A noise spectral estimation method based on VAD and recursive averaging using new adaptive parameters for non-stationary noise environments

著者	Nakayama Kenji, Higashi Shoya, Hirano Akihiro
journal or publication title	2008 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2008
page range	184187
year	2009-02-01
URL	http://hdl.handle.net/2297/18069

doi: 10.1109/ISPACS.2009.4806668

2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008)
Swissôtel Le Concorde,Bangkok,Thailand

A Noise Spectral Estimation Method Based on VAD and Recursive Averaging Using New Adaptive Parameters for Non-Stationary Noise Environments

Kenji Nakayama Shoya Higashi Akihiro Hirano
Graduate School of Natural Science and Technology, Kanazawa Univ., Japan
E-mail:nakayama@t.kanazawa-u.ac.jp

***Abstract*— A noise spectral estimation method, which is used in spectral suppression noise cancellers, is proposed for highly non-stationary noise environments. Speech and non-speech frames are detected by using the entropy-based voice activity detector (VAD). An adaptive normalization parameter and a variable threshold are newly introduced for the VAD. They are very useful for rapid change in the noise spectrum and power. Furthermore, a recursive averaging method is applied to estimating the noise spectrum in the non-speech frames. In this method, an adaptive smoothing parameter is proposed, based on speech presence probability. Simulations are carried out by using many kinds of noises, including white, babble, car, pink, factory and tank, which are changed from one to the other. The segmental SNR is improved by $2.0 \sim 3.8$dB, and noise spectral estimation error is improved by $3.2 \sim 4.7$dB for the white noise and the babble noise, which are changed from one to the other.**

I. Introduction

A spectral suppression technique is a hopeful approach to noise cancellers used in a mobile phone [1]. In this approach, it is very important to estimate a spectral gain, used to suppress the noise spectrum. Several methods, including MMSE STSA [2], MMSE LSA [3] and Joint MAP [4], have been proposed. Furthermore, performance of the spectral suppression technique is highly dependent on accuracy of the noise spectral estimation [5],[6]. There exist many kinds of noises. In highly non-stationary noise environments, power and spectrum of the noises can be dynamically changed. The noise spectral estimation should adapt this kind of changes quickly. Several noise spectral estimation methods have been proposed for non-stationary noise environments [7]∼[12].

In this paper, a noise spectral estimation method, which uses voice activity detection (VAD) and recursive spectral estimation, is proposed. An adaptive normalization parameter is proposed in the VAD. Furthermore, an adaptive smoothing parameter is proposed in the recursive averaging method, in order to estimate the noise spectrum in non-speech frames. Computer simulations by using speech signal and many kinds of noises will be shown.

II. Spectral Suppression Noise Canceller

Figure 1 shows a blockdiagram of the spectral suppression noise canceller. Spectra of speech and noise are assumed to be statistically independent. Let $s(m)$, $n(m)$ and $x(m)$ be noise free speech, noise and noisy speech, respectively.

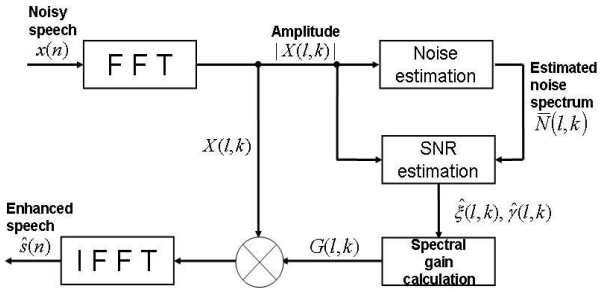

Fig. 1. Blockdiagram of spectral suppression noise canceller.

