NOISE SUPPRESSION WITH HIGH SPEECH QUALITY BASED ON WEIGHTED NOISE ESTIMATION AND MMSE STSA

Masanori Kato, Akihiko Sugiyama and Masahiro Serizawa

Multimedia Research Laboratories, NEC Corporation
1-1, Miyazaki 4-chome, Miyamae-ku, KAWASAKI 216-8555
m-kato@df.jp.nec.com

ABSTRACT

A noise suppression algorithm with high speech quality based on weighted noise estimation and MMSE STSA is proposed. The proposed algorithm continuously updates the noise estimate by noisy speech weighted in accordance with an estimated SNR. The spectral gain is modified with the estimated SNR so that it can better utilize the improvement in noise estimation. Subjective evaluation results show that five-grade mean opinion scores of the new algorithm are improved by as much as 0.93 and 0.35, compared with the original MMSE STSA and the EVRC noise suppression algorithm, respectively.

1. INTRODUCTION

Applications of speech coding and speech recognition have been exploding these days. Among these are cellular phones and car navigation systems, to name a few. One of the challenges in those applications is that they are often used in noisy environment. The speech quality is seriously degraded in noisy environment, resulting in uncomfortable communication or a lower recognition rate. This is because speech coding and speech recognition have been developed in noise-free environment. A remedy for this problem is a noise suppressor.

A variety of noise suppression algorithms can be found in the literature [1]. Most of the noise suppression algorithms are based on STSA (short time spectral amplitude) analysis. STSA is most widely used for its computational advantage. Among others, MMSE (minimum mean square error) STSA proposed by Ephraim and Malah [2] is the most popular STSA based algorithm. It minimizes the mean squared error of the estimated short time spectral amplitude. It is reported that MMSE STSA can provide good noise suppression without unpleasant residual noise called "musical noise" [3, 4].

Another STSA-based popular algorithm is the one employed for EVRC (Enhanced Variable Rate Codec) [5] which is the North American CDMA digital cellular phone standard. This is the most successful algorithm whose quality has been proven to be good through commercial products. Nevertheless, the quality may not be sufficiently good for a wide range of SNRs which were not given much attention when it was standardized.

This paper proposes a noise suppression algorithm with good speech quality for a wide range of SNRs. The proposed algorithm continuously estimates the noise with a noisy speech weighted by an estimated SNR. This makes more accurate SNR estimate available for gain calculation, resulting in good speech quality and sufficient noise suppression simultaneously. The spectral gain is modified so that the improved noise estimation can be utilized more effectively.

2. CONVENTIONAL ALGORITHM

2.1. MMSE STSA[2]

Figure 1 shows the structure of the original MMSE STSA. It mainly consists of six functions; short-time Fourier analysis, noise estimation, *a posteriori* and *a priori* SNR estimation, spectral gain calculation and short-time synthesis.

Assuming that the clean speech $s(t)$ is degraded by an additive noise $d(t)$, the noisy speech $x(t)$ is given by

$$x(t) = s(t) + d(t), \tag{1}$$

where t is a time index. In short-time Fourier analysis, the noisy speech $x(t)$ is first segmented into frames of M samples. An analysis window with a 50 % overlap is applied to the segmented noisy speech $x_n(t)$ in frame n. The discrete Fourier transform of the windowed noisy speech is computed to output its spectral amplitude $|X_n(k)|$ and phase $\angle X_n(k)$, where n and k refer to the analysis frame and the frequency bin index. Noise suppression is applied only to the spectral amplitude of the noisy speech in each frequency bin. The amplitude $|X_n(k)|$ of the noisy speech is multiplied by a spectral gain $G_n(k)$ to obtain the amplitude $|Y_n(k)|$ of the enhanced speech. In short-time synthesis, the enhanced speech spectrum $Y_n(k)$ is first constructed with $\angle X_n(k)$ and $|Y_n(k)|$. After the inverse discrete Fourier transform of $Y_n(k)$ is calculated, the enhanced speech $y_n(t)$ is obtained by performing the overlap-add processing.

