Improved methods for noise spectral estimation and adaptive spectral gain control in noise spectral suppressor

doi: 10.1109/ISPACS.2007.4445812

Improved Methods for Noise Spectral Estimation and Adaptive Spectral Gain Control in Noise Spectral Suppressor

Kenji Nakayama Hirokazu Suzuki Akihiro Hirano
Graduate School of Natural Science and Technology, Kanazawa Univ.
Kakuma-machi, Kanazawa, 920-1192, Japan
Tel: +81-76-234-4896, Fax:+81-76-234-4900
E-mail:nakayama@t.kanazawa-u.ac.jp

***Abstract*— In this paper, new approaches to noise spectrum estimation and spectral gain control are proposed for noise spectral suppressors. First, the speech absent frames are detected by using spectral entropy. In the speech absent frames, a weighting factor used in estimating the noise spectrum is modified so as to emphasize effect of the noisy speech signal. Next, a spectral gain is more reduced by multiplying a factor in order to suppress effects of the noise in the speech absent frames. Furthermore, in the speech present frames, in order to reduce signal distortion, the spectral gain is controlled to be unity based on an SNR calculated by using a ridgeline spectrum. Finally, the original noisy speech is added to the estimated speech in some ratio. This ratio is controlled by the long term averaged SNR of the estimated noise and the noisy speech. Computer simulations by using speech signals, the white noise, the car noise and the bubble noise, which are available in public, have been carried out for the conventional methods and the proposed method. The proposed method can improve a segmental SNR and speech quality compared to the conventional methods. Especially, it is useful for the bubble noise.**

I. Introduction

Mobile communication systems are used in a variety of environments, under which communication quality is affected by many kinds of noises. ETSI recommended the minimum performance requirements for noise suppresser application to the AMR speech encoder [1].

A spectral suppression technique is a hopeful approach, and many kinds of noise spectral suppressors have been developed for mobile phones. Several methods for estimating a spectral gain, which is used to suppress the noise spectrum, have been proposed. They include MMSE STSA [2], MMSE LSA [3] and Joint MAP [11]. Performance of the noise suppressor based on the spectral suppression technique is highly dependent on accuracy of the noise spectral estimation. A weighted noise spectral estimation method, which can follow the noise spectrum change, has been proposed [8],[9]. This approach satisfies the requirements recommended by ETSI [10]. In this method, however, the noise spectrum estimation is not adequate. Especially, the estimation errors are relatively large for bubble noises, whose spectrum may dynamically changes.

In this paper, the speech absent and present frames are precisely detected by using a voice activity detector [12]. In the speech absent frames, the noise spectrum is more accurately estimated. Furthermore, the spectral gain is adaptively controlled depending on whether the speech absent or present frames. Computer simulations by using speech signal, the white noise, the car noise and the bubble noise will be shown in order to confirm usefulness of the proposed methods.

II. Spectral Suppression Technique

Let $x(n)$, $d(n)$ and $y(n)$ be a clean speech, noise and their mixed signal, that is noisy speech, respectively.

$$y(n) = x(n) + d(n) \tag{1}$$

Let $X(m,k)$, $D(m,k)$ and $Y(m,k)$ be the FFT of $x(n)$, $d(n)$ and $y(n)$, respectively. m is a frame number, and k is a frequency number. They are related by

$$\begin{aligned} Y(m,k) &= X(m,k) + D(m,k) && (2)\\ Y(m,k) &= R(m,k)e^{j\theta(m,k)} && (3)\\ X(m,k) &= A(m,k)e^{j\alpha(m,k)} && (4) \end{aligned}$$

A prior SNR $\xi(m,k)$, which is a ratio of the clean speech spectrum and the noise spectrum, and a posterior SNR $\gamma(m,k)$, which is a ratio of the noisy speech spectrum and the noise spectrum, are expressed by

$$\begin{aligned} \xi(m,k) &= \frac{\lambda_x(m,k)}{\lambda_d(m,k)} && (5)\\ \gamma(m,k) &= \frac{|Y(m,k)|^2}{\lambda_d(m,k)} && (6)\\ \lambda_x(m,k) &= E\{\left|X(m,k)\right|^2\} && (7)\\ \lambda_d(m,k) &= E\{\left|D(m,k)\right|^2\} && (8) \end{aligned}$$

