IEICE TRANS. FUNDAMENTALS, VOL.E103–A, NO.10 OCTOBER 2020

LETTER

4th Order Moment-Based Linear Prediction for Estimating Ringing Sound of Impulsive Noise in Speech Enhancement

Naoto SASAOKA[†a)], ***Senior Member***, **Eiji AKAMATSU**[†], ***Nonmember***, **Arata KAWAMURA**[††], **Noboru HAYASAKA**[†††], ***Members***, ***and*** **Yoshio ITOH**[†], ***Senior Member***

SUMMARY Speech enhancement has been proposed to reduce the impulsive noise whose frequency characteristic is wideband. On the other hand, it is challenging to reduce the ringing sound, which is narrowband in impulsive noise. Therefore, we propose the modeling of the ringing sound and its estimation by a linear predictor (LP). However, it is difficult to estimate the ringing sound only in noisy speech due to the auto-correlation property of speech. The proposed system adopts the 4th order moment-based adaptive algorithm by noticing the difference between the 4th order statistics of speech and impulsive noise. The brief analysis and simulation results show that the proposed system has the potential to reduce ringing sound while keeping the quality of enhanced speech.

key words: *speech enhancement, impulsive noise, adaptive filter, high order statistics*

1. Introduction

For stationary noise or non-stationary noise whose frequency characteristic is gently changed, a short time spectral amplitude analysis, e.g., minimum mean square estimation - short time spectral amplitude, is widely used as a single channel speech enhancement [1]. On the other hand, there is the impulsive noise occurring by hitting an object. It has an initial peak sound whose frequency characteristics is wideband, and then changes to a ringing sound whose is narrowband associated with a natural oscillation of the object.

Speech enhancement methods for the impulsive noise without a ringing sound have been proposed [2], [3]. These methods take advantage of flat frequency characteristics. These methods are not suitable to reduce ringing sound. Unfortunately, the ringing sound interrupts speech conversation than an initial peak sound because the ringing sound keeps for a long time. Therefore, the speech enhancement methods based on a zero-phase signal have been proposed to reduce not only an initial peak sound not also a ringing sound in [4], [5]. However, these methods reduce high pitch ringing sound only and degrade the quality of enhanced speech due to using a zero-phase signal.

Manuscript received January 16, 2020.
Manuscript publicized April 2, 2020.

[†]The authors are with the Department of Electrical Engineering and Computer Science, Faculty of Engineering, Tottori University, Tottori-shi, 680-8552 Japan.
[††]The author is with the Faculty of Information Science and Engineering, Kyoto Sangyo University, Kyoto-shi, 603-8555 Japan.
[†††]The author is with the Department of Engineering Informatics, Osaka Electro-Communication University, Neyagawa-shi, 572-8530 Japan.
a) E-mail: sasaoka@tottori-u.ac.jp

DOI: 10.1587/transfun.2020EAL2005

Therefore, we focus on a ringing sound reduction in this letter. First of all, we propose the model of the noise generation process by an all-pole filter. Therefore, a ringing sound can be estimated by an LP. However, the 2nd order moment-based adaptive algorithm, e.g., a least mean square (LMS) algorithm, cannot avoid degrading the quality of enhanced speech due to the auto-correlation of speech. Therefore, we take advantage of the difference between the 4th order statistics of speech and impulsive noise. Since the normalized kurtosis of impulsive noise is enough higher than that of speech in a short-time analysis [6], an LP can estimate only a ringing sound by introducing a least mean fourth (LMF) adaptive algorithm based on 4th order moment [7]. Although the literature [7] proved that LMF algorithm improves the estimation accuracy of system identification in case that an input signal and disturbance are Gaussian and non-Gaussian respectively, it did not verify the behavior of the LMF algorithm on an LP when its input signal is non-Gaussian. Thus we show the effectiveness of the LMF algorithm on an LP by a brief analysis in this letter.

2. Linear Prediction for Ringing Sound

The impulsive noise converts an initial peak sound due to a deformation of an object to a ringing sound due to the natural mode radiation [8]. The initial peak sound has a wideband component, and the ringing sound is composed of a narrowband component, which has some resonance frequencies. Besides, about an experimental study on impulsive sound, which does not include a ringing sound, Yoshimura et al. showed that the impulsive noise is super-Gaussian. Its normalized kurtosis is distributed from 50 to 150 or is higher than 150 [6]. Consequently, the ringing sound is assumed to occur by exciting an all-pole filter with plural resonance peaks by non-Gaussian wideband noise in this letter. Because an LP can estimate an all-poll filter [9], we can obtain the ringing sound as its output signal.