$$x(m) = s(m) + n(m) \tag{1}$$

The Fourier transform of $x(m)$, $s(m)$ and $n(m)$ in the lth frame and at the kth frequency bin are expressed by

$$X(l,k) = S(l,k) + N(l,k) \tag{2}$$

The prior SNR $\xi(l,k)$, a ratio of the clean speech power to the noise power, and the posterior SNR $\gamma(l,k)$, a ratio of the noisy speech power to the noise power, are defined by

$$\xi(l,k) = \frac{E[|S(l,k)|^2]}{E[|N(l,k)|^2]} \tag{3}$$

$$\gamma(l,k) = \frac{|X(l,k)|^2}{E[|N(l,k)|^2]} \tag{4}$$

Actually, the noisy speech signal $x(n)$ is only available. The prior SNR $\xi(l,k)$ is estimated as follows [2]:

$$\begin{aligned}\hat{\xi}(l,k) &= \alpha\gamma(l-1,k)G^2(l-1,k) \\ &+ (1-\alpha)P[\gamma(l,k)-1]\end{aligned} \tag{5}$$

where $0 < \alpha < 1$ and $P[x]$ satisfies

$$P[x] = \begin{cases} x, & x > 0 \\ 0, & otherwise \end{cases} \tag{6}$$

The posterior SNR $\gamma(l,k)$ can be estimated by using the noise spectrum estimation $\bar{N}(l,k)$ as follows:

$$\hat{\gamma}(l,k) = \frac{|X(l,k)|^2}{\bar{N}(l-1,k)} \tag{7}$$

978-1-4244-2565-5/08/$25.00 ©2008 IEEE

How to estimate $N(l,k)$ is a main issue in this paper. A spectral gain $G(l,k)$ is estimated by using the prior SNR $\hat{\xi}(l,k)$ and the posterior SNR $\hat{\gamma}(l,k)$, and is used to suppress the noise spectrum included in the noisy speech. In order to calculate $G(l,k)$, we employ MMSE STSA method [2] and Joint MAP method [4] in this paper.

III. Conventional Rapid Adaptation Method

In this section, a conventional noise spectral estimation method proposed for non-stationary noise environments is briefly descrived [9]-[12]. A voice activity detector (VAD) [7] is applied in this method.

A. Voice Activity Detector (VAD)

The VAD discriminates the speech frames and the non-speech frames based on the following entropy $H(l)$

$$P_r(l,k) = \frac{X_{energy}(l,k)}{\sum_{k=1}^{2M} X_{energy}(l,k)} \tag{8}$$

$$H(l) = -\sum_{k=1}^{2M} P_r(l,k) \cdot \log(P_r(l,k)) \tag{9}$$

$$X_{energy}(l,k) = |X(l,k)|^2 \tag{10}$$

The entropy $H(l)$ has a large value in the non-speech frames compared to the speech frames. We assume several frames at the beginning to be the non-speech frames. An average of the entropy, estimated in these frames, denoted $H_{av}(0)$ is used as the threshold, with which the following frames are discriminated as the speech or the non-speech frames. Actually, $H_{av}(0)$ is scaled by a constant $c(<1)$.

$$H(l) > cH_{av}(0) \quad \rightarrow \quad \text{Non-speech frame}$$
$$H(l) < cH_{av}(0) \quad \rightarrow \quad \text{Speech frame}$$

$H(l)$ is not accurate and cannot discriminate the non-speech frame and the speech frame, when the spectra of the speech and the noise are small and large, respectively. In order to improve this problem, a positive constant C has been introduced in $P_r(l,k)$ as follows [8]:

$$P_{rc}(l,k) = \frac{X_{energy}(l,k)+C}{\sum_{k=1}^{2M} X_{energy}(l,k)+C} \tag{11}$$

$$H_c(l) = -\sum_{k=1}^{2M} P_{rc}(l,k) \cdot \log(P_{rc}(l,k)) \tag{12}$$

B. Noise Spectral Estimation Method

The conventional noise spectral estimation method is briefly described here [9],[10]. The noisy speech is discriminated into the non-speech frames or the speech frames by using the VAD. The noise spectrum is estimated in the non-speech frames by

$$\bar{N}(l,k) = \lambda \cdot \bar{N}(l-1,k) + (1-\lambda) \cdot |X(l,k)|^2 \tag{13}$$