In the above equations, only the noisy speech spectrum $|Y(m,k)|^2$ is available, and the other spectra should be estimated. The prior SNR $\xi(m,k)$ can be estimated by [2].

$$\begin{aligned} \hat{\xi}(m,k) &= \alpha_{SNR}\gamma(m-1,k)G^2(m-1,k)\\ &+ (1-\alpha_{SNR})P[\gamma(m,k)-1] && (9)\\ P[x] &= \begin{cases} x, & x>0 \\ 0, & otherwise \end{cases} && (10) \end{aligned}$$

1-4244-1447-4/07/$25.00 ©2007 IEEE

α_{SNR}, satisfying $0 < \alpha_{SNR} < 1$, controls a tradeoff between noise suppression and signal distortion reduction. $G(m-1,k)$ is a spectral gain function at the previous frame $m-1$. In order to suppress musical noise and make the residual noise to be natural, $\hat{\xi}(m,k)$ should be bounded by [4],

$$\hat{\xi}(m,k) = \begin{cases} \hat{\xi}(m,k), & \hat{\xi}(m,k) > \xi_{MIN} \\ \xi_{MIN}, & otherwise \end{cases} \quad (11)$$

The posterior SNR $\gamma(m,k)$ can be calculated by using the estimated noise spectrum. By using these estimated prior and posterior SNRs, the spectral gain $G(m,k)$ at the mth frame is calculated, and the noise spectrum is suppressed by

$$\hat{X}(m,k) = G(m,k)Y(m,k) \quad (12)$$

$G(m,k)$ is calculated by MMSE STSA [2], MMSE LSA [3] and Joint MAP [11].

III. Noise Spectrum Estimation

Several kinds of estimation methods for the noise spectra have been proposed. A most simple method is to estimate the noise spectrum in the beginning several frames ($m \le T_0$), where the speech is assumed to be absent [2].

$$\lambda_d(m,k) = \begin{cases} |Y(m,k)|^2 & m \le T_0 \\ \frac{1}{T_0}\sum_{m=1}^{T_0}|Y(m,k)|^2 & T_0 < m \end{cases} \quad (13)$$

The second method is based on the voice activity detection (VAD) technique. The noise is assumed to be stationary compared to the speech. The noise spectrum is updated in the speech absent frames [12].

$$\lambda_d(m,k) = \begin{cases} \alpha\lambda_d(m-1,k) + (1-\alpha)|Y(m,k)|^2 \\ \quad \text{Speech absent frame} \\ \lambda_d(m-1,k) \\ \quad \text{Speech present frame} \end{cases} \quad (14)$$

α is a scaling factor, satisfying $0 < \alpha < 1$.

The third method is the weighted noise estimation method [8],[9]. The noise spectrum is continuously estimated by using the noisy speech, which is weighted following the estimated posterior SNR, that is $\hat{\gamma}(m,k)$. It is possible to avoid over estimation and high tracking performance for nonstationary noise. The weighted noise estimation consists of the posterior SNR estimation, calculating the weighting function $W(m,k)$ and the averaging.

First, $\gamma(m,k)$ is estimated by using $|Y(m,k)|^2$ and $\lambda_d(m-1,k)$ of the previous $(m-1)th$ frame as follows:

$$\hat{\gamma}(m,k) = \frac{|Y(m,k)|^2}{\lambda_d(m-1,k)} \quad (15)$$

Next, based on $\hat{\gamma}(m,k)$, the weighting factor $W(m,k)$ is calculated following Fig.1, where over estimation can be avoided for relatively high SNR.

IV. An Improved Noise Spectrum Estimation Method

In the proposed method, the noise spectrum is more precisely estimated in the speech absent frame. For this purpose, a voice activity detector (VAD) is applied. Especially, in order to estimate the noise spectrum in the nonstationary environment, the VAD using the spectral entropy [12] is employed.

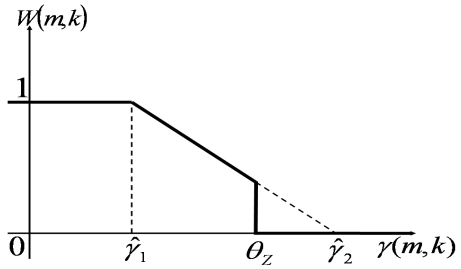

Fig. 1. Weight function $W(m,k)$ related to $\hat{\gamma}(m,k)$.

