NOISE REDUCTION USING SIMULTANEOUS MASKING PROPERTY AND SNR VARIATION FOR VARIOUS NOISE CORRUPTIONS

Ching-Ta Lu[1] and Kun-Fu Tseng[2]
[1]Department of Information Communication, Asia University, Wufeng 413, Taiwan, R.O.C.
[2]Department of Electronic Engineering, Chin Min Institute of Technology, Miaoli 351, Taiwan, R.O.C.

ABSTRACT

A speech enhancement algorithm adapted by both intra-frame masking properties of the human auditory system and inter-frame SNR variation is proposed to enhance a speech signal corrupted by colored noise. Herein, we employ a gain factor adapted by the SNR variation to reduce the spectral variation over successive frames, so the effect of musical residual noise can be mitigated. In addition, the masking property of the human ears is also employed to adapt the gain factor, enabling the imperceptive residual noise with energy below the noise masking threshold to be retained. The speech distortion is therefore reduced by preserving more noisy speech signals. Experimental results show that the proposed scheme can efficiently reduce the effect of musical residual noise by rendering residual noise perceptually white.

Index Terms— speech enhancement, perceptual, SNR variation, auditory masking, colored noise.

1. INTRODUCTION

Many speech enhancement algorithms have been proposed to improve speech quality [1]-[6]. However, most methods still suffer from annoying musical residual noise in the case of colored noise corruption. This musical residual noise is caused by randomly spaced spectral peaks that come and go in successive frames, and occur at random frequencies [1]. Some novel schemes attempted to reduce the effect of musical residual noise by the human auditory system [2], [3]. This auditory system is based on the fact that the human ears cannot perceive residual noise when this noise level falls below the noise masking threshold (NMT). Lu and Wang [2] proposed a wavelet-domain optimal linear estimator which incorporated the masking properties of the human auditory system to make the residual noise inaudible. In addition, Lu also derived a smoothing factor as a second stage to reduce the effect of musical residual noise [3]. An accurate estimate of the a priori SNR is critical for eliminating the musical noise. Hasan et al. [4] presented a method to find an improved estimate of the a priori SNR. Hence, this estimated SNR was applied to a subtraction-based algorithm, thus, allowing the effect of musical residual noise to be reduced. Improved results were obtained in terms of speech quality measures for various types of noise at different SNR levels.

Based on the above findings, utilizing either the noise masking properties or the improved estimate of the a priori SNR to adapt a speech enhancement system is beneficial to enhance speech signals degraded by colored noise. However, musical residual noise still exists after denoising. In this paper, we propose to incorporate both the masking properties and the a priori SNR to adapt a gain factor. In turn, employing this gain factor to enhance a noisy speech signal would render the residual noise perceptually white. The effect of musical residual noise is accordingly reduced. Unlike the gain factor adapted by the a priori SNR [1], [4], the proposed gain factor is not only adapted by the a priori SNR, but also adapted by the noise masking threshold (NMT). Experimental results show that the proposed approach outperforms the modified Ephraim and Malah suppression rule [5] and a method adapted by the a priori SNR and the SNR variation of consecutive two frames [4] for enhancing a speech signal corrupted by various kinds of noise.

2. GAIN FACTOR ADAPTED BY NMT AND SNR

A noisy speech signal $y(m,n)$ can be modeled as the sum of clean speech $s(m,n)$ and additive noise $d(m,n)$ in the frame m of the time domain, i.e., $y(m,n)=s(m,n)+d(m,n)$. The spectral estimate of speech signal $\hat{S}(m,\omega)$ is obtained by multiplying a gain factor $g(m,\omega)$ with the noisy spectrum $Y(m,\omega)$ of a subband, i.e. $\hat{S}(m,\omega)=g(m,\omega)\cdot Y(m,\omega)$

A spectral distortion measure $E(m,\omega)$ is defined as the difference between the short-term spectra of clean speech $S(m,\omega)$ and of enhanced speech $\hat{S}(m,\omega)$. This spectral distortion specifies the performance of a speech enhancement system, and is given by $E(m,\omega)=\hat{S}(m,\omega)-S(m,\omega)=E_S(m,\omega)+E_R(m,\omega)$, where the spectra of speech distortion $E_S(m,\omega)$ and that of residual noise $E_R(m,\omega)$ are expressed as

$$E_S(m,\omega)=[g(m,\omega)-1]\cdot S(m,\omega) \quad (1)$$

$$E_R(m,\omega)=g(m,\omega)\cdot D(m,\omega) \quad (2)$$

where $S(m,\omega)$ and $D(m,\omega)$ represent the spectra of speech

and of noise signals, respectively.

