lambda}_d^{SD}(t,k)},\ \hat{\gamma}(t,k) = \frac{|Y(t,k)|^2}{\hat{\lambda}_d^{SD}(t,k)}. \tag{17}$$

4. EXPERIMENTS AND RESULTS

The proposed adaptive noise power estimation technique was evaluated with an objective quality experiment under various noise conditions. The experimental data comprised 48 test phrases, where 24 were spoken by two male speakers and the other 24 were spoken by two female speakers. Each phrase consisted of two different meaningful sentences and lasted 8 s. The noise power estimation was performed for each frame of 10 ms duration with a sampling rate of 8 kHz. Four types of noise sources such as babble, office, street, and white noise from the NOISEX-92 database were added to the clean speech waveform at SNRs of 5, 10 and 15 dB. Table 1 shows the perceptual evaluation of speech quality (PESQ)

Table 1. PESQ results for the proposed algorithm with respect to the conventional method (SEGSD) and the IMCRA method under various background noise environments.

Environments		Method		
Noise	SNR (dB)	IMCRA [6]	SEGSD [3]	Proposed
Babble	5	2.438±0.07	2.477±0.08	2.509±0.08
	10	2.784±0.06	2.834±0.08	2.854±0.08
	15	3.066±0.06	3.110±0.07	3.116±0.07
Office	5	2.484±0.07	2.460±0.07	2.490±0.08
	10	2.810±0.06	2.796±0.07	2.815±0.07
	15	3.087±0.05	3.073±0.07	3.084±0.07
Street	5	2.806±0.05	2.734±0.07	2.749±0.07
	10	3.063±0.05	2.994±0.06	3.004±0.06
	15	3.310±0.05	3.231±0.06	3.242±0.06
White	5	2.402±0.08	2.210±0.07	2.271±0.07
	10	2.730±0.05	2.542±0.07	2.594±0.07
	15	2.968±0.04	2.850±0.06	2.887±0.06

Table 2. Overall quality (C_{ovl}) results for the proposed algorithm with respect to the conventional method (SEGSD) and the IMCRA method under various background noise environments.

Environments		Method		
Noise	SNR (dB)	IMCRA [6]	SEGSD [3]	Proposed
Babble	5	2.354±0.08	2.789±0.08	2.832±0.08
	10	2.780±0.08	3.217±0.07	3.249±0.07
	15	3.131±0.10	3.526±0.07	3.538±0.07
Office	5	2.715±0.07	2.893±0.08	2.923±0.08
	10	3.104±0.07	3.275±0.08	3.294±0.07
	15	3.417±0.07	3.570±0.07	3.583±0.07
Street	5	3.172±0.06	3.192±0.07	3.209±0.07
	10	2.473±0.07	2.497±0.06	3.509±0.06
	15	3.742±0.07	3.759±0.06	3.773±0.06
White	5	2.351±0.10	2.465±0.08	2.534±0.09
	10	2.697±0.11	2.834±0.08	2.895±0.08
	15	2.964±0.11	3.163±0.08	3.208±0.08

scores for various noise types and at various noise levels. The proposed method consistently achieves more improvement compared to the conventional method (the SEGSD method [3]). Its advantage is more significant in nonstationary and low-SNR noise environments. And, our approach achieved a higher improvement or was at least comparable to the previous adaptive noise power estimation technique (denoted by the IMCRA [6]) except the stationary noise such as white.

On the other hand, we used the well-known objective speech quality check method (called composite measure [5]) having the significant correlation with subjective quality as a combination of various representative objective quality measures, which was proposed in [5]. Specifically, the composite measure represents a five point scale of background noise intrusiveness and the overall quality using the mean opinion score (MOS) scale. Table 2 presents the results of the overall quality. These results show that the proposed method consistently results in superior performance compared to the SEGSD method and the IMCRA method. This finding is much valuable when considering the previous PESQ results, which show that our approach is effective for both the noise signal and the speech signal in terms of the subjective quality. The proposed method leads to better results in not only non-stationary noise such as babble, but also stationary noise such as white noise. This is attributable to the on-line noise adaptation for better subjective quality on a frame-by-frame basis, which depends on the spectral difference. Note that the spectral difference of the white noise can vary as shown in Fig. 1, and eventually requires adaptive updating of the noise power.

5. CONCLUSIONS

The proposed technique comprises steps for deriving the spectral difference and the adaptive smoothing parameter for updating the noise power. This is clearly different from conventional techniques because we restrict the updating of the noise estimator during speech absence or adapt the smoothing parameter according to the speech absence probability. Compared to the conventional method, the proposed noise estimate responds more efficiently to noise variation, when integrated into an MMSE-based speech enhancement system and yields effective performance under various noisy conditions.

6. REFERENCES

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

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, Jul. 2001.

[3] 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.

[4] I. Cohen, 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.

[5] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466-475, Sep. 2003.