Noisy speech $x(n)$ at time index n is represented as

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

where $s(n)$ and $d(n)$ are respectively speech and impulsive noise. The proposed method estimates a ringing sound from noisy speech by an LP. The estimated ringing sound of the LP $\hat{d}(n)$ is given by

$$\hat{d}(n) = \mathbf{x}^T(n)\mathbf{h}(n) \tag{2}$$

$$\mathbf{x}(n) = \begin{bmatrix} x(n-1) & \cdots & x(n-M) \end{bmatrix}^T \quad (3)$$

$$\mathbf{h}(n) = \begin{bmatrix} h_1(n) & h_2(n) & \cdots & h_M(n) \end{bmatrix}^T \quad (4)$$

where $\mathbf{x}(n)$ and $\mathbf{h}(n)$ are respectively a tap input vector and coefficient vector of the LP. M is the number of tap coefficients of the LP. $h_i(n)$ is i-th tap coefficient of the LP. The error signal of the LP is represented as

$$e(n) = x(n) - \hat{d}(n). \quad (5)$$

3. 4th Order Moment-Based Adaptive Algorithm

In this section, we consider an adaptive algorithm for estimating a ringing sound only in noisy speech. In the case of the adaptive algorithm based on 2nd order moment, LP estimates not only ringing sound but also speech because the speech in tap inputs has high auto-correlation.

The initial peak sound has a high normalized kurtosis, as explained in Sect. 2. In contrast, the normalized kurtosis of speech concentrates at about zero in the short-time analysis [6], although speech is known to be super-Gaussian with high kurtosis in a long-time analysis [10]. That is the reason there is a possibility to be only periodic voiced sound or unvoiced sound, whose source is white noise, in a short-time analysis duration. Thus the proposed method adopts an LMF adaptive algorithm, which is based on 4th order moment and converges on a solution without local minima because of a convex cost function. The LMF algorithm is given by [7]

$$\mathbf{h}(n+1) = \mathbf{h}(n) + 4\mu e^3(n)\mathbf{x}(n), \quad (6)$$

where μ is a step size.

We will analysis the influence of speech on an LMF algorithm. The error signal of the LP $e(n)$ is represented as

$$e(n) = e_d(n) + e_s(n) \quad (7)$$

where $e_d(n) = d(n) - \mathbf{d}^T(n)\mathbf{h}(n)$ and $e_s(n) = s(n) - \mathbf{s}^T(n)\mathbf{h}(n)$ represent error components about impulsive noise and speech respectively. $\mathbf{d}(n)$ and $\mathbf{s}(n)$ are impulsive noise vector and speech vector in a tap input vector $\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{d}(n)$. The cost function of the LMF algorithm is expressed by

$$\begin{aligned} E[e^4(n)] &= E[\{e_d(n) + e_s(n)\}^4] \quad (8) \\ &\approx E[e_d^4(n)] + 6E[e_d^2(n)]E[e_s^2(n)] \end{aligned}$$

where the all odd order moments are zero assuming that probability density functions of speech and impulsive noise are symmetry around zero. In addition, the 4th order moment of impulsive noise is assumed to be enough greater than that of speech. When speech is absent, $E[e^4(n)]$ becomes $E[e_d^4(n)]$ and then the LP estimates ringing sound. When the speech is present and the impulsive noise is absent, $E[e^4(n)]$ is assumed to be about zero because the normalized kurtosis concentrates at about zero and the speech has smaller kurtosis than impulsive noise. Thus, the speech does not influence the tap coefficients of the LP. When the input signal includes speech and impulsive noise, the LP can estimate the ringing sound without bias free under the condition that the 4th order moment of impulsive noise is enough greater than the power of speech. Thus, speech is estimated as the error signal $e(n)$ by minimizing the 4th order moment of the error signal.