On the other hand, in the speech frames, the noise spectrum is estimated by the following recursive equation.

$$\bar{N}(l,k) = \rho(l,k) \cdot \bar{N}(l-1,k) + (1-\rho(l,k)) \cdot |X(l,k)|^2 \tag{14}$$

$$\rho(l,k) = a_d + (1-a_d) \cdot P_{sp}(l,k) \tag{15}$$

$P_{sp}(l,k)$ is a probability of including the speech in the noisy speech signal, that is a speech presence probability, and is given by

$$P_{sp}(l,k) = \frac{|X(l,k)|^2}{P_{min}(l,k)} \tag{16}$$

$P_{min}(l,k)$ is the minimum of the noisy speech spectrum, as shown in the following. First, the averaged spectrum of the noisy speech $P(l,k)$ is obtained by

$$P(l,k) = \eta P(l-1,k) + (1-\eta)|X(l,k)|^2 \tag{17}$$

η is a smoothing factor. Next, $P_{min}(l,k)$ is updated by the following equations.

$$P_{min}(l,k) = \gamma \cdot P_{min}(l-1,k) + \frac{1-\gamma}{1-\beta}(P(l,k) - \beta \cdot P(l-1,k)), \text{ if } P_{min}(l-1,k) \le P(l,k) \tag{18}$$

$$P_{min}(l,k) = P(l,k), \text{ if } P_{min}(l-1,k) > P(l,k) \tag{19}$$

β and γ are determined by experience.

IV. A New Noise Spectral Estimation Method

A. New Adaptive Parameter and Threshold for VAD

In the conventional method, as shown in Eq.(11), $P_{rc}(l,k)$ is normalized by using a constant C. The constant C is highly dependent on SNR of the noisy speech signal, and should be optimized. In this paper, we propose a new adaptive parameter, which is controlled by the difference between the maximum and the mean of $|X(l,k)|$ as follows:

$$P_{new}(l,k) = \frac{X_{energy}(l,k)+C_{new}(l)}{\sum_{k=1}^{2M} X_{energy}(l,k)+C_{new}(l)} \tag{20}$$

$$C_{new}(l) = \max\{|X(l,k)|\} - \text{mean}\{|X(l,k)|\} \tag{21}$$

$$H_{new}(l) = -\sum_{k=1}^{2M} P_{new}(l,k) \cdot \log(P_{new}(l,k)) \tag{22}$$

Since the highly non-stationary noise environments are considered, the noise power can be drastically changed. In order to adapt the noise power change, we also propose an adaptive threshold $H_{av}(l)$ for $H_{new}(l)$ in this paper. $H_{av}(l)$ is calculated by using the entropy in just before five non-speech frames. Therefore, $H_{av}(l)$ can be adjusted to the recent noise power. Actually, $cH_{av}(l), c=0.95$ is used for the threshold.

B. A New Noise Spectral Estimation Method

1) Non-Speech Frames: The proposed noise spectral estimation method is based on the recursive equation as shown in Eq.(14). How to control the smoothing parameter is very important. The conventional smoothing parameter $\rho(l,k)$ given by Eq.(15) is not useful for non-stationary noise environments. In this paper, we propose a new smoothing parameter, which is adjusted based on $P_{sp}(l,k)$ given by Eq.(16). The proposed noise spectral estimation is given by

$$\bar{N}(l,k) = \rho_{new}(l,k) \cdot \bar{N}(l-1,k) + (1-\rho_{new}(l,k)) \cdot |X(l,k)|^2 \tag{23}$$

$$\rho_{new}(l,k) = \frac{1}{1+\exp(-r \cdot (P_{sp}(l,k) - t \cdot T_p(l,k)))} \tag{24}$$

t is a constant. $T_p(l,k)$ is an adaptive threshold in the lth frame, and is calculated in the speech frames as follows:

$$T_p(l,k) = \frac{|X(l,k)|^2_{mean}}{\bar{N}_{mean}(l-1,k)} \quad (25)$$

$$|X(l,k)|^2_{mean} = E[|X(i,k)|^2] \quad (26)$$

$$\bar{N}_{mean}(l-1,k) = E[\bar{N}(i,k)] \quad (27)$$

(i $\in$ all speech frames, up to lth and $(l-1)th$ frames)

