A RESTRICTED IMPACT NOISE SUPPRESSOR IN ZERO PHASE DOMAIN

Arata KAWAMURA

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

ABSTRACT

This paper proposes an impact noise suppression method in zero phase (ZP) domain. The signal in ZP domain (ZP signal) is obtained by taking IDFT of the pth power of a spectral amplitude. We previously proposed an impact noise suppressor in ZP domain for reducing impact noise signals, even if they are accompanied with damped oscillation. Unfortunately, the previous method causes speech degradation in non-impact noise segments, because this method performs noise reduction in all the segments. Since an impact noise exists only a short duration, we restrict the noise suppression procedure so that it cannot be applied to the non-impact noise segments. The restriction is achieved by using the ratio of the first to the second peak values of the ZP signal. In non-impact noise segments, this ratio becomes much larger than one. Thus, we can improve speech quality of the extracted signal when the restriction works well. Simulation results show that the proposed method improves about 15dB of SNR for a speech signal mixed with clap noise with SNR$=0$dB.

Index Terms— Zero Phase Signal, Speech Enhancement, Noise Suppression, Impact Noise, Damped Oscillation

1. INTRODUCTION

Noise reduction techniques are required to extract a speech signal from an observed signal which includes noise. Many noise reduction methods have been proposed in the last decades [1]– [12]. Most of them are of stationary noise suppression. Recently, some researchers have focused on impulsive noise suppression which is difficult to be achieved, because we do not know when it occurs. When the target noise can be expressed as an ideal impulse, we can remove it by using a simple median filter [2], which is preferably used in the image processing area. The median filter is not suitable for reducing a real-environmental impulsive noise. A phase randomization method has been proposed for suppressing auto-focus noise caused in a camera[8, 9]. This method randomizes the phase spectrum of the observed signal, while the impulsive noise has a linear phase characteristic. Then, the impulsive noise can not form a peak in time domain, and it is spread over in a long time. This method requires an impulsive noise existence estimator, because the phase randomization also breaks the phase spectrum of speech, i.e., speech quality is degraded. As shown in [8, 9], a noise detection method was established by using spectral peak and hangover of speech. Since this detection method relies on the periodicity of speech, it may not be effective for an impact noise which is accompanied with damped oscillation. Talman *et al.* established an impact noise reduction method under the assumption that the noise event is repeatedly occurred and the time duration of each event is short compared to speech phonemes [10]. This method is based on capturing similar patterns of noise. Hence, it can reduce even if the impact noise is accompanied with damped oscillation. Unfortunately, an impact noise whose pattern is time variant may not be reduced.

As an attractive method, there exists an impact noise suppressor in zero phase (ZP) domain [11, 12]. A signal in ZP domain (ZP signal) is obtained by taking IDFT of the pth power of spectral amplitude, typically $p = 1$. When the spectral amplitude is approximately flat, its ZP signal has values at only around the origin [1]. On the other hand, a ZP signal for a periodic signal becomes also a periodic signal whose period is unchanged and its maximum value always exists at the origin. Voiced speech signals can be approximated by a periodic signal. When a voiced speech signal is mixed with an impact noise, we can reduce it by replacing some samples around the origin with some samples around the second or latter period of the ZP signal [11]. Since the impact noise with damped oscillation cannot be reduced by the simple replacement technique of the ZP signal, because the ZP signal of damped oscillation is periodic as well as speech. To solve this problem, we have previously proposed a noise suppressor, under the assumption that the pitch of the damped oscillation is much higher than the pitch of human speech [12]. This method utilizes pitch detection, and estimates the spectral amplitude of the damped oscillation. Subtracting the estimated damped oscillation from the observed signal, we can obtain a speech mixed with impact noise which does not include damped oscillation. Then, applying the replacement technique to it, we achieve speech enhancement. As reported in [12], we can suppress an impact noise even if it is accompanied with a damped oscillation. However, the extracted speech has some degradation,

[1] For example, the ZP signal of an impulse or white signal is corresponding to a delta function.

since the noise suppression process is applied in all samples of the observed signal, even if the observed signal does not include noise.

In this paper, we also utilize the damped oscillation cancellation and the ZP signal replacement technique for suppressing impulsive noise signals, and derive a method that can improve the speech quality of the extracted speech signal. For improving the speech quality, we should restrict the use of the ZP signal replacement processing. In the proposed method, the ZP signal replacement processing is performed except when the observed signal consists of only voiced speech. The proposed method utilizes a ratio of the value at the origin to the peak value of the second period of the ZP signal. When the ratio is around one, we can determine that the observed signal is a voiced speech signal, because a perfect periodic signal gives the ratio identical to unity. In this case, we do not apply the ZP signal replacement technique to the observed signal. Otherwise, we use the ZP signal replacement technique. As a result, we can avoid redundant processes, and the speech deterioration reduces. This restriction is convenience since it can be simply implemented in the conventional ZP replacement scheme.