4. Computer Simulations

The performance of the proposed speech enhancement method was evaluated by computer simulation. In this paper, the number of tap coefficients M was set to 1088. To evaluate the speech enhancement ability and the quality of enhanced speech, we used overall signal to noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) [11]. The input and output overall SNRs are respectively defined as follows:

$$SNR_{in} = 10\log_{10}\frac{\sum_{n=1}^{N'} s^2(n)}{\sum_{n=1}^{N'} d^2(n)} \quad [\mathrm{dB}] \quad (9)$$

$$SNR_{out} = 10\log_{10}\frac{\sum_{n=1}^{N'} s^2(n)}{\sum_{n=1}^{N'} \{s(n) - e(n)\}^2} \quad [\mathrm{dB}] \quad (10)$$

where N' is the number of samples. All sound data prepared in simulations were sampled at 16 kHz. The five male and five female speech data contained in the Acoustic Society of Japan-Japanese Newspaper Article Sentences (ASJ-JNAS) [12] were used. As the impulsive noise, we used cup noise in the RWCP sound scene database in real acoustical environments. The noisy speech $x(n)$ was composed of clean speech $s(n)$ and impulsive noise $d(n)$ in a -5 dB input overall SNR environment. The impulsive noise occurred only once. We calculated the improvements in SNR and PESQ, which were the difference of an output object measure from an input objective measure. We carried out 100 independent computer simulations in which the generation timing of the impulsive noise changes at random, and we averaged the improvements in SNR and PESQ respectively.

Figure 1 shows the average improvement in SNR vs. step size. Since the suitable step size depends on the structure of an LP and an adaptive algorithm, we compare the maximum average improvements in SNR obtained by speech enhancement systems. Comparing the LP using an LMF algorithm with the LP using an LMS algorithm, the maximum average improvement in SNR was increased from 6.21 dB to 9.63 dB by the proposed system. The average improvement in PESQ vs. step size is shown in Fig. 2. The proposed system increases maximum average improvement in PESQ from 0.35 to 0.44 compared to the LP using an LMF algorithm with the LP using an LMS algorithm. From the improvements in SNR and PESQ shown in Figs. 1 and 2, it can be seen that the LP has the potential to enhance speech corrupted by impulsive noise regardless of an LMS or LMF algorithm. Besides, an LMF algorithm improves the speech

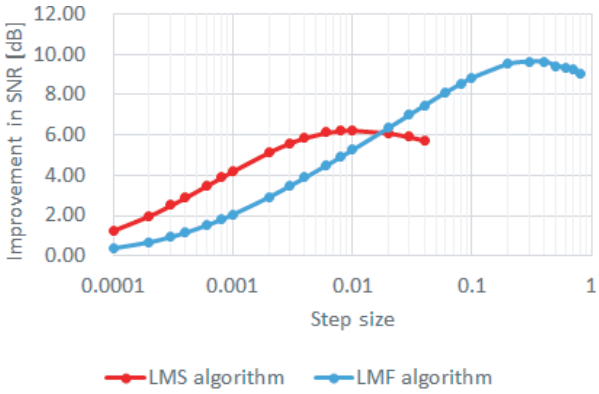


Fig. 1 Average improvement in SNR vs. Step size.

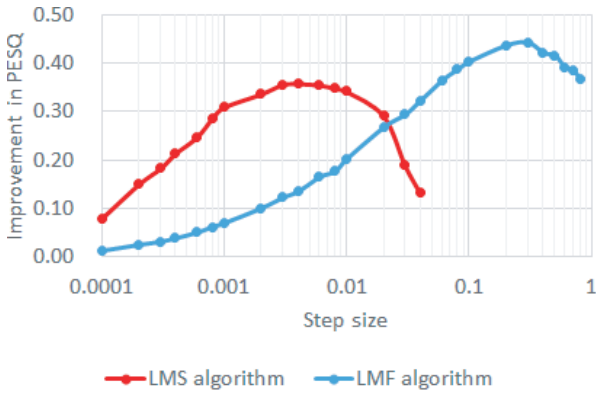
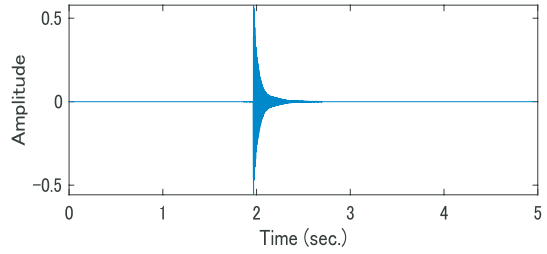
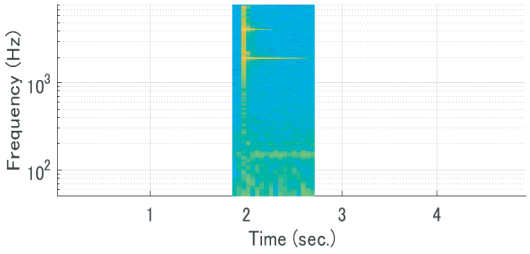


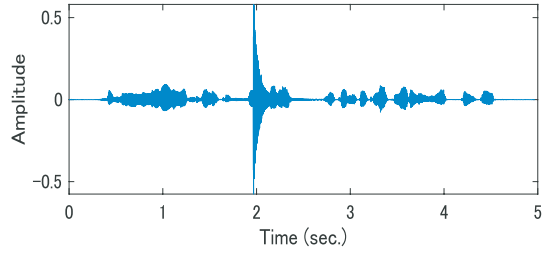
Fig. 2 Average improvement in PESQ vs. Step size.



