AN AUTO-FOCUSING NOISE SUPPRESSOR FOR CELLPHONE MOVIES BASED ON PEAK PRESERVATION AND PHASE RANDOMIZATION

Ryoji Miyahara† *and Akihiko Sugiyama*

Information and Media Processing Laboratories
NEC Corporation
† Internet Terminal Division, NEC Engineering
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan
aks@ak.jp.nec.com

ABSTRACT

This paper proposes an auto-focusing noise suppressor (AF-NS) for cellphone movies. For relatively small clicks compared to the target signal, it consists of magnitude suppression and complementary phase modification. The input signal is analyzed in the frequency domain to detect and preserve important spectral components including peaks and their vicinities. All other components are suppressed to the environmental signal level that is estimated during absence of the important components. Residual spikes by auto-focusing noise are successfully suppressed by phase randomization. Subjective evaluation results demonstrate that the proposed AF-NS achieves scores of 1.6 and 1.8 in the 7-grade modified CMOS with statistically significant differences compared to the input noisy signal and an AF-NS with no phase randomization.

Index Terms— Cellular phone, Movie, Auto-focusing noise, Noise suppressor, Phase randomization

1. INTRODUCTION

With the dissemination of cellular phones, it is becoming more and more common to use them for video recording. In video recording, spike-like auto-focusing (AF) noise as shown in Fig. 1 generated by mechanical components and captured by microphones often contaminates the target signal. This is more serious with inexpensive piezoelectric actuators. Conventional cellular phones disable auto-focusing function in the movie mode when its noise is intolerable. Drawbacks are out-of-focus images of quickly moving objects often encountered in watching sports and athletic meets. In addition, high definition (HD) video formats are available in most of the high-end cellphones. In such formats, only a slight defocus is visible and gives an impression of serious degradation. Therefore, to record HD video of quickly moving objects with suitable high audio quality, AF noise suppression is a must function.

For the purpose of communication, noise suppressors have been extensively used for suppressing the undesirable noise and enhancing the target speech [1]-[9]. However, AF noise is a series of clicks in comparison with more general, slowly changing environmental signal in communication. It means that presence of the AF noise is discontinuous and needs detection so that suppression can be applied only upon detection. Because clicks happen suddenly with no advance information and last for a very short time, its detection is challenging. Moreover, conventional click suppression algorithms [10] -[15] keep the phase of the input noisy speech. Even when this phase is combined with the true magnitude of the target signal, which is superior to any conventional magnitude estimation, there are residual clicks as shown in the following section.

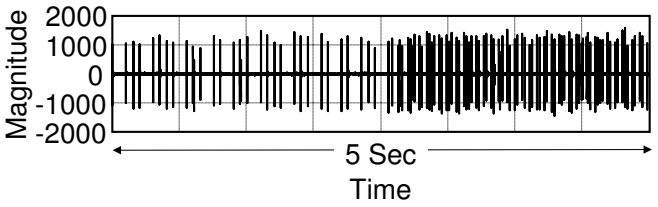

Fig. 1. An example of auto-focusing noise. This will be mixed with the target signal.

This paper proposes an auto-focusing noise suppressor (AF-NS) for cellphone movies. Instead of click detection and suppression, it performs suppression of the input signal components to the environmental signal level except important spectral components including peak frequencies and their vicinities. The following section discusses why conventional noise suppressors are not applicable to AF noise. Section 3 presents an AF-NS algorithm with phase randomization. In Section 4, objective and subjective evaluation results are demonstrated to support good performance.

2. CONVENTIONAL NOISE SUPPRESSORS

Generally in conventional noise suppressors, averaging is employed in noise estimation for higer accuracy [1]-[9]. It does not reflect values by clicks which exists for a short duration. Minimum statistics [17] and its variants relies on a minimum value that does not respond to local maxima by clicks. Therefore, conventionsl noise suppressors designed for communications is not suitable for clicks by AF noise.