A. Voice Activity Detector

The specral entropy is given by

$$Y_{energy}(m,k) = |Y(m,k)|^2 \quad (16)$$

$$P_r(m,k) = \frac{\left(Y_{energy}(m,k)+C\right)}{\sum_{k=1}^{2M}\left(Y_{energy}(m,k)+C\right)} \quad (17)$$

$$H(m) = -\sum_{k=1}^{2M} P_r(m,k)\log\left(P_r(m,k)\right) \quad (18)$$

$2M$ is the number of frequency points. If $H(m)$ is higher than the threshold, then this frame is classified into the speech absent frame, and if $H(m)$ is lower than the threshold, then this frame is regarded as the speech present frame.

B. Noise Spectrum Estimation

In the speech absent frame, the noise spectrum is estimated by using a new weight function shown in Fig.2. This weight

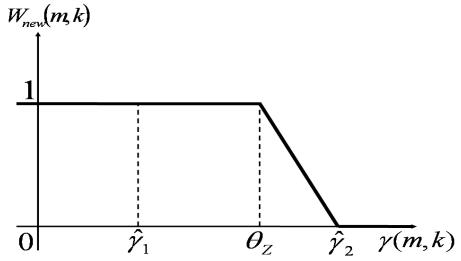

Fig. 2. A new weight function used to estimate noise spectrum.

function is modified from the conventional weight function [8],[9] in order to put a stress on the noisy speech.

In order to avoid over estimation of the noise spectrum, when a long term SNR is enough high, the conventional weigh function shown in Fig.1 is used. The long term SNR is evaluated as follows:

$$\bar{\lambda}_y(m) = \alpha_{LT}\bar{\lambda}_y(m-1) + (1-\alpha_{LT})\frac{1}{2M}\sum_{k=1}^{k=2M}|Y(m,k)|^2 \quad (19)$$

α_{LT} is a forgetting factor. Equation (19) is carried out in the speech present frames. By using $\bar{\lambda}_y(m)$ and $\lambda_d(m,k)$, the long term SNR is calculated by

$$SNR_{LT}(m) = \frac{2M\bar{\lambda}_y(m)}{\sum_{k=1}^{2M}\lambda_d(m,k)} - 1 \quad (20)$$

V. A New Adaptive Control Method for Spectral Gain

In the spectral suppression technique, there exits a trade-off between the large residual noise due to a lack of noise suppression and speech distortion caused by over suppression. In this paper, an adaptive control method for the spectral gain is proposed.

A. Modification of Spectral Gain

In the speech absent frames, the spectral gain is more reduced in order to well suppress the noise spectrum.

$$G(m,k) = \begin{cases} G_{sup}G(m,k) & \text{in speech absent frames} \\ G(m,k) & \text{in speech present frames} \end{cases} \quad (21)$$

Next, in order to reduce speech distortion due to over suppression, the minimum values G_{floor} and G_{min} in the speech present frames are set to be lager than those in the speech absent frames.

B. Adaptive Control of Spectral Gain Based on Speech Spectrum Estimation

A ridgeline spectrum and an offset-SNR are estimated as follows [14]:

1) The peak amplitude $|Y_{max}(k)|$ is updated by using the noisy speech spectrum $|Y(m,k)|$ as follows:
   If $|Y_{max}(k)| < |Y(m,k)|$, then $|Y_{max}(k)| = |Y(m,k)|$
2) When the noisy speech spectrum $|Y(m,k)|$ is close to the peak amplitude $|Y_{max(k)}|$, the ridgeline spectrum $|Y_{edge}(m,k)|$ is updated as follows:
   If $|Y(m,k)| > \beta_{max}|Y_{MAX}(k)|$, then $|Y_{edge}(m,k)| = \mu_r|Y_{edge}(m-1,k)| + (1-\mu_r)|Y(m,k)|$
3) An offset-SNR(SNR_{offset}) is calculated based on a ratio of the ridgeline spectrum and the estimated noise spectrum, which are averaged in the frequency domain. $SNR_{offset} = \frac{\sum_{k=1}^{2M}(1-\alpha_{offset})|Y_{edge}(m,k)|}{\sum_{k=1}^{2M}\sqrt{\lambda_d(m,k)}}$

β_{max}, μ_r and α_{offset} are positive constants less than 1.

SNR_{offset} is the posterior SNR by which the speech distortion does not occur. In the proposed method, SNR_{offset} is used as follows: In the speech present frames, which are detected by the VAD, if the posterior SNR is higher than SNR_{offset}, then the spectral gain is set to be $G(m,k) = 1$.