2. IMPACT NOISE SUPPRESSION BASED ON ZERO PHASE SIGNAL

Let $s(n)$ and $d(n)$ be a speech signal and an additive impulsive noise signal at time n. The observed signal $x(n)$ is given as

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

The observed signal $x(n)$ is transformed into frequency domain by segmentation and windowing with a window function $h(n)$. The DFT coefficient of the observed signal $x(n)$ at frame index l and frequency bin k is calculated as

$$X_l(k) \quad = \sum_{n=0}^{N-1} x(lQ+n)h(n)e^{-j2\pi nk/N}, \tag{2}$$

where N denotes the DFT frame size, and the window is shifted by Q samples to compute the next DFT. The observed spectrum $X_l(k)$ can be represented as $|X_l(k)|e^{j\angle X_l(k)}$, where $|\cdot|$ and $\angle\{\cdot\}$ denote spectral amplitude and phase, respectively. Since the DFT is a linear transformation, we have

$$X_l(k) \quad = \quad S_l(k) + D_l(k), \tag{3}$$

where $S_l(k)$ and $D_l(k)$ denote the DFT coefficients obtained from $s(n)$ and $d(n)$, respectively. The ZP signal is obtained by taking the IDFT of the pth power of spectral amplitude given as [11]

$$x_0(n) = \frac{1}{N}\sum_{k=0}^{N-1} |X_l(k)|^p e^{j2\pi nk/N}. \tag{4}$$

Obviously, the pth power of spectral amplitude $|X_l(k)|^p$ is reconstructed by the DFT of the ZP signal $x_0(n)$.

In the same manner as many stationary noise suppression scenario [1]–[7], we assume that the speech spectral phase is identical to the observed spectral phase. Then, we have

$$|X_l(k)| = |S_l(k)| + |D_l(k)|, \tag{5}$$

and hence its ZP signal with $p = 1$ is given as

$$x_0(n) = s_0(n) + d_0(n), \tag{6}$$

where $s_0(n)$ and $d_0(n)$ denote the ZP signals of $s(n)$ and $d(n)$, respectively.

Under the assumption that a voiced speech signal is periodic and an impact noise signal has almost a flat spectral amplitude, we have

$$x_0(n) = \begin{cases} s_0(n) + d_0(n), & 0 \le n \le L \\ s_0(n), & \text{otherwise} \end{cases}, \quad (n = 0, 1, \cdots, N/2) \tag{7}$$

where $x_0(N/2+m) = x_0(N/2-m)$ $(m = 1, 2, \cdots, N/2-1)$. The parameter L $(< N/2)$ is a natural number. The impact noise usually provides a small number of L. Hence, when the period of the speech signal is T, we have an estimated ZP speech signal as the following replacement technique [11].

$$\hat{s}_0(n) \approx \begin{cases} g(n,T)x_0(T+n), & 0 \le n \le L \\ x_0(n), & \text{otherwise} \end{cases}, \tag{8}$$

where $\hat{s}_0(n)$ is the estimated speech ZP signal, and $g(n,T)$ is a scaling function which compensates the decay of the ZP signal caused by segmenting and windowing the observed signal. The function $g(n,T)$ is given as the inverse function of the window function as shown in [11]. Taking the DFT of $\hat{s}_0(n)$ gives the estimated speech spectral amplitude $|\hat{S}_l(k)|$. The IDFT of the estimated speech spectral amplitude with the observed spectral phase, i.e., $|\hat{S}_l(k)|e^{j\angle X_l(k)}$, provides the estimated speech signal $\hat{s}(n)$ in time domain.

The noise suppression system is shown in Fig. 1, when $p = 1$. Here, we introduced a damped oscillation canceling method proposed in [12]. The damped oscillation cancellation is achieved by detecting its pitch frequency and suppressing their spectrum from the observed signal, under the assumption the pitch frequency of the damped oscillation is much higher than that of human. Then, $|X_l(k)|$ consists of a speech and an impact noise without damped oscillation. As reported in [12], the system shown in Fig. 1 is effective to suppress the impact noise with damped oscillation. However, the replacement method degrades the speech quality when the observed signal includes a speech signal.