The spectral gain is calculated with an estimated *a priori* SNR $\hat{\xi}_n(k)$ and *a posteriori* SNR $\hat{\gamma}_n(k)$. Their

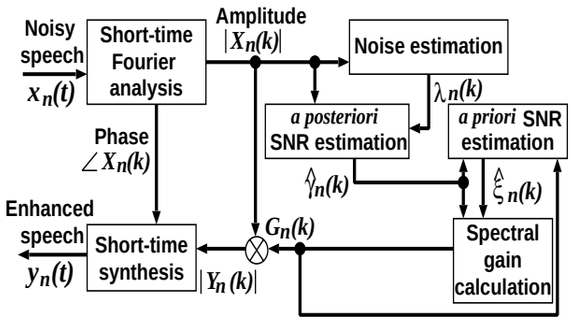

Figure 1: Structure of the conventional noise suppression (MMSE STSA).

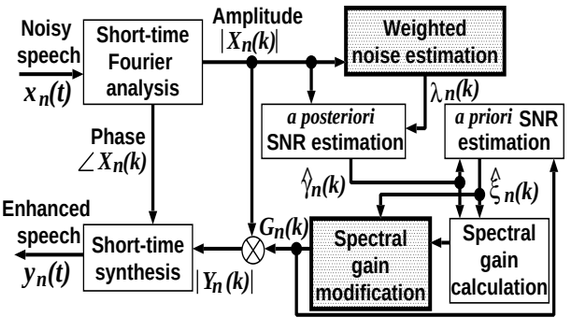

Figure 2: Structure of the proposed noise suppression.

estimates are obtained based on the estimated noise power spectrum $\lambda_n(k)$ which is calculated from the spectral amplitude of the noisy speech in the first nonspeech period.

2.2. Problem in Noise Estimation

The original MMSE STSA estimates the noise power spectrum based on the noisy speech only in the first nonspeech period where the pure noise is available. This means that, for a nonstationary noise, a change in noise characteristics cannot be tracked and the enhanced speech quality becomes poor.

As a continuous noise estimation which has the tracking capability, a noise estimation method based on minimum statistics [6] is widely used. The minimum value of the smoothed noisy speech power within a finite time-window length L_{MS} is used as the estimated noise. Because of the statistical nature, a larger L_{MS} provides more accurate noise estimation for a stationary noise. However, the tracking capability for a nonstationary noise is degraded. A short window, on the other hand, may introduce overestimation which results in poor speech quality for high SNRs, although it achieves better tracking capability. As a result, there is a tradeoff in the selection of L_{MS}. Therefore, it is not easy to select an appropriate window length for good tracking capability without overestimation.

3. PROPOSED ALGORITHM

To achieve good tracking capability without overestimation for various nonstationary noise sources, the proposed noise suppression algorithm employs a noise estimation with a weighting factor based on the estimated SNR. The weighting factor makes continuous noise estimation possible without overestimation even in speech periods. As a result, the weighted noise estimation tracks the change of the noise characteristics in both speech and nonspeech periods. To obtain a suitable spectral gain for the new noise estimation, the spectral gain is modified in accordance with the SNR. Figure 2 shows the structure of the proposed noise suppression.

3.1. Weighted noise estimation

The weighted noise estimation mainly consists of three steps; SNR estimation, weighting factor calculation and averaging. The noisy speech is weighted by a weighting factor calculated based on the estimated SNR. The estimated noise is obtained as an average of the weighted noisy speech.