C. Adding Original Noisy Speech

In the spectral suppression technique, the noise spectrum can be suppressed, however, at the same time, the speech itself is usually distorted. By adding the original noisy speech to the estimated speech, the speech becomes more natural at the expense of the remaining noise.

In this paper, a rate of adding the original noisy speech is controlled by the long term average $\gamma_{av}(m,k)$ of the posterior SNR as follows:

$$\gamma_{av}(m,k) = \alpha_{av}\gamma(m,k) + (1-\alpha_{av})\gamma(m-1,k) \quad (22)$$

$$\tilde{G}(m,k) = \begin{cases} m_1 + (1-m_1)G(m,k), \\ \quad SNR_{th1} < \gamma_{av}(m,k) \\ m_2 + (1-m_2)G(m,k), \\ \quad SNR_{th2} < \gamma_{av}(m,k) \leq SNR_{th1} \\ G(m,k), \\ \quad otherwise \end{cases} \quad (23)$$

$$\hat{x}(m,n) = IFFT\big[\tilde{G}(m,k)|Y(m,k)|e^{j\theta(m,k)}\big] \quad (24)$$

$\hat{x}(m,n)$ is the estimated speech in the mth frame. The parameters are determined by experience as follows: $m_1 = 0.75$, $m_2 = 0.15$, $SNR_{th1} = 12$dB and $SNR_{th2} = 5$dB.

VI. Simulations and Discussions

A. Speech and Noise Data

The speech of Japanese sentence of 30,000 samples with a sampling frequency of 10kHz is used. White noise, car noise and bubble noise are used, which are available in public [13]. The number of FFT samples is 256, and the Hamming window is used to extract a frame.

B. Ideal SNR

In order to evaluate performance of the proposed method, an ideal SNR is introduced here. The true noise spectrum $|D(m,k)|$ is used to calculate the spectral gain $G(m,k)$. Letting this ideal spectral gain be $G_{tn}(m,k)$, the estimated speech is given by

$$\hat{x}_{ideal}(m,n) = IFFT\big[G_{tn}(m,k)|Y(m,k)|e^{j\theta(m,k)}\big] \quad (25)$$

C. Performance Evaluation

The segmental SNR is used. The signal is divided into an each 12 msec segment. SNR is calculated in each segment, and are averaged over a long term. The segmental SNR for the input and the output, that is before and after noise suppression, are given by

$$\text{Input:}SNR_{seg} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{n=Nl}^{Nl+N-1}x^2(n)}{\sum_{n=Nl}^{Nl+N-1}d^2(n)} \quad (26)$$

$$\text{Output:}SNR_{seg} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{n=Nl}^{Nl+N-1}x^2(n)}{\sum_{n=Nl}^{Nl+N-1}(\hat{x}(n)-x(n))^2} \quad (27)$$

N is the number of samples in each segment, L is the number of the segments. The lower and upper bounds are set to be -35dB and 35dB, respectively.

Taking quality of speech into account, the log-spectral distortion (LSD) is also employed [5].

$$LSD = \frac{1}{L}\sum_{m=1}^{L}\left(\frac{1}{2M}\sum_{k=1}^{2M}\Big(\log\frac{|X(m,k)|+\delta}{|\hat{X}(m,k)|+\delta}\Big)^2\right)^{\frac{1}{2}} \quad (28)$$

δ is a very small value, $2M$ is the frame length. SNR_{seg} compares the time domain waveforms, which includes both amplitude and phase components. On the other hand, LSD uses only an amplitude response, which is important for hearing by human ears.

D. Spectral Gain Calculation Methods

The following three kinds of methods, including MMSE STSA, OM-LSA and Joint MAP are used for calculating the spectral gain. SNR_{seg} and LSD are shown in Table 1 through Table 4. SNR before the noise suppression, are set to be 0, 6 and 12 dB. The higher SNR_{seg} and the lower LSD mean good noise suppression.

TABLE I

SNR_{seg} IN DB FOR WHITE NOISE.

Input SNR_{seg}[dB]	0	6	12
Ideal			
MMSE STSA	8.03	5.66	3.54
OM-LSA	7.94	5.96	3.44
Joint MAP	8.09	5.87	3.96
Conventional Methods			
MMSE STSA	6.37	4.54	2.41
OM-LSA	6.24	4.49	2.41
Joint MAP	6.24	4.75	3.05
Proposed Methods			
MMSE STSA	7.53	5.17	3.24
OM-LSA	7.64	5.22	3.31
Joint MAP	7.46	5.12	3.29

TABLE II

LSD FOR WHITE NOISE.