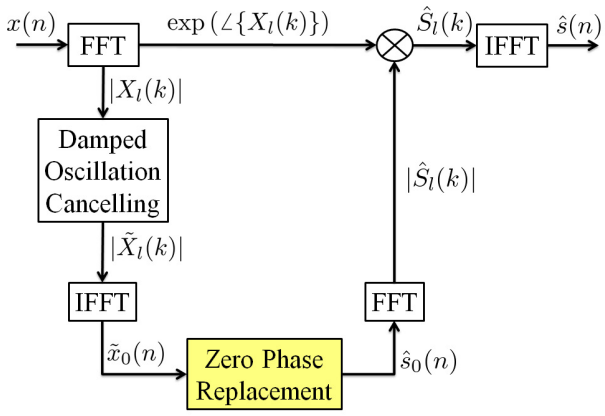
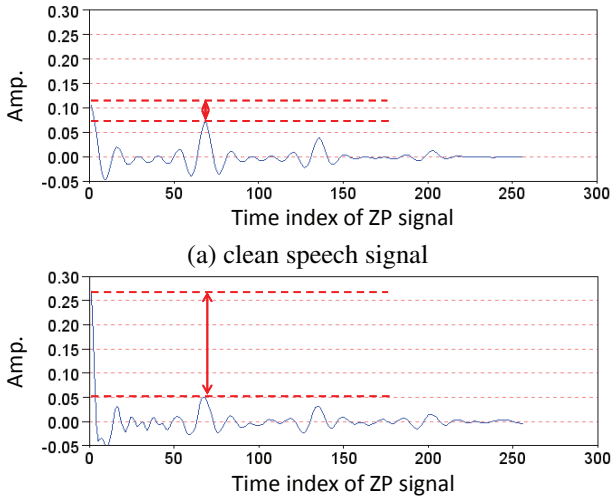


Fig. 1. Noise suppression system.



(a) clean speech signal

(b) speech mixied with clap noise.

Fig. 2. ZP signals of speech and noisy speech.

3. RESTRICTED IMPACT NOISE SUPPRESSOR IN ZP DOMAIN

In this section, we investigate a method to improve the speech quality of the estimated speech by the ZP signal replacement method. The ZP signal replacement method degrades the speech quality even if the observed signal includes only speech, because an actual speech signal is not a perfect periodic signal. The improvement of the speech quality is achieved by restricting the ZP signal replacement procedure so that it does not work when the observed signal includes only speech.

To judge whether the observed signal includes only speech or not, we utilize the ratio of the value at the origin to the peak value in at the second period of the ZP observed signal. When the analysis frame includes only speech signal, the ratio approaches to one. In this case, we skip the replacement procedure. Figure 2 illustrates the difference between the ZP signals of speech and noisy speech, where the noisy signal was a clean speech signal mixed with a clap noise. From the upper side in Fig. 2, we see that the difference between the first and second peak values is small in comparison to the result in the lower side. Here, we also see that both of the ZP signals gradually decreases. This effect was caused from the window function. To detect the present frame includes only speech or not, we introduce a threshold given as

$$\hat{s}_0(n) \approx \begin{cases} g(n,T)x_0(T+n), & \frac{x_0(0)}{sc(0|T)x_0(T)} < \alpha, \\ x_0(n), & \text{otherwise} \end{cases}, \quad (9)$$

where $n = 0, \cdots, L$, and α is the threshold. When the threshold α is around one, we can expect that the observed signal is voiced speech without impact noise.

We carried out some simulations of impulsive noise suppression to search an appropriate value of α. In the simulation, 100 male and 100 female speech signals were used as a clean speech signal from a Japanese speech database [14]. Impact noise signals were "cup", "bottle", "china" from a database [15], an artificially generated impulse signal, and a clap noise recorded by us in a seminar room. The sampling frequency was 16kHz, the DFT size and the frame shift size were $N = 512$ and $Q = 256$, respectively, and Hanning window was used for windowing. We set the number of replacement in the ZP signal as $L = 20$, and $p = 1.3$ [12]. The speech period was estimated as the sample index which supports the maximum value of the observed ZP signal except around the origin. The noise suppression capability was evaluated by using the SNR, SD (Spectral Distance), and ISD (Itakura-Saito Distance) defined as [16]

$$\text{SNR} = 10\log_{10}\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1}\{s(n)-\hat{s}(n)\}^2}, \quad (10)$$

$$\text{SD} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{N}\sum_{k=0}^{N-1}\left(|S_m(k)| - |\hat{S}_m(k)|\right)^2, \quad (11)$$

$$\text{ISD} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{N}\sqrt{\sum_{k=0}^{N-1}\left(\log\frac{|S_m(k)|}{|\hat{S}_m(k)|} + \frac{|\hat{S}_m(k)|}{|S_m(k)|} - 1\right)^2}, \quad (12)$$

where M is the number of samples, and $\hat{S}_m(k)$ is the estimated speech spectral amplitude at mth frame. When the noise reduction is effectively performed, the output SNR becomes larger, and the output SD and the output ISD become smaller. We changed α from 0 to 50 with 1 gap. In all the simulations, the SNR of the observed signal was set to 0dB. The simulation results are summarized in Fig. 3, where the horizontal axis denotes α and the vertical axis denotes the averaged values of each evaluation criterion. We see from these results that values around 13 provided the maximum SNR.