In the first step, the estimated SNR $\tilde{\gamma}_n(k)$ is obtained from the power spectrum of the noisy speech $|X_n(k)|^2$ and the estimated noise $\lambda_{n-1}(k)$ as follows.

$$\tilde{\gamma}_n(k) = \frac{|X_n(k)|^2}{\lambda_{n-1}(k)} \tag{2}$$

In the second step, the weighting factor $W_n(k)$ is calculated by

$$W_n(k) = \begin{cases} 1, & \tilde{\gamma}_n(k) < \tilde{\gamma}_1 \\ \frac{\tilde{\gamma}_n(k)-\tilde{\gamma}_2}{\tilde{\gamma}_1-\tilde{\gamma}_2}, & \tilde{\gamma}_1 \le \tilde{\gamma}_n(k) \le \tilde{\gamma}_2 \\ 0, & \tilde{\gamma}_2 < \tilde{\gamma}_n(k) \end{cases}, \tag{3}$$

where $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are constants. This nonlinear function is designed such that the weighting factor is almost inversely proportional to the estimated SNR. Overestimation for high SNRs does not happen by appropriately suppressing the contribution of the noisy speech to the estimate.

The weighted noisy speech $z_n(k)$ and its average $\lambda_n(k)$, which is used as the estimated noise, are given by

$$z_n(k) = W_n(k)|X_n(k)|^2, \tag{4}$$

$$\lambda_n(k) = \frac{trace\{\boldsymbol{Z}_n(k)\}}{\Psi(\boldsymbol{Z}_n(k))}, \tag{5}$$

where

$$\boldsymbol{Z}_n(k) = \begin{cases} [z_n(k), \tilde{\boldsymbol{Z}}_{n-1}(k)], & n \le T_{init} \\ [z_n(k), \tilde{\boldsymbol{Z}}_{n-1}(k)], & \tilde{\gamma}_n(k) < \theta_Z \\ \boldsymbol{Z}_{n-1}(k), & otherwise \end{cases}, \tag{6}$$

$$\tilde{\boldsymbol{Z}}_0(k) = \mathbf{0}_{1\times(L_Z-1)}, \tag{7}$$

$$\tilde{\boldsymbol{Z}}_n(k) = \boldsymbol{Z}_n(k)\,[\,\mathbf{I}_{L_Z-1}\ \mathbf{0}^T_{1\times(L_Z-1)}]^T. \tag{8}$$

Table 1: Parameters for proposed noise suppression.

Parameter	Value	Parameter	Value
M	128	T_{init}	4 frames
L_Z	20	θ_G	10 dB
$\tilde{\gamma}_1$	0 dB	G_{mod}	-1.0 dB
$\tilde{\gamma}_2$	10 dB	G_{floor}	-6.8 dB
θ_Z	7 dB		

$\Psi(\boldsymbol{Z}_n(k))$ is the number of non-zero elements in $\boldsymbol{Z}_n(k)$ and $trace\{\cdot\}$ is an operator to take the sum of its diagonal elements. L_Z and $\mathbf{I}_{L_Z-1}$ are the number of samples for the average and the identity matrix of size $L_Z - 1$, respectively. (6) means that $\boldsymbol{Z}_n(k)$ is updated only when the estimated SNR is lower than a threshold θ_Z, or the frame index is smaller than or equal to T_{init}. An inappropriate weighted noisy speech sample by an unreliable SNR estimate is eliminated by θ_Z to obtain a better value of $\lambda_n(k)$. $W_n(k) = 1$ for $0 < n \leq T_{init}$ under the assumption that the speech does not start in the first T_{init} frames. Noise estimation is performed independently for each bin, enabling more precise results depending on the SNR at each bin.

3.2. Spectral gain modification

The spectral gain is modified in two ways; conditional scaling and limitation. Conditional scaling further suppresses the residual noise for high SNRs resulting in clearer enhanced speech. The spectral gain $G_n(k)$ is multiplied by a scaling factor G_{mod} only when the estimated *a priori* SNR $\hat{\xi}_n(k)$ is smaller than a threshold θ_G. $G_{mod} < 1$ makes the value of $G_n(k)$ smaller to further suppress the noise for low SNRs.

The minimum value of the spectral gain is limited with G_{floor}. If $G_n(k)$ is smaller than G_{floor}, the original spectral gain $G_n(k)$ is replaced with G_{floor}. Therefore, excessive suppression, which causes speech distortion, can be avoided.