In the non-stationary noise environments, the noise spectrum can be changed. The estimation by using Eq.(13) does not work well. In the proposed method given by Eqs.(23) and (24), the adaptive threshold $T_p(l,k)$ is introduced for each frame. The smoothing parameter $\rho_{new}(l,k)$ can be adaptively controlled based on the probability of including the speech in the noisy speech signal. As a result, the noise spectrum in the non-stationary noise environments can be well estimated.

2) Speech Frames: In the conventional method, the recursive estimation was used as shown by Eq.(14). However, this method does not work well for the non-stationary noise environments. Estimation accuracy is poor.

In this paper, we apply the weighted noise spectral estimation method [5],[6]. The noisy speech spectrum is weighted by the weight function $W(l,k)$, which is determined based on the posterior SNR $\hat{\gamma}(l,k)$ as shown in Fig.2. The noisy speech spectrum is suppressed in the high SNR region in order to suppress over estimation of the noise spectrum. The weighted spectrum is expressed by

$$z(l,k) = W(j,k)|X(l,k)|^2 \quad (28)$$

The noise spectrum is estimated by averaging $z(l,k)$ over

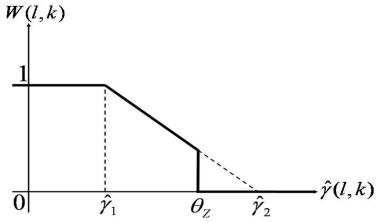

Fig. 2. Weight function $W(l,k)$.

several frames. Figure 2 means that in the begining frames and in the low SNR region, that is $\hat{\gamma}(l,k) < \theta_z$, $z(l,k)$ is included as the noise components. On the other hand, after the beginning frames and in the high SNR region, that is $\hat{\gamma}(l,k) > \theta_z$, $z(l,k)$ is not included as the noise components, rather the previous average of $z(l,k)$ is used.

V. Simulation and Discussions

A. Evaluation Measures

1) Normalized Estimation Error:

$$\varepsilon(l) = 10\log_{10}\left(\frac{\sum_{k=0}^{M}\left||N(l,k)|^2 - |\bar{N}(l,k)|^2\right|}{\sum_{k=0}^{M}|N(l,k)|^2}\right) \quad (29)$$

$$\bar{\varepsilon} = \frac{1}{L}\sum_{l=1}^{L}\varepsilon(l) \quad (30)$$

L is the number of all frames. The smaller value of ε means the higher acculate estimation.

2) Segmental SNR: SNR at the input and the output is evaluated by the following segmental SNR.

$$SNR_{seg} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{n=N_l}^{N_l+N-1}s^2(n)}{\sum_{n=N_l}^{N_l+N-1}(s(n)-\hat{s}(n))^2} \quad (31)$$

N is a length of the interval, where the segmental SNR is evaluated. The actual length is 12ms.

3) Ideal Estimation: In order to evaluate accuracy of the proposed method, we employ the ideal estimation by using the true noise spectrum. Let $G_{tl}(l,k)$ be the spectral gain obtained by using the true noise spectrum. The ideal noise suppressed output signal is obtained by

$$\hat{s}(n) = IFFT[G_{tl}(l,k)X(l,k)] \quad (32)$$

B. Simulation and Discussions

1) Effects of Adaptive Normalization Parameter in VAD: The adaptive normalization parameter $C_{new}(l)$ in the VAD proposed in Sec.IV-A is evaluated here. Figure 3 shows the normalized noise spectral estimation error in average $\bar{\varepsilon}$, with respect to the C value, which is changed from 0 to 50. Furthermore, many kinds of noises are used, including white, babble, car, pink, factory and tank. In the first interval from 0 to 10,000 samples, the white noise is used, and in the second interval, from 10,001 to 30,000 samples, the white and the non-white noises are used. The input SNR_{seg} is 3dB. As shown in this figure, the optimum value of C, which