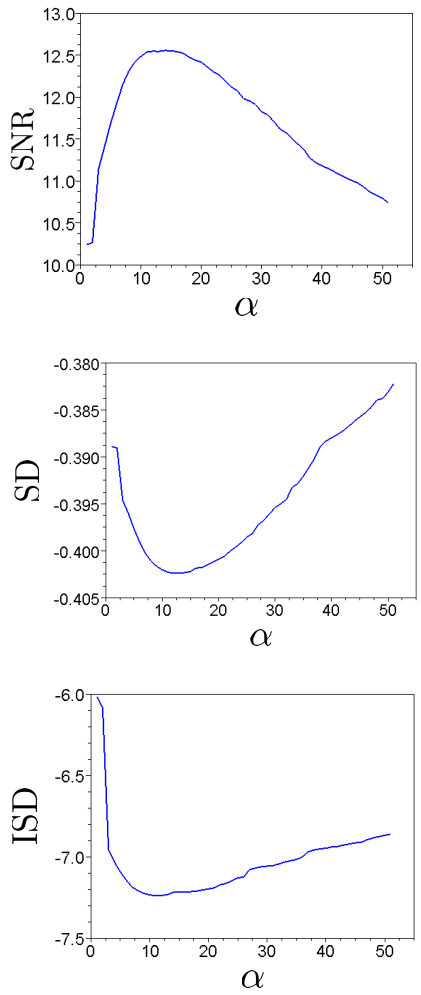


Fig. 3. Evaluation results for α. Top panel shows SNR, middle panel shows SD, and bottom panel shows ISD. These are averaged values for various combinations of speech and noise.

Hence, we set $\alpha = 13$ in the following simulation. Although an investigation of the relation between p and α may give more good results, we put it on our future work.

Figure 4 shows an example of the impact noise detection results with $\alpha = 13$. Here, the top panel shows an observed signal which is a speech mixed with a clap noise. The middle panel shows the impulsive noise detection results, where 'Noise' denotes the detected noise segments, and 'No Noise' denotes the speech or silent segments. The bottom panel shows the extracted speech signal. We see from Fig. 4 that the most of clap noise signals were detected and removed by the proposed method. From the results around the frame index 50, we see that noise segment was not detected, but the clap noise was eliminated. This fortunate result might be caused from an effect of the pre-processing, i.e., the damped

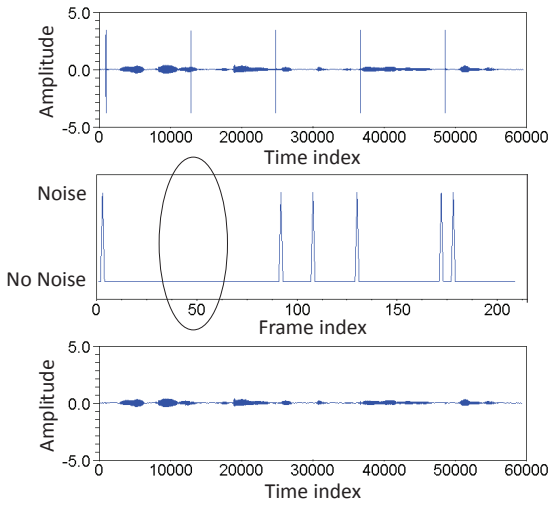


Fig. 4. Noise suppression results. Top panel shows an observed signal, middle panel shows noise detection results, and bottom panel shows the extracted speech.

oscillation cancellation [12] often removes not only damped oscillation but also an impact part.

4. EVALUATION FOR NOISE SUPPRESSION CAPABILITY

In this section, we compared the impulsive noise suppression capability of the proposed method with one of the conventional method [12]. The conditions of both methods were set to the same in the previous simulation except of α. The parameter α was set to 13 in the proposed method.

