LETTER

A Robust Room Inverse Filtering Algorithm for Speech Dereverberation Based on a Kurtosis Maximization

Jae-woong JEONG[†a)], ***Nonmember***, **Young-cheol PARK**[††], ***Member***, **Dae-hee YOUN**[†††], ***and*** **Seok-Pil LEE**[††††], ***Nonmembers***

SUMMARY In this paper, we propose a robust room inverse filtering algorithm for speech dereverberation based on a kurtosis maximization. The proposed algorithm utilizes a new normalized kurtosis function that nonlinearly maps the input kurtosis onto a finite range from zero to one, which results in a kurtosis warping. Due to the kurtosis warping, the proposed algorithm provides more stable convergence and, in turn, better performance than the conventional algorithm. Experimental results are presented to confirm the robustness of the proposed algorithm.

key words: *kurtosis maximization, normalized kurtosis, room inverse filtering, voiced/unvoiced (V/UV) classification*

1. Introduction

Gillespie *et al.* [1] proposed a speech dereverberation algorithm using kurtosis of the linear prediction (LP) residual of speech received by a microphone. By maximizing kurtosis, the algorithm estimates an inverse filter of the room impulse response (RIR). Based on this algorithm, Wu *et al.* [2] proposed a two-stage algorithm for single-microphone reverberant speech enhancement. In Wu's algorithm, an inverse filter for reducing coloration effects of early reflections was estimated by a kurtosis maximization in the first stage. Then, the influence of long-term reverberation was minimized by spectral subtraction in the second stage. After the works by Gillespie *et al.* [1] and Wu *et al.* [2], several online dereverberation methods have been proposed [3].

In this paper, we propose a new room inverse filtering algorithm based on a kurtosis maximization. A major contribution of the proposed algorithm is the convergence robustness which was not obtained in the conventional kurtosis maximization implemented in the time-domain [4]. The robustness of the proposed algorithm stems from a new performance function that implements kurtosis warping, which involves a nonlinear mapping of the unbounded kurtosis to the bounded one within a range from zero to one. The kurtosis warping changes the gradient of the performance function so that it is steep at the transient state and gentle at the steady state. Due to this property, the proposed algorithm provides faster convergence speed and, at the same time, more stable steady state performance than the conventional algorithm. In this paper, we also consider a selective adaptation scheme for practical implementation of the kurtosis maximization algorithm. It is shown that, by allowing adaptation only for voiced speech segments, more robust convergence can be obtained. It is noted that our algorithm focuses on reducing coloration effects as the first stage of Wu's algorithm [2], and a dereverberation system can be constructed by combining the algorithm with spectral subtraction.

The remaining sections are organized as follows: Sections 2 and 3 describe the conventional and the new room inverse filtering algorithms, respectively, based on a kurtosis maximization. Section 4 discusses considerations for practical implementation with actual speech inputs. Performance evaluations and concluding remarks are provided in Sects. 5 and 6, respectively.

2. Room Inverse Filtering Algorithm Using a Kurtosis Maximization

A block diagram of the room inverse filtering system using a kurtosis maximization is depicted in Fig. 1. The reverberant speech signal received by a microphone is $x(n)$ and its LP residual is $x_r(n)$. Assuming that $\mathbf{h}(n) = [h(0), h(1), \ldots, h(L-1)]^T$ is an inverse filter of length L, the LP residual of the inverse-filtered output speech $\widehat{s_r}(n)$ is computed by convolving $\mathbf{h}(n)$ and $\mathbf{x}_r(n) = [x_r(n), x_r(n-1), \ldots, x_r(n-L+1)]^T$, i.e., $\widehat{s_r}(n) = \mathbf{h}^T(n)\mathbf{x}_r(n)$. The inverse filter $\mathbf{h}(n)$ is estimated by maximizing the kurtosis of $\widehat{s_r}(n)$. Finally, we can obtain inverse-filtered speech $\widehat{s}(n)$ by directly applying the replica of $\mathbf{h}(n)$ to the reverberant speech received by a microphone, $x(n)$.