Assuming the noise signal is additive, and is uncorrelated with a speech signal. The gain factor $g(m,\omega)$ can be optimized by minimizing the short-term spectral energy associated with the speech distortion, subject to a constraint on the short-term spectral energy related to residual noise below the noise masking threshold (NMT) [2]:

$$\min_{g(m,\omega)}\left\{E_S^2(m,\omega)\right\} \quad \text{subject to the constraint } E_R^2(m,\omega) \le T(m,\omega) \tag{3}$$

where $T(m,\omega)$ is the NMT corresponding to the frequency bin ω. The NMTs are all identical in a critical band.

In order to derive the gain factor $g(m,\omega)$, a cost function $J_P(m,\omega)$ can be formulated in terms of the speech distortion and the residual noise [2]:

$$J_P(m,\omega) = E_S^2(m,\omega) + \mu(m,\omega)\cdot[E_R^2(m,\omega) - T(m,\omega)] \tag{4}$$

where $\mu(m,\omega)$ is the Lagrangian multiplier.

Substituting (1) and (2) into (4), then to partially differentiate $J_P(m,\omega)$ with respect to the gain factor $g(m,\omega)$, an optimal gain factor can be derived as

$$g^*(m,\omega) = \frac{\gamma_{prio}(m,\omega)}{\gamma_{prio}(m,\omega) + \mu(m,\omega)} \tag{5}$$

where $\gamma_{prio}(m,\omega) = |S(m,\omega)|^2 / |D(m,\omega)|^2$ is defined as the a priori SNR.

The optimal Lagrangian multiplier $\mu(m,\omega)$ can be derived as [2]

$$\mu(m,\omega) = \max\left(\sqrt{\frac{|S(m,\omega)|^2}{T(m,\omega)}}\cdot\sqrt{\frac{1}{\gamma_{prio}(m,\omega)}} - 1,\ 0\right)\cdot\gamma_{prio}(m,\omega) \tag{6}$$

Substituting (6) into (5), a perceptual gain factor $g_P(m,\omega)$ can be derived as

$$g_P(m,\omega) = \frac{1}{1+\max\left(\sqrt{\frac{|S(m,\omega)|^2}{T(m,\omega)}\cdot\frac{1}{\gamma_{prio}(m,\omega)}} - 1,\ 0\right)} \tag{7}$$

The a priori SNR $\gamma_{prio}(m,\omega)$ is unknown and is critical to the perceptual gain factor $g_P(m,\omega)$ in (7). An accurate estimate of the a priori SNR can significantly reduce the musical residual noise produced by a speech enhancement system. An improved estimate of the a priori SNR can be obtained by a time-frequency varying averaging factor $\alpha(m,\omega)$, given as [4]

$$\hat{\gamma}_{prio}(m,\omega) = \alpha(m,\omega)\cdot\hat{\gamma}_{prio}(m-1,\omega) + [1-\alpha(m,\omega)]\cdot P[\gamma_{post}(m,\omega) - 1]\cdot \tag{8}$$

where $\hat{\gamma}_{prio}(m-1,\omega) = |\hat{S}(m-1,\omega)|^2 / |\hat{D}(m-1,\omega)|^2$. $\gamma_{post}(m,\omega) = |Y(m,\omega)|^2 / |\hat{D}(m,\omega)|^2$ is defined as the a posteriori SNR. $|\hat{S}(m-1,\omega)|^2$ and $|\hat{D}(m-1,\omega)|^2$ represent the spectral estimates of speech and of noise in the frame (m-1). $P[\cdot]$ denotes positive half-wave rectification.

The choice of the averaging factor $\alpha(m,\omega)$ in (8) is critical. As $\alpha(m,\omega)$ approaches unity, the SNR varies slightly over the successive frames, enabling the effect of musical residual noise to be reduced. The sacrifice is that $\hat{\gamma}_{prio}(m,\omega)$ with slight variation would fail to respond to the change of a speech signal. On the contrary, if $\alpha(m,\omega)$ approaches zero, the SNR variation goes high. The spectral magnitudes of residual noise vary rapidly in successive frames, resulting in the effect of musical residual noise. Therefore, the SNR estimate $\hat{\gamma}_{prio}(m,\omega)$ given in (8) should be as close as possible to the a priori SNR $\gamma_{prio}(m,\omega)$. A minimum-mean-square-error estimator was proposed to optimize $\alpha(m,\omega)$ by minimizing the error $J_\alpha(m,\omega)$ [4],

$$J_\alpha(m,\omega) = E\{[\hat{\gamma}_{prio}(m,\omega) - \gamma_{prio}(m,\omega)]^2 \mid \hat{\gamma}_{prio}(m-1,\omega)\} \tag{9}$$

given the SNR estimate of previous frame $\hat{\gamma}_{prio}(m-1,\omega)$.