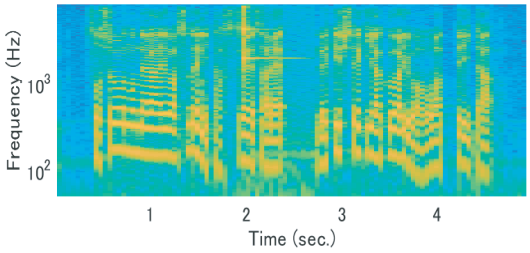
(a) Waveform of impulsive noise



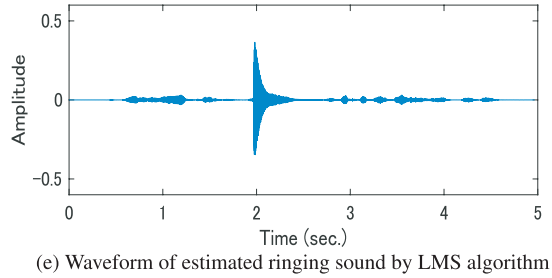
(b) Spectrogram of impulsive noise



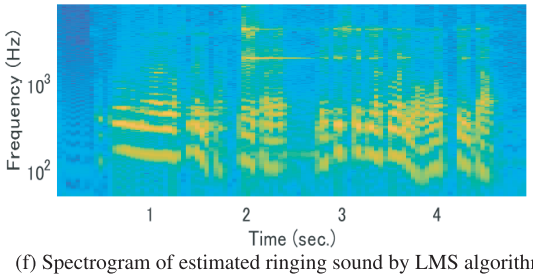
(c) Waveform of noisy speech



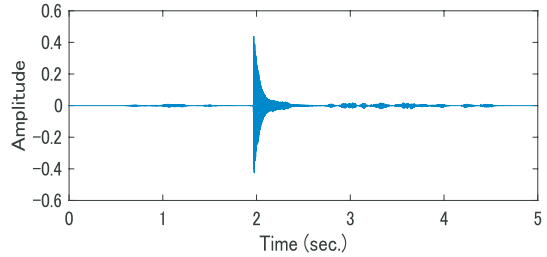
(d) Spectrogram of noisy speech



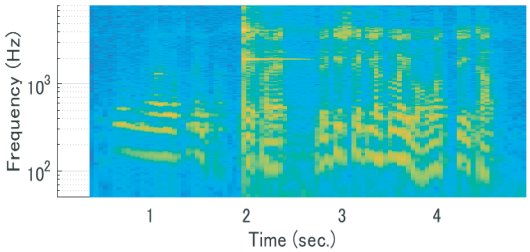
(e) Waveform of estimated ringing sound by LMS algorithm



(f) Spectrogram of estimated ringing sound by LMS algorithm



(g) Waveform of estimated ringing sound by LMF algorithm



(h) Spectrogram of estimated ringing sound by LMF algorithm

Fig. 3 Waveforms and spectrograms of estimated ringing sound.

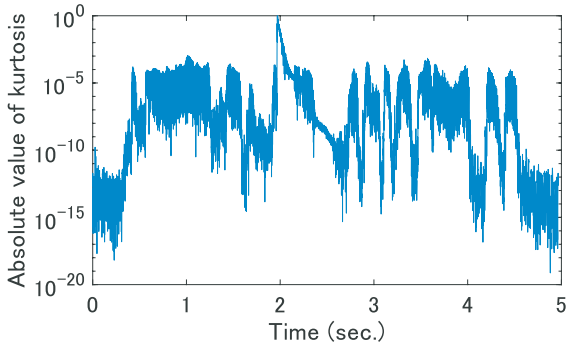


Fig. 4 Kurtosis of noisy speech.

enhancement ability than an LMS algorithm.