4. EVALUATION

The proposed noise suppression algorithm was compared with the conventional noise suppression algorithms in terms of noise estimation accuracy and subjective quality of the enhanced speech. Both speech and noise had been sampled at 8 kHz before they were digitally mixed to generate the noisy speech. Four kinds of background noise sources (babble, office, street and vehicle) were used. Hamming window was used as the analysis window. In MMSE STSA, 0.98 and 0.20 were used for the weighting factor α of decision-directed estimation and the probablity q of signal absence, respectively. Other parameters are shown in Table 1.

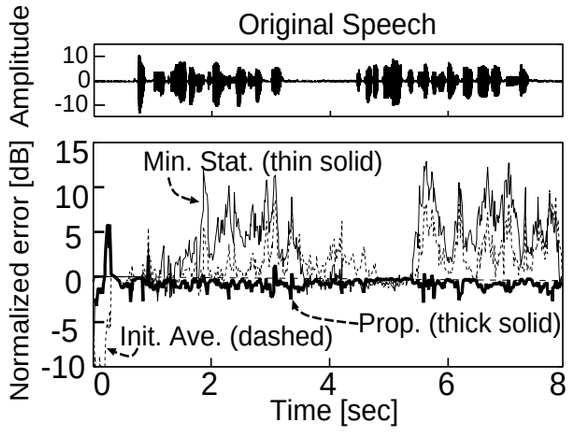

Figure 3: Original speech and Estimation error.

4.1. Objective evaluation for noise estimation

Noise estimation accuracy was evaluated frame by frame based on the normalized estimation error ε_n given by

$$\varepsilon_n = 10 \log_{10} \left(\frac{\sum_{k=0}^{M-1} \left| |D_n(k)|^2 - \lambda_n(k) \right|}{\sum_{k=0}^{M-1} |D_n(k)|^2} \right). \quad (9)$$

For the Minimum Statistics, $L_{MS} = 50$ is used as specified in [6]. The initial averaging, which is the original noise estimation used for the MMSE STSA, estimates the noise power spectrum in the initial 20 frames.

Figure 3 shows the normalized estimation error of the evaluated noise estimation methods for the 5-dB babble noise with the corresponding clean speech. The proposed noise estimation (Prop.) is clearly more accurate, compared with either the Minimum Statistics (Min. Stat.) or the initial averaging (Init. Ave.).

4.2. Subjective evaluation

A listening test was carried out for the proposed method, the original MMSE STSA combined with the Minimum Statistics and the EVRC noise suppression (EVRC/NS). A five-grade MOS (Mean Opinion Score) based on the absolute category rating [7] was used in the test. Twelve listeners evaluated the noise-suppressed speech which was encoded and decoded by the EVRC. Each listener scored between one and five, with five being the best. Six speech signals were used as the clean speech. The noise was added to the speech with different SNRs ($0, 5, 10$ and 15 dB) to produce the noisy speech.

Figure 4 shows the listening test result. Scores of the proposed method are higher than those of the original MMSE STSA and the EVRC noise suppression in most conditions. The difference between the scores of the proposed algorithm and those of the original MMSE STSA is statistically significant under 70 % of all tested conditions. The maximum difference was 0.93. When the proposed method is compared to the EVRC noise suppression, the difference of their scores is statistically