Substituting (8) into (9), the error function $J_\alpha(m,\omega)$ can be decomposed as

$$\begin{aligned} J_\alpha(m,\omega) &= \alpha^2(m,\omega)\cdot\hat{\gamma}_{prio}(m-1,\omega) + [\alpha(m,\omega)-1]^2\cdot \\ &E\{[\gamma_{post}(m,\omega)-1]^2\} + E\{\gamma_{prio}^2(m,\omega)\} + 2\cdot\alpha(m,\omega)\cdot \\ &[1-\alpha(m,\omega)]\cdot\hat{\gamma}_{prio}(m-1,\omega)\cdot E\{\gamma_{post}(m,\omega)-1\} - \\ &2\cdot[1-\alpha(m,\omega)]\cdot E\{[\gamma_{post}(m,\omega)-1]\cdot\gamma_{prio}(m,\omega)\} \\ &-2\cdot\alpha(m,\omega)\cdot\hat{\gamma}_{prio}(m-1,\omega)\cdot E\{\gamma_{prio}(m,\omega)\} \end{aligned} \tag{10}$$

The first and the second moments of the a posteriori SNR are given as

$$E\{\gamma_{post}(m,\omega)\} = \gamma_{prio}(m,\omega) + 1 \tag{11}$$

and

$$E\{\gamma_{post}^2(m,\omega)\} = 2\cdot\gamma_{prio}^2(m,\omega) + 2\cdot\gamma_{prio}(m,\omega) + 1. \tag{12}$$

In addition, we also use the relationship [6],

$$E\{|S(m,\omega)|^4\}/\sigma_D^4(m,\omega) = 2\cdot\gamma_{prio}^2(m,\omega). \tag{13}$$

which follows from the definition of the fourth moment with the assumption that speech spectral magnitude $|S(m,\omega)|$ has a Rayleigh distribution.

Substituting (11), (12) and (13) into (10), hence, differentiating the error function with respect to $\alpha(m,\omega)$, and setting the result to zero, an optimal smoothing factor is then derived as

$$\alpha^*(m,\omega) = 1/\left(1+\left(\frac{\hat{\gamma}_{prio}(m-1,\omega) - \gamma_{prio}(m,\omega)}{\gamma_{prio}(m,\omega)}\right)^2\right) \tag{14}$$

As the priori SNR $\gamma_{prio}(m,\omega)$ is unknown, an approximate value of $\alpha^*(m,\omega)$ given in (14) is essential. For simplicity, we assume that the a posteriori SNR $\gamma_{post}(m,\omega)$ is greater than unity. The a priori SNR can be approximately computed by $\gamma_{prio}(m,\omega) \approx \gamma_{post}(m,\omega) - 1$.

In turn, substituting the optimal averaging factor $\alpha^*(m,\omega)$ into the term of $\alpha(m,\omega)$ in (8) yields an improved estimate of the a priori SNR $\hat{\gamma}_{prio}(m,\omega)$ given as

$$\hat{\gamma}_{prio}(m,\omega) = \frac{\gamma_{prio}(m,\omega)\cdot\{\hat{\gamma}_{prio}(m-1,\omega)\cdot\gamma_{prio}(m,\omega) + \Delta_{m,m-1}^2(\omega)\}}{\gamma_{prio}^2(m,\omega) + \Delta_{m,m-1}^2(\omega)} \tag{15}$$

where

$$\Delta_{m,m-1}(\omega) = \gamma_{prio}(m,\omega) - \hat{\gamma}_{prio}(m-1,\omega). \quad (16)$$

$\Delta_{m,m-1}(\omega)$ represents the SNR variation for the frames m and m-1 in the subband ω.

In order to understand the smoothed version of SNR is varied by the SNRs of previous frame $\hat{\gamma}_{prio}(m-1,\omega)$ and of current frame $\gamma_{prio}(m,\omega)$. The relationship among them is demonstrated in Fig. 1. In the case of noise-dominated region, the a posteriori SNR $\gamma_{post}(m,\omega)$ tends to uniform variation. The smoothing factor $\alpha^*(m,\omega)$ given in (14) would approach unity. Applying this smoothing factor to estimate the SNR given in (8) would reduce the variation of the SNR estimate. In turn, applying this SNR to adapt a speech enhancement algorithm can reduce the spectral variation in successive frames. Accordingly, the spectra of residual noise are less annoying while the effect of musical residual noise is reduced. In the case of speech-dominated region, $\alpha^*(m,\omega)$ attains a small value, enabling $\hat{\gamma}_{prio}(m,\omega)$ given in (15) to respond to the change of speech signal. The speech quality is therefore maintained.