Detection of clicks, which is not guaranteed to be perfect, is another issue. Some literatures [18]-[23] assume large clicks comaprable in magnitude to the target signal. Detection becomes more difficult when clicks with much smaller magnitudes than the target signal are buried. If there is even a single failure of detection, that click is easily audible, leading to lower subjective quality. Moreover, even if 100% detection is acieved and the clean magnitude is combined, an example in Fig. 2 indicates that there are residual clicks which are audible. It should be noted that this is an ultimate ideal case and no other conventional transient noise suppression outperforms this result. Therefore, a new approach is needed that does not rely on detection but may utilize the relatively small magnitude

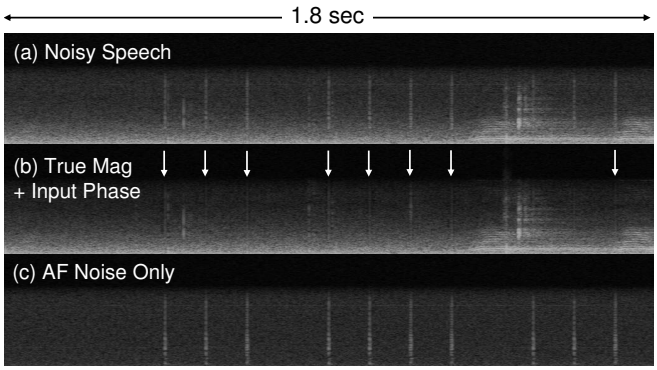

Fig. 2. Spectrogram of (a) Noisy speech, (b) True magnitude with the noisy phase, (c) AF noise only.

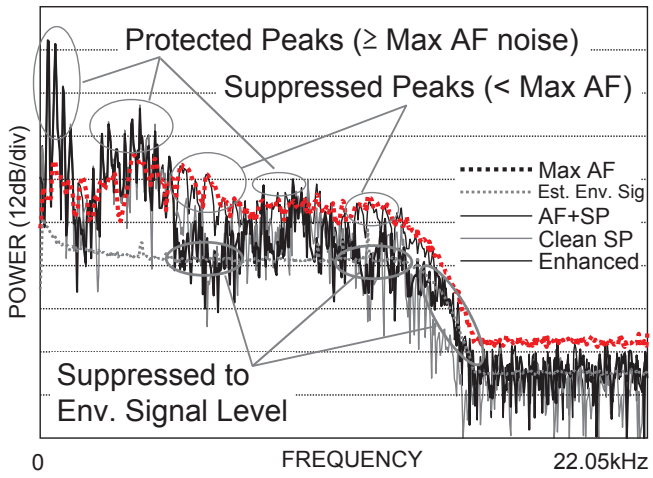

Fig. 3. Overview of magnitude suppression by spectrum.

of clicks to the target signal. Fig. 2 suggests that the phase is somehow modified for good AF-noise suppression.

3. PROPOSED AUTO-FOCUSING NOISE SUPPRESSOR (AF-NS)

3.1. Underlying Concept of Proposed AF-NS

The proposed AF-NS takes a totally diffferent approach from conventional noise suppressors. For the imperfect detection problem, it has no detection in the algorithm and tries to suppress clicks as a part of magnitude spectrum. Figure 3 depicts an overview of magnitude suppression by spectrum. Magnitude peaks and their vicinities larger than the maximum of the AF noise, which can be obtained in advance, are preserved. Smaller magnitude components and non-peaks are suppressed to an estimated environmental signal level. For the phase modification issue, it applies phase randomization for small-magnitude frequency bins whose magnitude have been suppressed. Randomization is motivated by removal of any undesirable characteristics that make residual clicks audible. It is effective for further reducuction of the residual noise, that is perceived behind the target signal due to the same phase characteristics as that of the noisy signal. Other frequency componentss use the noisy phase as conventional algorithms because contribution of the noise in the phase is indominant in those frequency bins.