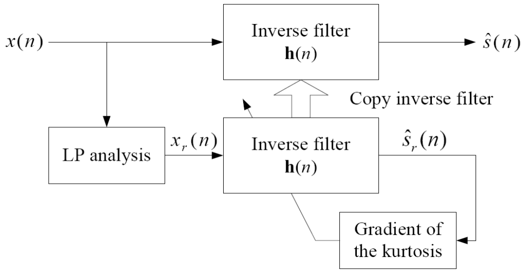


Fig. 1 Block diagram of the room inverse filtering algorithm using a kurtosis maximization.

Manuscript received December 3, 2009.
Manuscript revised January 14, 2010.

[†] The author is with Yonsei Univ., Seoul, Korea.
[††] The author is with Yonsei Univ., Wonju, Kwangwon, Korea.
[†††] The author is with Yonsei Univ., Biometrics Engineering Research Center, Seoul, Korea.
[††††] The author is with Korea Electronics Technology Institute, Sungnam, Kyoungki, Korea.
a) E-mail: woong@dsp.yonsei.ac.kr

DOI: 10.1587/transinf.E93.D.1309

In [1], adaptation equation was derived by maximizing the general form of the normalized kurtosis:

$$C(n) = \frac{E\{\widehat{s}_r^{\,4}(n)\}}{E^2\{\widehat{s}_r^{\,2}(n)\}} - 3. \tag{1}$$

Using the statistic approximation [5], the gradient of $C(n)$ with respect to $\mathbf{h}(n)$ is obtained as

$$\begin{aligned}\frac{\partial C(n)}{\partial \mathbf{h}(n)} &= 4\left(\frac{E\{\widehat{s}_r^{\,2}(n)\}\widehat{s}_r^{\,2}(n) - E\{\widehat{s}_r^{\,4}(n)\}}{E^3\{\widehat{s}_r^{\,2}(n)\}}\right)\\ &\quad \times \widehat{s}_r(n)\mathbf{x}_r(n) \\ &= f(n)\mathbf{x}_r(n).\end{aligned} \tag{2}$$

Consequently, the adaptation equation of the inverse filter can be derived as

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu f(n)\mathbf{x}_r(n), \tag{3}$$

where μ is a step size parameter to control a convergence speed. $E\{\widehat{s}_r^{\,m}(n)\}, m = 2, 4$ can be estimated as

$$E\{\widehat{s}_r^{\,m}(n)\} = \beta E\{\widehat{s}_r^{\,m}(n-1)\} + (1-\beta)\widehat{s}_r^{\,m}(n), \tag{4}$$

where β $(0 \le \beta < 1)$ is a forgetting factor.

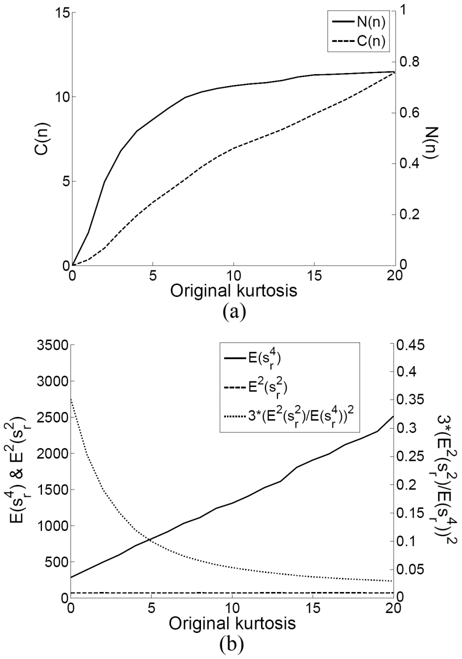
As mentioned previously, the time-domain approach has a problem of slow convergence speed [4]. It is because kurtosis of the input signal is widely diversified, which makes it difficult to set stability bounds of the adaptive algorithm. Although a frequency-domain approach was utilized [2] to solve this problem, we propose a time-domain approach as another solution.