Substituting (15) into (7), the proposed gain factor can be derived as shown in (17). This gain factor $g^*(m,\omega)$ is varied with the intra-frame NMT $T(m,\omega)$, and the inter-frame SNR variation $\Delta_{m,m-1}(\omega)$. If the SNR variation $\Delta_{m,m-1}(\omega)$ is much greater than the a priori SNR $\gamma_{prio}(m,\omega)$, the smoothing factor given in (14) tends to zero. The smoothing effect is only slight in estimating the SNR. It enables the sudden change of speech signal to be preserved, such as unvoiced speech signals. Conversely, if the SNR variation $\Delta_{m,m-1}(\omega)$ is much less than the a priori SNR $\gamma_{prio}(m,\omega)$, the smoothing factor $\alpha^*(m,\omega)$ given in (14) approaches unity. As the optimal smoothing factor $\alpha^*(m,\omega)$ increases, the smoothing effect increases in estimating the SNR. Substituting this SNR estimate to (17) yields a higher value of gain factor, enabling the noisy speech signal to be retained. Accordingly, the speech distortion decreases.

3. EXPERIMENTAL RESULTS

In the experiments, speech signals are Mandarin Chinese spoken by five female and five male speakers. Noisy speech signals are obtained by adding a clean speech signal with F16-cockpit, factory, and babble (speech-like) noise signals which are extracted from the Noisex-92 database. Three SNR levels, including 0 dB, 5 dB and 10dB, are used to evaluate the performance of a speech enhancement system.

The minimum statistics algorithm is performed to estimate the power of noise for each frequency bin [7]. Both modified Ephraim and Malah Suppression Rule (modEMSR) and Hasan's method [4] were conducted for comparisons.

Table 1 presents the performance comparisons in terms of the modified Bark spectral distortion (MBSD). The minimal MBSD corresponds to the best speech quality [8]. In the cases of factory and babble noise corruption, the modEMSR method slightly outperforms Hasan's method. The proposed method obtained much lower values of the MBSD than the other two methods. It is attributed to that the proposed method employs the NMT $T_j^i(m)$ as a major parameter to adapt the gain factor given in (17). In addition, the proposed method also determined the gain factor based on the critical band which matches the perception of the human ears. These two reasons enable the proposed method to substantially outperform the other two methods in terms of the MBSD.

Figure 2 shows the spectrograms of a speech signal corrupted by factory noise (Fig. 2(b)) with *Avg*_SegSNR = 0 dB. Thus, three methods do not over-attenuate the noisy speech signal. The enhanced speech signals demonstrated in Figs. 2(c), 2(d) and 2(e), are not suffered from serious speech distortion. In addition, the spectrograms also reveal fine structure of spectra in speech-activity regions. Therefore, a muffled signal is absent at the output of each speech enhancement method. Comparing the spectrograms in speech-pause regions, Hasan's method shown in Fig. 2(d) is better able to remove the background noise than the other two methods shown in Figs. 2(c) and 2(e). However, the spectral variation in Fig. 2(d) for Hasan's method is the largest, causing chirps in the enhanced speech signal. In Fig. 2(e), the spectrogram of residual noise tends to perceptually white for the proposed method, enabling the residual noise to sound less annoying. Consequently, adapting a gain factor by both inter-frame SNR variation and intra-frame masking property is beneficial to render the residual noise perceptually white.

4. CONCLUSIONS

Integrating both the intra-frame masking properties of the human ears and the inter-frame SNR variation to adapt the gain factor of a subband was proposed. Experimental results show that this gain factor cannot only remove the background noise, but also renders residual noise perceptually white for the colored noise corruptions. Accordingly, the proposed approach offers less annoying residual noise than a method adapted by the SNR only or by the inter-frame SNR variation, such as the modified EMSR and Hasan's methods.

$$g^*(m,\omega) = \frac{1}{1+\max\left(\sqrt{\frac{|S(m,\omega)|^2}{T(m,\omega)} \cdot \frac{\gamma_{prio}^2(m,\omega)+\Delta_{m,m-1}^2(\omega)}{\gamma_{prio}(m,\omega)\cdot[\hat{\gamma}_{prio}(m-1,\omega)\cdot\gamma_{prio}(m,\omega)+\Delta_{m,m-1}^2(\omega)]}} - 1,\ 0\right)} \quad (17)$$

Table 1. Performance comparisons of modified Bark spectral distortion for the enhanced speech in various noises.