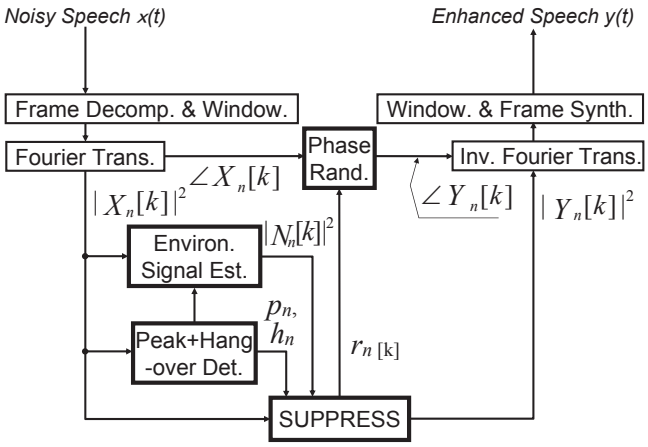

Fig. 4. Blockdiagram of the proposed AF noise suppressor

Figure 4 illustrates a blockdiagram of the proposed auto-focusing noise suppressor. The input noisy signal is decomposed into frames of L samples and applied an windowing function before it is converted to a frequency-domain signal by Fourier transform. Magnitude of the frequency-domain signal is provided to Environmental Signal Estimation (Environ. Signal Est.), Peak and Hangover Detection (Pk+Ho Det.), and Suppression (SUPPRESS). Phase goes to Phase Randomization (Phase Rand.). Environmental signal is estimated in frequency bins that are not detected as peaks. Peaks are detected by the center power and the width as is explained later in details. Other frequency bins are considered as noise and their magnitudes are suppressed to the estimated environmental-signal level. Hangover is detected in Peak+Hangover Det. and treated separately from peaks. An overview of magnitude suppression is illustrated in Fig. 5.

Peaks are detected as follows:

1. Find peak frequency bins for $k = 1 \sim N/2-2$ whose power $|X_n[k]|^2$ is higher than neighboring ones where n and k represent the block index and the frequency index. It is to find bins that satisfy

$$|X_n[k]|^2 > |X_n[k+1]|^2 \text{ and } |X_n[k]|^2 > |X_n[k-1]|^2. \quad (1)$$

2. For each k at a peak, search for a power more than σ_L or σ_H smaller than the peak power, within M_L or M_H adjacent bins on the left- and the right-hand side. Frequency bins between those closest to the peak at k on both sides will be considered as peak bins. It is to examine, for $l = 1 \sim M_L$ and $j = 1 \sim M_H$, if the following inequalities are satisfied.

$$|X_n[k]|^2 > |X_n[k-l]|^2 + \sigma_L, \quad (2)$$
$$|X_n[k]|^2 > |X_n[k+j]|^2 + \sigma_H. \quad (3)$$

   Frequency bins between those with the minimum l in (2) and the minimum j in (3) are determined as peak bins and a peak-position index $p_n[m]$ is set to 1 for $k - \min l \le m \le k + \min j$. It means that each peak has a certain bandwidth and is not an isolated freqnency bin.

3. All peaks are labeled with a peak flag $p_n[k] = 1$ after being evaluated if they are above the maximum value of the AF noise which is obtaied in advance. Otherwise, they are labeled as nonpeaks with $p_n[k] = 0$.

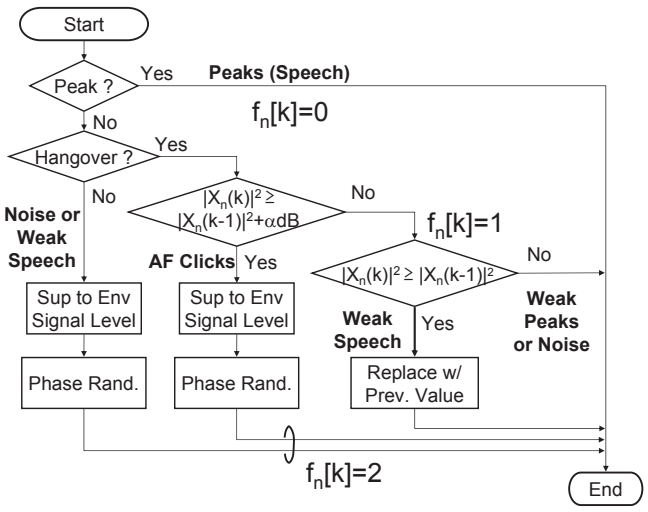

Fig. 5. Overview of magnitude suppression.