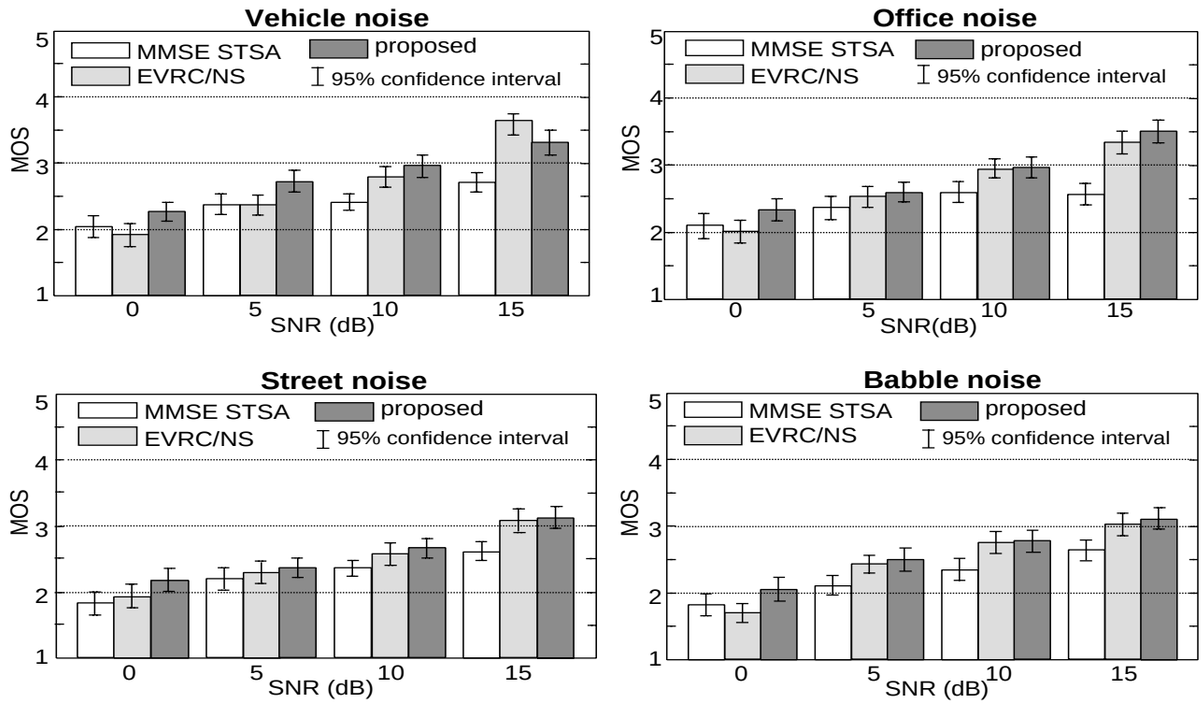

Figure 4: Results of listening test with codec.

significant under 25 % of all tested conditions. The maximum difference was 0.35. Similar results were also obtained in the test without codec [8].

5. CONCLUSION

A noise suppression algorithm based on weighted noise estimation and MMSE STSA has been proposed. The proposed algorithm continuously updates the noise estimate by noisy speech weighted in accordance with an estimated SNR. The spectral gain is modified with the SNR so that it better fits the new noise estimate for higher speech quality. In the subjective evaluation with a five-grade mean opinion score (MOS), the scores of the proposed algorithm are improved by as much as 0.93 and 0.35, compared with the original MMSE STSA and the EVRC noise suppression algorithm, respectively. Under 90 % of all tested conditions, the proposed algorithm outperforms either or both of the conventional algorithms with a statistically significant MOS difference.

6. ACKNOWLEDGMENT

The authors would like to thank Tatsuya Miyamoto of Nippon Net System Limited for his help in the simulations and experiments.

7. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, no. 3, pp. 197-210, Jun. 1978.

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

[3] O. Cappe, "Elimination of the musical noise phenomenon with Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Processing, vol. 2, no. 2, pp. 345-349, Apr. 1994.

[4] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," Proc. ICASSP'96, pp.629-632, May 1996.

[5] TIA/EIA/IS-127, "Enhanced Variable Rate Codec," Jan 1997.

[6] R. Martin, "Spectral subtraction based on minimum statistics," *EUSIPCO '94*, pp.1182-1185, Sep. 1994.

[7] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," 1996.

[8] M. Kato, A. Sugiyama and M. Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA," Technical Report of IEICE, DSP/IE/MI2001-8, pp.53-60, Apr. 2001.