Noise type	SNR (dB)	modified Bark spectral distortion noisy	modEMSR	Hasan	proposed
F16	0	16.71	6.96	6.56	2.69
	5	8.01	3.02	2.93	1.09
	10	3.47	0.99	1.05	0.35
factory	0	16.05	6.96	6.99	3.62
	5	7.32	2.89	2.89	1.51
	10	3.08	1.01	1.12	0.56
babble	0	13.65	7.03	7.36	3.59
	5	6.15	2.73	2.97	1.43
	10	2.41	0.99	1.15	0.52

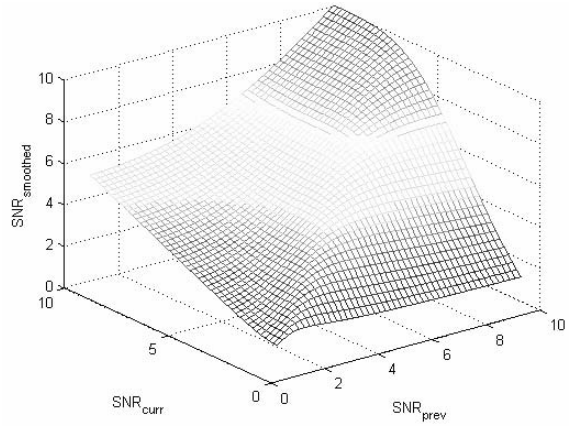

Fig. 1. Smoothed SNR versus the SNRs of previous (SNRprev) and current (SNRcurr) frames.

ACKNOLEDGEMENTS

This research was sponsored by the National Science Council, Taiwan, under contract number NSC 96-2221-E-243-001.

REFERENCES

[1] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 1, pp. 59-67, Jan. 2004.

[2] C. -T. Lu and –H. C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electron. Lett.*, vol.40, no.6, pp.394-396, Mar. 2004.

[3] C. –T. Lu, "Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing," *Pattern. Recog. Lett.*, vol.28, pp. 1300-1306, Aug. 2007.

[4] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Processing Lett.*, vol. 11, no. 4, pp. 450-453, Apr. 2004.

[5] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. IEEE Workshop Statistical Signal Processing*, pp. 496-499, 2001.

[6] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 328-337, July 1998.

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

[8] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Proc. IEEE ICASSP*, 1998.

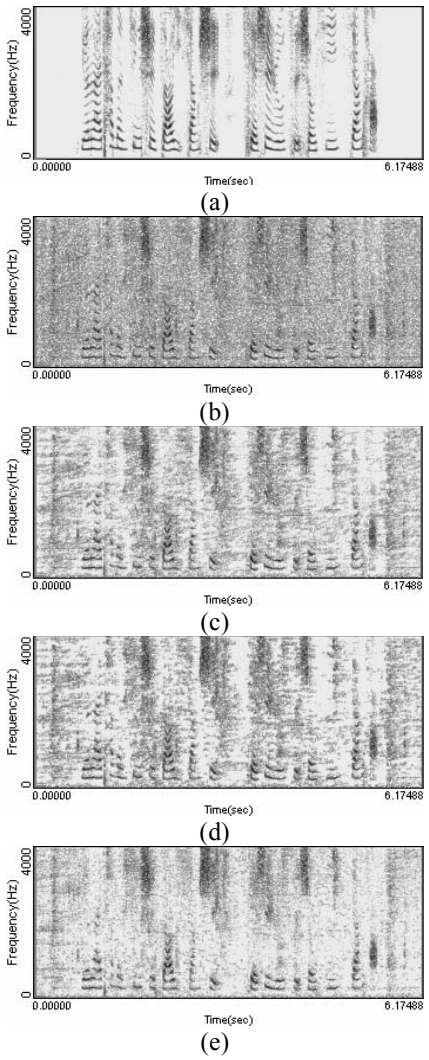

Fig. 2. Spectrograms of clean speech spoken by a female speaker, (a) clean speech, (b) noisy speech (corrupted by factory noise with average SegSNR = 0 dB), (c) enhanced speech (modified EMSR), (d) enhance speech (Hasan's method), and (e) enhanced speech (the proposed method).