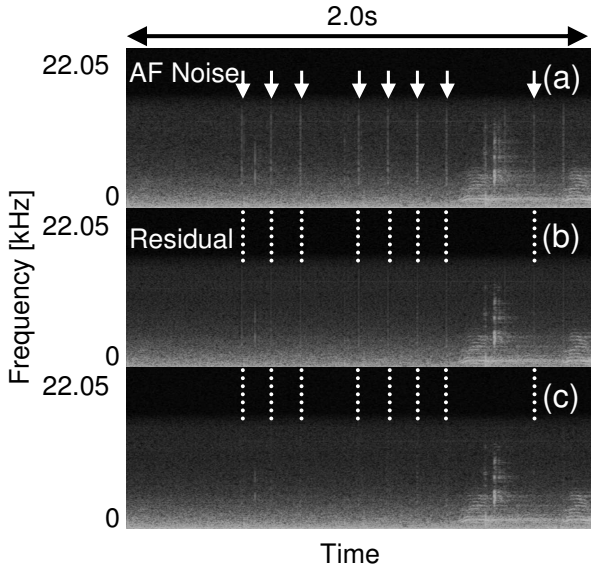

Fig. 6. Input (a) and output signals with (b) weak [24] and (c) full (proposed) phase randomization.

Hangover is determined when there is any peak in a past period to fill gaps in a speech section. A hangover index $h_n[k]$ is set as

$$h_n[k] = \begin{cases} 1 & \sum_{n-Q+1}^{n} p_n[k] > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where an integer Q is a hangover period.

An estimate of the environmental signal $\hat{\lambda}_n^2[k]$ is updated based on a first-order leaky integration (recursive filter) with a leaky factor γ in non-peak frequency bins.

For a simple description, a suppression flag $f_n[k]$ that indicates detailed suppression is introduced. It has three values, 0, 1, and 2, each representing "preserve," "reverb," and "suppress." For peak bins and non-peak-non-hangover bins, $f_n[k]$ is defined by

$$f_n[k] = \begin{cases} 0 & p_n[k] = 1 \\ 2 & p_n[k] + h_n[k] = 0 \end{cases}. \tag{5}$$

It performs magnitude discrimination between bins to be preserved and those to be suppressed. In non-peak-hangover bins, assuming short-time stationarity, they are processed as

$$f_n[k] = \begin{cases} 2 & |X_n[k]|^2 \geq |X_{n-1}[k]|^2 + \alpha\text{dB} \\ 0 & |X_n[k]|^2 < |X_{n-1}[k]|^2 \\ 1 & \text{otherwise} \end{cases}. \tag{6}$$

Eq. (6) is to identify clicks for suppression by sharp increase of $|Xn[k]|^2$. Decrease and moderate increase are to be reverbed and preserved, respectively.

Based on the suppression flag $f_n[k]$, magnitude of the noise suppressed signal $|Y_n[k]|^2$ is obtained by

$$|Y_n[k]|^2 = \begin{cases} |X_n[k]|^2, r_n[k] = 0 & f_n[k] = 0 \\ |X_{n-1}[k]|^2, r_n[k] = 0 & f_n[k] = 1 \\ \hat{\lambda}_n^2[k], r_n[k] = 1 & f_n[k] = 2 \end{cases}. \tag{7}$$

For $f_n[k] = 2$, a randomization index $r_n[k]$ is set to 1 and the phase is randomized. Otherwise, $r_n[k]$ is set to 0 to preserve the noisy-speech phase.

The input noisy signal phase $\angle X_n[k]$ is randomized based on $r_n[k]$ in Phase Rand. to obtain the enhanced signal phase $\angle Y_n[k]$ as

$$\angle Y_n[k] = \angle X_n[k] + r_n[k] \cdot \phi_n[k], \tag{8}$$

where $\phi_n[k]$ is a random value between $\pm\pi$. Weak randomization [24] with $\phi_n[k]$ between $\pm\pi/4$ turned out to be insufficient. $|Y_n[k]|^2$ and $\angle Y_n[k]$ are used to reconstruct the enhanced signal at the output.

It should be noted that the proposed AF noise suppressor is designed in the forward/inverse Fourier transform framework. For different kinds of noise such as environmental signal, the Fourier transform pair can be shared and the suppression part can be cascaded with other existing noise suppressors.

4. EVALUATION

4.1. Objective Evaluation

The AF noise was recorded in a real cellphone with a sampling frequency of 44.1 kHz and mixed with different environmental signals. The frame size L and the FFT size N were set to 512 and 1024, respectively. Other parameters, optimized for several different commercial cellphone handsets, are summarized in Tab. 1.