Table 1 shows averaged values of 200 speech signals obtained for each impact noise signal. We see from Table 1 that the proposed method is superior to the conventional method for impulse and clap noise signals. On the other hand, for "cup", "bottle", and "china", the results of the proposed method were almost the same to the conventional method. These noise signals were accompanied with damped oscillation. We expect from these results that the effectiveness of the damped oscillation cancellation is much larger than the effectiveness of the impact noise reduction. In other words, the proposed method effectively improves the noise reduction capability for impact noise without damped oscillation, and it holds the capability for the impact noise with damped oscillation. The SNR results of the conventional and proposed methods are 11.9dB and 15.2dB, respectively for clap noise which does not have noticeable damped oscillation. We see from these results that the proposed method improved more than 3dB in comparison to the conventional method.

Table 1. Evaluation results for impact noise ($\alpha = 13$).

	SNR		S. Dist.		I-S. Dist.	
	Conv.[12]	Prop.	Conv.[12]	Prop.	Conv.[12]	Prop.
Impulse	7.0	11.3	-0.34	-0.40	-10.84	-12.60
Clap	11.9	15.2	-0.42	-0.45	-5.65	-5.91
Cup	9.6	9.5	-0.36	-0.36	-8.61	-8.58
Bottle	12.7	12.5	-0.38	-0.38	-3.81	-3.79
China	11.3	11.2	-0.41	-0.41	-7.16	-7.16

5. CONCLUSION

In this paper, we have utilized the ZP signal replacement technique for impact noise suppression, and proposed a method to improve the speech quality. In the proposed method, the ZP signal replacement processing was performed except when the observed signal consists of only voiced speech. The proposed method has utilized a ratio of the value at the origin to the maximum value of the second period in the observed ZP signal. When the ratio is around one, we can judge the observed signal is voiced speech. In this case, we skip the ZP signal replacement method. Otherwise, we applied the ZP signal replacement method to the observed signal. As a result, we could avoid redundant processes, and the speech quality improved. Simulation result has shown that the proposed method improved the SNR of 15.2 dB, while the conventional method improved the SNR of 11.9 dB, when the input SNR=0dB for clap noise.

6. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.

[2] M. Muneyasu and A. Taguchi, Nonlinear digital signal processing, Asakura Publishing Company, Tokyo, 1999.

[3] Y. Ephraim and D. Malah, " Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, " IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109–1121, Dec. 1984.

[4] P.J. Wolfe and S.J. Godsill, " Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, " EURASIP Journal on Applied Signal Processing, vol.10, pp.1043–1051, Oct. 2003.

[5] T. Lotter and P. Vary, " Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model, " EURASIP Journal on Applied Signal Processing, vol.7, pp.1110–1126, July 2005.

[6] P. Vary and R. Martin, Digital Speech Transmission, John Wiley & Sons, Ltd, UK, 2006.

[7] W. Thanhikam, A. Kawamura, Y. Iiguni, " Speech Enhancement Based on Real-Speech PDF in Various Narrow SNR Intervals " IEICE Trans. Fundamentals, vol.E95-A, no.3, pp.623–630, Mar. 2012.

[8] A. Sugiyama and R. Miyahara, "Phase Randomization - A New Paradigm for Single-Channel Signal Enhancement," Proc. of ICASSP 2013, pp.7487–7491, May. 2013.

[9] R. Miyahara and A. Sugiyama, "An Auto-Focusing Noise Suppressor for Cellphone Movies Based on Peak Preservation and Phase Randomization," Proc. of ICASSP 2013, pp.2785–2789, May. 2013.

[10] R. Talman, I. Cohen, and S. Gannot, "Transient Noise Reduction Using Nonlocal Diffusion Filters," IEEE Trans. Audio, Speech, and Language Processing, vol.19, no.6, pp.1584–1599, Aug. 2011.

[11] W. Thanhikam, A. Kawamura,and Y. Iiguni, "Stationary and Non-stationary Wide-Band Noise Reduction Using Zero Phase Signal " IEICE Trans. Fundamentals, vol.E95-A, no.5, pp.843–852, May. 2012.

[12] S. Kohmura, A. Kawamura, and Y. Iiguni, "An Efficient Zero Phase Noise Reduction Method for Impact Noise with Damped Oscillation" Proc. of ICASSP 2013, pp.892–895, May. 2013.

[13] S. Itabashi, Sound Engneering, Morikita Publishing Company, Tokyo, 2005.

[14] "ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)," Speech Resources Consortium, http://research.nii.ac.jp/src/en/JNAS.html.

[15] "RWCP Sound Scene Database in Real Acoustical Environments (RWCP-SSD)," Speech Resources Consortium, http://research.nii.ac.jp/src/en/RWCP-SSD.html.

[16] S. Furui, Digital speech processing, Tokai University Press, Tokyo, 1985.