The estimated ringing sound is shown in Fig. 3 to verify the factor in the improvement of SNR and PESQ. In this figure, the spectrogram shows the power spectrum from 50 Hz to 8,000 Hz. As a speech signal, we used one of the male speech datasets. Figure 3(a) and (b) respectively represent the waveform and spectrogram of impulsive noise in a −5 dB input SNR environment, respectively. Figure 3(c) and (d) respectively depict the waveform and spectrogram of the noisy speech in a −5 dB input SNR environment. The waveform and spectrogram of estimated ringing sound by the LP with an LMS algorithm are shown in Fig. 3(e) and (f), respectively. The step size for the LMS algorithm was set to 0.008. The improvements in SNR and PESQ were 6.44 dB and 0.32, respectively. The waveform and spectrogram of the ringing sound estimated by the LP with an LMF algorithm are shown in Fig. 3(g) and (h), in which the improvements in SNR and PESQ were 10.15 dB and 0.62, respectively. The step size for the LMF algorithm was set to 0.3. The waveforms and spectrograms of the estimated ringing sounds show that the LP can estimate ringing sound, and the LMF algorithm prevents the estimation of speech components, especially harmonics of speech.

Figure 4 shows the absolute value of kurtosis of noisy speech. The kurtosis $k(n)$ at time n is given from 4th order and 2nd order moments $M_4(n)$ and $M_2(n)$ by

$$k(n) = M_4(n) - 3\{M_2(n)\}^2 \tag{11}$$

$$M_4(n) = \alpha M_4(n-1) + (1-\alpha)x^4(n-1) \tag{12}$$

$$M_2(n) = \alpha M_2(n-1) + (1-\alpha)x^2(n-1) \tag{13}$$

where α is a forgetting factor. α was set to 0.9 in this simulation. We used the data shown in Fig. 3(c) as noisy speech. The absolute value is normalized by maximum absolute value of kurtosis. The kurtosis is maximum when the impulsive noise is present according to Fig. 4. Compared to the kurtosis of impulsive noise, the kurtosis of speech is sufficiently small. Thus, the proposed method can estimate the ringing sound while preventing the LP from estimating speech.

5. Conclusions

In this letter, we have proposed the speech enhancement for ringing sound by an LP. An all-pole filter models the ringing sound, and then the ringing sound is obtained as the output of the LP. Unfortunately, the LP estimates not only the ringing sound but also speech. Therefore, the proposed system adopts the LMF algorithm due to the difference between 4th order statistics of impulsive noise and speech. From the brief analysis and simulation results, the LP with an LMF algorithm has the potential to estimate the ringing sound while avoiding estimating the speech component. In our future work, we will research the effectiveness evaluation for various impulsive noise, and adaptive algorithm or structure of an LP for further improvement of the enhanced speech quality.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17K00272.

References

[1] P.C. Loizou, Speech Enhancement, CRC Press, 2007.
[2] R.C. Nongpiur, "Impulse noise removal in speech using wavelets," 2008 IEEE Int. Conf. Acoustics, Speech, and Signal Process., pp.1593–1596, March 2008.
[3] A. Sugiyama, R. Miyahara, and P. Kwangsoo, "Impact-noise suppression with phase-based detection," Proc. EUSIPCO2013, pp.1–5, Sept. 2013.
[4] S. Kohmura, A. Kawamura, and Y. Iiguni, "A zero phase noise reduction method with damped oscillation estimator," IEICE Trans. Fundamentals, vol.E97-A, no.10, pp.2033–2042, Oct. 2014.
[5] A. Kawamura, N. Hayasaka, and N. Sasaoka, "Impact and high-pitch noise suppression based on spectral entropy," IEICE Trans. Fundamentals, vol.E99-A, no.4, pp.777–787, April 2016.
[6] T. Yoshimura, F. Asano, H. Asoh, and N. Kitawaki, "Investigation of voise/sound activity classifier using distribution models of fourth-order statistics," IPSJ SIG Technical Report, HI-109, July 2004 (in Japanese).
[7] E. Walach and B. Widrow, "The least mean fourth (LMF) adaptive algorithm and its family," IEEE Trans. Inf. Theory, vol.IT-30, no.2, pp.275–283, March 1984.
[8] A. Akey, "A review of impact noise," J. Acoustic Soc. Am., vol.64, no.4, pp.977–987, Oct. 1978.
[9] S. Haykin, Adaptive Filter Theory, Prentice Hall, 1996.
[10] T.W. Won, Independent Component Analysis, Kluwer Academic Publishers, 1998.
[11] ITU, Perceptual evaluation of speech quality, and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech coders, ITU-T Recommendation, P.862, 2000.
[12] http://research.nii.ac.jp/src/eng/index.html