Input SNR_{seg}[dB]	0	6	12
Ideal			
MMSE STSA	1.46	1.37	1.30
OM-LSA	1.45	1.35	1.27
Joint MAP	1.41	1.28	1.17
Conventional Methods			
MMSE STSA	1.57	1.32	1.15
OM-LSA	1.58	1.33	1.14
Joint MAP	1.58	1.30	1.11
Proposed Methods			
MMSE STSA	1.44	1.27	1.13
OM-LSA	1.43	1.29	1.12
Joint MAP	1.44	1.26	1.12

TABLE III

SNR_{seg} IN DB FOR BUBBLE NOISE.

Input SNR_{seg}[dB]	0	6	12
Ideal			
MMSE STSA	5.83	3.90	2.14
OM-LSA	5.69	3.86	2.11
Joint MAP	5.82	4.14	2.70
Conventional Methods			
MMSE STSA	2.88	2.07	0.68
OM-LSA	2.91	2.12	0.73
Joint MAP	2.95	2.33	1.37
Proposed Methods			
MMSE STSA	4.92	3.38	1.75
OM-LSA	5.28	3.59	1.94
Joint MAP	5.06	3.56	1.96

The proposed method can provide good performance of noise suppression, which are close to the ideal results. Especially, for the bubble noise, SNR_{seg} can be improved by about 2 dB. Simulation results for the car noise are omitted here due to page limitation. The improvements in SNR_{seg} and LSD are similar to those of the white noise case.

TABLE IV

LSD FOR BUBBLE NOISE.

Input SNR_{seg}[dB]	0	6	12
Ideal			
MMSE STSA	1.25	1.14	1.00
OM-LSA	1.22	1.12	1.00
Joint MAP	1.13	0.99	0.84
Conventional Methods			
MMSE STSA	1.23	1.07	0.93
OM-LSA	1.22	1.05	0.93
Joint MAP	1.20	1.00	0.85
Proposed Methods			
MMSE STSA	1.19	1.10	0.97
OM-LSA	1.19	1.11	0.93
Joint MAP	1.17	1.06	0.97

VII. CONCLUSIONS

Several improved techniques are proposed for the noise spectrum estimation and the adaptive spectral gain control in the noise spectral suppressor. The segmental SNR and the log-spectral distortion (LSD) are evaluated by using the speech and several noises. The proposed method is superior to the conventional methods in all noise environments. Especially, SNR_{seg} can be drastically improved for the bubble noise.

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] O.Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppresor", IEEE Trans. Speech Audio Process., vol.2, no.2, pp.345-349, April 1994.

[5] D.L.Wang and J.S.Lim, "The unimportance of phase in speech enhancement", IEEE Trans. Acoust. Speech and Signal Processing, vol.ASSP-30, no.4, pp.679-681, Aug. 1982.

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

[7] I.Cohen, "Optimal speech enhancement under signal presence unsertainty using log-spectral amplitude estimator", IEEE Signal Processing Letters, vol.9, no.4, pp.113-116, 2002. robust speech recognition", Proc. IEEE ICASSP-95, Detroit, pp.153-156, May 1995.

[8] A.Sugiyama, T.P.Hua and M.Kato, "Noise suppression with synthesis windowing and pseudo noise injection", Proc. ICASSP'02, pp.545-548, May 2002.

[9] M.Katou, A.Sugiyama and M.Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA", IEICE Trans. Fundamental, vol.E85-A, no.7, pp.1710-1718, July 2002.

[10] "Test results of NEC AMR-NS solution based on TS 26.077", 3GPP Tdoc S4-020415, July 2002.

[11] 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.

[12] 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, Feb. 2006.

[13] A.Varga and H.J.M.Steeneken,"Assessment for automatic speech recognition: II.NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", Speech Commun., vol.12, no.3, pp.247-251, July 1993.

[14] T.Otani, K.Suzuki, Y.Ota, S.Sassaki and F.Amano,"A noise suppresser using speech spectrum estimations (in Japanese)", IEICE, 21th Signal Processing Symposium, Kyoto, Japan, C8-3, Nov. 2006.