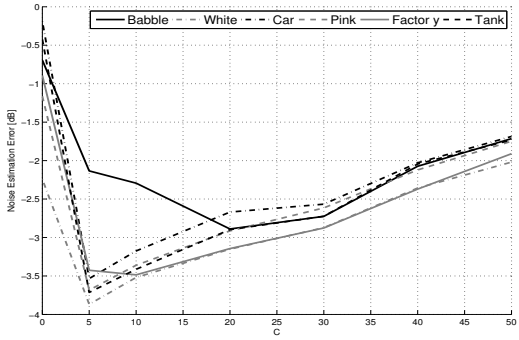

Fig. 3. Normalized noise spectral estimation error in average $\bar{\varepsilon}$, with respect to C.

gives the minimum estimation error, is different for the noise property. $C_{new}(l)$ can provide small estimation errors, that is -2.62, -3.69, -3.60, -3.74, -3.58, -3.79 dB for the noises in the second interval, that is babble, white, car, pink, factory, tank, respectively. These results are close to the minimum values.

2) Normalized Estimation Error and Segmental SNR: The babble noise and the white noise are used in the first interval from 0 to 10,000 samples and the second interval, from 10,001 to 30,000 samples, respectively. The input segmental SNR is 6dB and -1.4dB in the first and the second intervals, respectively. The normalized estimation error $\varepsilon(l)$ is shown in Fig.4.

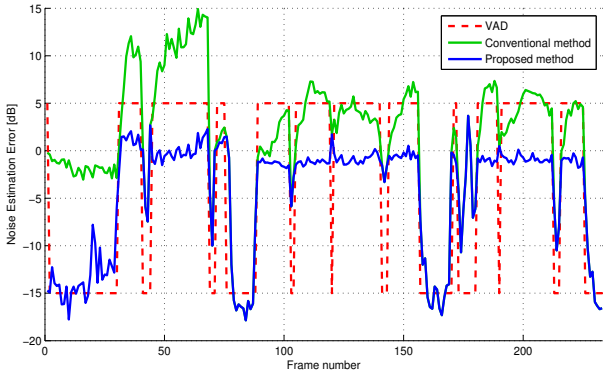

Fig. 4. Normalized noise spectral estimation error $\varepsilon(l)$

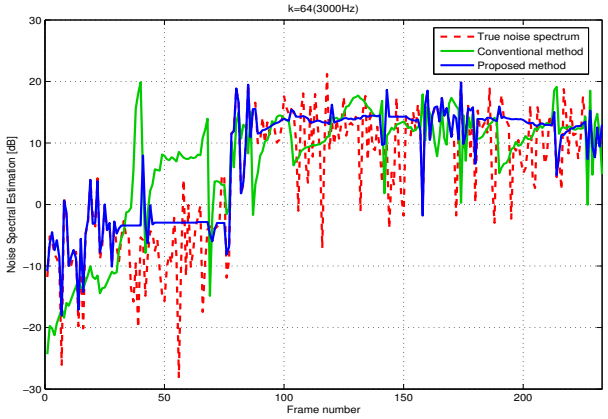

Fig. 5. Noise spectral estimation for 3kHz component.

A red dashed line indicates the output of the VAD. High and low levels mean the speech and the non-speech frames, respectively. The green line and the blue line indicate the errors of the conventional method and the proposed method, respectively. The error can be well reduced from the conventional method. Furthermore, the noise spectrum is accurately estimated in the non-speech frames.

Estimation of the time varying noise spectrum for the 3kHz component are shown in Fig.5. The horizontal axis is the frame number. The red dotted line means the true noise spectrum, the green line is that of the conventional method and the blue line is that of the proposed method. At around 80th frame, the noise is changed from the babble noise to the white noise. The proposed method can estimate the noise spectrum in both the babble and the white noise environments. Furthermore, the proposed method can quickly adapt the change of the noise.