3. Proposed Room Inverse Filtering Algorithm with a New Normalized Kurtosis

The conventional normalized kurtosis was obtained by normalizing the kurtosis by squared variance. In this paper, we define a new normalized kurtosis by normalizing kurtosis by the fourth moment of the signal:

$$N(n) = 1 - 3\frac{E^2\{\widehat{s}_r^{\,2}(n)\}}{E\{\widehat{s}_r^{\,4}(n)\}}. \tag{5}$$

A distinctive feature of the new normalized kurtosis function is that its levels are confined to a finite range. A Gaussian-distributed signal has a kurtosis value of zero, whereas a signal distributed more sharply than Gaussian has a kurtosis greater than zero, i.e., $E\{x^4\} - 3E^2\{x^2\} > 0$. Since speech signals, in general, have distributions sharper than Gaussian, the inequality $E\{x^4\} > 3E^2\{x^2\}$ almost always holds. Thus, it can be deduced that $0 < N(n) < 1$. Figure 2 (a) shows the relationship between the input kurtosis and the normalized kurtoses, $C(n)$ and $N(n)$, according to variation of the input kurtosis from 0 (corresponding to Gaussian) to 20 (corresponding to typical voiced speech). Random noises with these kurtoses were generated through the Pearson system [6]. As can be seen, $C(n)$ maps kurtoses from 0 to 20 almost linearly to a range from 0 to 11, whereas $N(n)$ maps the same kurtoses nonlinearly to a range from 0 to around 0.8. In fact, the range of $C(n)$ is unlimited since $E\{x^4\}$ is monotonically increasing while $E^2\{x^2\}$ is almost uniform, as can be seen from Fig. 2 (b), whereas the kurtosis limit corresponding to the case $E\{x^4\} \gg 3E^2\{x^2\}$ is mapped to $N(n) \approx 1$. Thus, the new function implements a kurtosis warping in which the entire range of kurtosis from zero to infinity is nonlinearly mapped onto a range from zero to one. The mapping is expanded at a low kurtosis region and compressed at a high kurtosis region.



(a)

(b)

Fig. 2 Variations according to increase of kurtosis: (a) $N(n)$ (solid) and $C(n)$ (dashed), (b) the normalization terms (solid and dashed) and the term $3\left(E^2\{\widehat{s}_r^{\,2}(n)\}/E\{\widehat{s}_r^{\,4}(n)\}\right)^2$ (dotted).

Now, similar to the case of $C(n)$, the gradient of $N(n)$ is obtained as

$$\begin{aligned}\frac{\partial N(n)}{\partial \mathbf{h}'(n)} &= 12\left(\frac{E\{\widehat{s}_r^{\,2}(n)\}\widehat{s}_r^{\,2}(n) - E\{\widehat{s}_r^{\,4}(n)\}}{E^2\{\widehat{s}_r^{\,4}(n)\}}\right)\\ &\quad \times E\{\widehat{s}_r^{\,2}(n)\}\widehat{s}_r(n)\mathbf{x}_r(n) \\ &= g(n)\mathbf{x}_r(n),\end{aligned} \tag{6}$$

and the adaptation equation is obtained as

$$\mathbf{h}'(n+1) = \mathbf{h}'(n) + \mu' g(n)\mathbf{x}_r(n), \tag{7}$$

where μ' is also a step size parameter. For abbreviation, the algorithms based on a kurtosis maximization using the conventional normalized kurtosis, $C(n)$, and the new normalized kurtosis, $N(n)$ are respectively referred to as *KM-CNK* and *KM-NNK* hereafter.

Comparing Eqs. (2) and (6), it can be seen that the solution vectors obtained by setting gradients to zero are the same except the case where $\widehat{s}_r(n)$ is zero. Therefore, it can be said that both *KM-CNK* and *KM-NNK* will converge to the same stationary point. However, different normalizations result in differences in detailed shapes of performance function and thereby gradients so that convergence characteristics will be different. As shown in Fig. 2 (a), $C(n)$ has almost constant gradient, whereas $N(n)$ has approximately logarithmic gradient with high gradients for low kurtosis (transient state) and low gradients for high kurtosis (steady state).

Assuming that the update terms in Eqs. (3) and (7) are the same, we can see that the step sizes μ and μ' are related by:

$$\mu = 3\mu' \left(\frac{E^2\{\widehat{s}_r^{\,2}(n)\}}{E\{\widehat{s}_r^{\,4}(n)\}} \right)^2 . \tag{8}$$

The term $3\left(E^2\{\widehat{s}_r^{\,2}(n)\}/E\{\widehat{s}_r^{\,4}(n)\}\right)^2$ decays almost exponentially as the kurtosis increases as can be seen from Fig. 2 (b). Consequently, we can consider the proposed algorithm as a variable step size version of the conventional algorithm. The update term of Eq. (7) will be relatively large in magnitude at the transient state and small as the algorithm converges to the stationary point. This is a desirable feature for obtaining both the fast convergence and stable steady-state response.