The output segmental SNR and the normalized estimation error for the different input segmental SNR are listed in Tables I and II. Compared to the conventional rapid adaptation method, the normalized noise spectral estimation error is improved by $3.2 \sim 4.7$dB, and the output segmental SNR is improved by $2.0 \sim 3.8$dB. From these results, the proposed method can accurately estimate the noise spectra in a variety of conditions.

TABLE I

SNR_{seg}[dB] OF OUTPUT SIGNAL $\hat{s}(n)$.

Input SNR_{seg}[dB]	0	3	6	9
MMSE STSA(Ideal)	10.58	12.34	14.32	16.45
MMSE STSA(Conven)	3.147	5.254	6.859	8.057
MMSE STSA(Proposed)	5.314	7.475	9.642	11.89
Joint MAP(Ideal)	10.58	12.36	14.33	16.42
Joint MAP(Conven)	3.180	5.612	7.450	8.904
Joint MAP(Proposed)	5.388	7.595	9.972	12.30

TABLE II

NORMALIZED ESTIMATION ERROR IN AVERAGE $\bar{\varepsilon}$.

Input SNR_{seg}[dB]	0	3	6	9
Conventional	-0.4373	-0.6667	0.4369	2.152
Proposed	-4.517	-3.864	-3.187	-2.521

VI. CONCLUSIONS

In this paper, a new noise spectral estimation method is proposed. The speech and the non-speech frames are discriminated by the VAD, in which a new adaptive normalization parameter and a adaptive threshold are proposed. A recursive noise estimation method is applied, in which a new adaptive smoothing parameter is proposed. Through simulations by using many kinds of the noises, the proposed method can accurately estimate the spectra of the many kinds of the noises, in both the speech and non-speech frames. Furthermore, it can quickly adapt to change of the noises.

REFERENCES

[1] "Minimum performance requirements for noise suppressor application to the AMR speech encode", 3GPP TS 06.77 V8.1.1, April 2001.
[2] Y.Ephraim and D.Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator", IEEE Trans. vol.ASSP-32, no.6, pp.1109-1121, Dec. 1984.
[3] Y.Ephraim and D.Malah, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator", IEEE Trans. vol.ASSP-33, no.2, pp.443-445, April 1985.
[4] T.Lotter and P.Vary, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-gaussian speech modeling", Proc. EUSIPCO-04, pp.1447-60, Sep. 2004.
[5] M.Katou, A.Sugiyama and M.Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA", IEICE (Japan) Trans. Fundamental, vol.E85-A, no.7, pp.1710-1718, July 2002.
[6] K.Nakayama, H.Suzuki and A.Hirano, "Improved methods for noise spectral estimation and adaptive spectral gain control in noise spectral suppressor", Proc. ISPACS'07, Xiamen, China, pp.97-100, Dec. 2007.
[7] J.Sohn and N.Kim, "Statistical model-based voice activity detection", IEEE signal Processing Letter, vol.6, no.1, pp.1-3, 1999.
[8] C.Jia and B.Xu, "An improved entropy-based endpoint detection algorithm", Proc. Int. Sympo. Chinese Spoken Language Processing, pp.1399-1402, Aug. 2002.
[9] I.Cohen and B.Berdugo, "Speech enhancement for non-stationary noise environments", Signal Processing, vol.81, no.11, pp.2403-2418, Nov. 2001.
[10] I.Cohen and B.Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", IEEE Signal Processing, Lett. vol.9, no.1, pp.12-15, 2002.
[11] B.F.Wu and K.C.Wang, "Noise spectrum estimation with entropy-based VAD in non-stationary environments", IEICE (Japan) Trans. Fundamentals, vol.E89-A, no2, pp.479-485, Feb. 2006.
[12] S.Rangachari, P.Loizou and Y.Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments", Proc. IEEE ICASSP'04, pp.305-308, 2004.