4. Implementation Considerations

For practical implementation of the algorithm, effects of input speech distribution on convergence can be considered. For estimation problems with Gaussian noise, adaptations using high-order statistics can't produce meaningful results [8]. Thus, a kurtosis maximization with an input of Gaussian-like speech has no benefits.

Moreover, for high kurtosis input, there will be a fair margin between kurtoses of the dry and reverberant signals so that the adaptive system has a good chance for effective learning of the input process. However, for low kurtosis input, there will be a small margin so that the adaptive system will try to improve kurtosis by modifying the input signal distribution, which will cause signal distortions at output.

In general, the excitations of speech signals are a quasi-periodic pulse train for voiced speech and random noise for unvoiced speech [7]. Therefore, it is highly likely that a kurtosis maximization with an input signal of unvoiced speech has little effect on performance. Consequently, we can say that a selective adaptation based on a V/UV decision of the input speech will be beneficial in practical systems. In this paper, we consider a selective adaptation for the proposed algorithm:

$$\mathbf{h}'(n+1) = \begin{cases} \mathbf{h}'(n) + \mu' g(n)\mathbf{x}_r(n), & \text{if } \mathbf{x} \in V, \\ \mathbf{h}'(n), & \text{otherwise.} \end{cases} \tag{9}$$

For the V/UV decision, various algorithms can be utilized. In this paper, we used a simple one based on pattern recognition with selective features, such as log energy and zero-crossing rate [9]. It is noted that this algorithm was designed for clean speech, and thus we would need more elaborate algorithms for reverberant speech. In our experiments, nonetheless, it is still useful to identify the effectiveness of the selective adaptation.

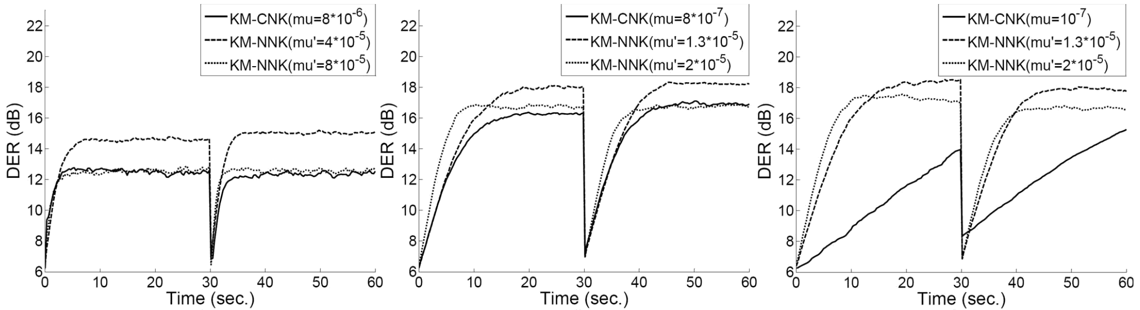
5. Experimental Results

Experiments were conducted using artificially generated noise signals. Random noises with kurtosis of 5, 10, and 20 were generated using the Pearson system [6]. These values, respectively, correspond to average kurtosis of typical unvoiced, mixed, and voiced speech. A room with dimensions of 3 m × 3 m × 2 m and a reverberation time (T_{60}) of 80 ms were assumed. RIR, $r(n)$, was generated using the image method [10] at an 8-kHz sampling rate. The filter was initialized as Dirac delta function and its length L was set to 100, with which most early reflections of the RIR could be covered. The forgetting factor β was 0.999. Since a kurtosis maximization expects to equalize early reflections of RIR, we measured direct-to-early reflection ratio (DER) of the inverse-filtered RIR ($z(n) = r(n) * h(n)$) as a performance metric, which is defined as

$$DER = 10\log_{10}\frac{p_D}{p_E}\ (\text{dB}), \tag{10}$$

where $p_D = \sum_{n=1}^{N_D} z^2(n)$ and $p_E = \sum_{n=N_D+1}^{N_E} z^2(n)$, and N_D and N_E are respectively the separation between direct part and early reflections and that between early and late reflections, which were set to 50 and 100.

Experimental results for noise inputs with various kurtosis are shown in Fig. 3, which were obtained by averaging 30 independent tests. To simulate the time-varying situation, the source-receiver distance was changed from 1 m to 0.5 m



(a) (b) (c)

Fig. 3 DERs for noise inputs with kurtosis (a) 5, (b) 10, and (c) 20.

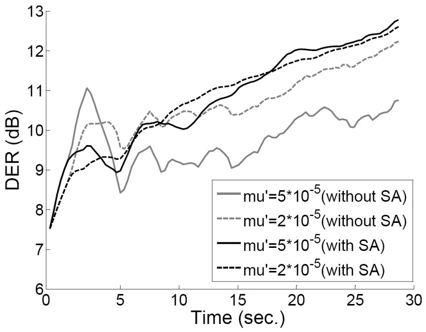


Fig. 4 DERs with (black) or without (gray) SA scheme.

at $t = 30$ s. Initial DERs of the generated RIRs were 6.6 and 10.2 dB, respectively. Various step sizes were tried to test robustness of the algorithms. For several cases, however, it was difficult to select a suitable step size that could make *KM-CNK* converge. On the other hand, *KM-NNK* easily attained convergence with a wide range of step sizes. For the convenience of comparison, we just present results obtained using the largest step sizes that could guarantee stable convergence for *KM-CNK* (see solid lines). Step-sizes for *KM-NNK* were then chosen to have similar transient or steady-state performance to *KM-CNK*. Results in Fig. 3 show that, for *KM-CNK*, the higher the input kurtoses, the smaller the step sizes that guarantee stable convergence. Also, *KM-NNK* converges much faster than *KM-CNK*, and *KM-NNK* arrives at higher DERs when step sizes were determined to have similar convergence speeds. The performance difference between the two algorithms is the most significant for the input with high kurtosis (kurtosis = 20). In this case, since *KM-CNK* could not converge even in 60 s, we could only confirm that *KM-NNK* provides similar performance for both 10 and 20 kurtoses.

In the next simulations, the selective adaptation (SA) of *KM-NNK* in Eq. (9) was tested. For the tests, a male speaker provided speech signals of number one to number ten in English, and these were recorded at a 16-kHz sampling rate. Figure 4 shows DERs with and without the SA scheme, which are for only the first RIR. Results show that the SA scheme can sustain the performance relatively well during the UV segments. On the other hand, the continuous adaptation scheme suffers drops of DER during UV segments. These degradation of performance are more serious for the larger step-size. Based on these results, we conclude that the SA based on a V/UV decision of input speech is advantageous when the algorithm is applied to speech signals.

6. Conclusions

In this paper, we proposed a new room inverse filtering algorithm for speech dereverberation based on a kurtosis maximization. Due mainly to the kurtosis-warping, the proposed algorithm proved to be more robust and had faster convergence than the conventional algorithm and its convergence characteristics for speech input could be improved by using the selective adaptation scheme. By combining the proposed algorithm with spectral subtraction, we can effectively construct a two-stage dereverberation system.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) through the Biometrics Engineering Research Center (BERC) at Yonsei University. (2009-0062989)

References

[1] B.W. Gillespie, H.S. Malvar, and D.A.F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," Proc. IEEE Int. Conf. Acoust., Speech Signal Process., vol.6, pp.3701–3704, May 2001.

[2] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," IEEE Trans. Audio Speech Language Process., vol.14, no.3, pp.774–784, May 2006.

[3] E.A.P. Habets, N.D. Gaubitch, and P.A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pp.4577–4580, April 2008.

[4] S. Haykin, Adaptive Filter Theory, fourth ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.

[5] O. Tanrikulu and A.G. Constantinides, "Least-mean kurtosis: A novel higher-order statistics based adaptive filtering algorithm," Electron. Lett., vol.30, no.3, pp.189–190, Feb. 1994.

[6] K. Pearson, "Second supplement to a memoir on skew variation," Phil. Trans. A 216, pp.429–457, 1916.

[7] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[8] P.I. Hübscher and J.C.M. Bermudez, "Properties of the KURTOSIS performance surface in linear estimation application to adaptive filtering," Proc. IEEE Int. Conf. Acoust., Speech Signal Process., vol.2, pp.837–840, May 2004.

[9] B.S. Atal and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-24, no.3, pp.201–212, June 1976.

[10] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol.65, no.4, pp.943–950, April 1979.