Figure 6 illustrates the output spectrogram of the proposed AF noise suppressor (AF-NS). Subfigures (a) through (c) represent the input noisy signal with AF click noise, the AF-NS output with weak phase randomization [24], and the AF-NS output with full phase randomization[1]. A bright dot represents a strong signal component. Bright vertical lines with downward arrows in (a) highlights positions of AF-noise clicks. It is observed in (b) that weak phase randomization [24] does not achieve sufficient suppression with visible and audible residual clicks. On the other hand, full phase randomization proposed in this paper successfully suppresses the AF noise as in (c).

Shown in Fig. 7 is effect of phase randomization for an AF noise. In this example, there is no speech in the noisy signal but only environmental signal (Noise 1) and AF noise. Fig. 7 (a) represents the noisy signal that contained environmental signal and AF noise. At the center of the ordinate, there is a white trajectory with many spikes. This is the AF noise to be suppressed. At positions of vertical

[1] with additive phase between $\pm\pi/4$ (weak) and $\pm\pi$ (full).

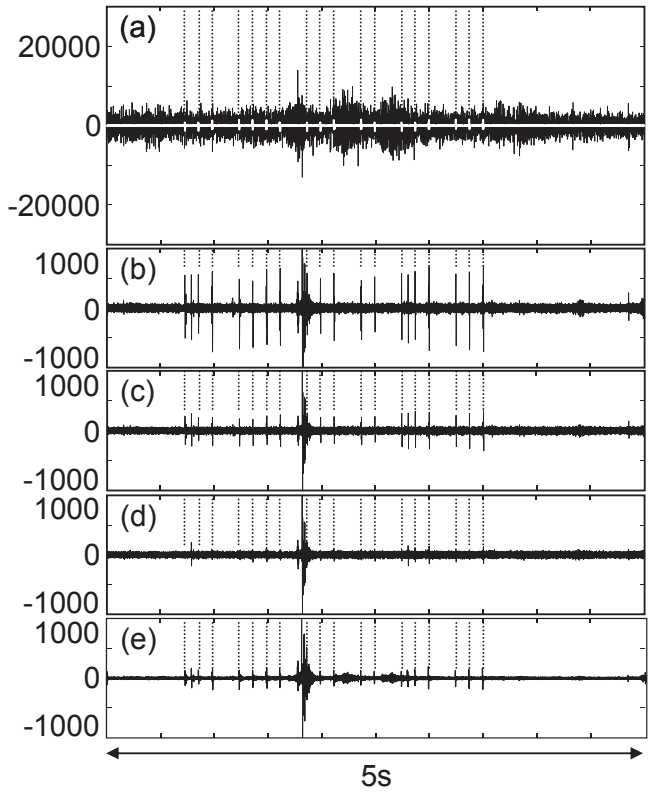

Fig. 7. Effect of phase randomization. (a) Noisy signal (speech+env. noise+zoom. noise), (b) Noisy signal after bandlimitation to $6-15$ kHz, (c) Enhanced signal w/o phase randomization after bandlimitation, (d) Enhanced signal w/ phase randomization after bandlimitation, (e) Enhanced signal by a communication NS [8] .

dotted lines that coincides with AF-noise spikes in white, AF noise has clicks. However, they are buried in the environmental signal and invisible, but audible. Those clicks become more visible when the signal is bandlimited to a frequency range from 6 to 15 kHz as in (b). Please note that there are some spikes that do not correspond to any vertical dotted line. These spikes do not come from AF noise but the environmental signal. All spikes are gone in (d) due to phase randomization in contrast to (c) without phase randomization. This difference is subjectively audible as is demonstrated in subjective evaluation. It is demonstrated in (e) that a communication NS [8] is not useful for AF-noise suppression as was described in Section 2.

4.2. Subjective Evaluation

The performance of the proposed AF-NS was evaluated by 7-grade modified CCR (Comparison Category Rating)[2][25] in comparison with the noisy signal. Conventional communication NSs were not included because it is not effective at all for AF noise as shown in Fig. 7. Male and female speech signals sampled at 44.1kHz were mixed with four different environmental signals in Tab. 2, which were also evaluated without speech. The speech models narration and the noise describes the environment in a typical movie scenario. The total number of evaluated signals was 72 including two reversed presentation orders with 14 subjects. An average signal-to-noise ratio (SNR) in noise sections was -2.0 dB. Other parameters were equal to those in the objective evaluations.

[2]The modified CCR method uses processed reference samples but without noise suppression whereas the standard CCR method uses unprocessed reference samples.

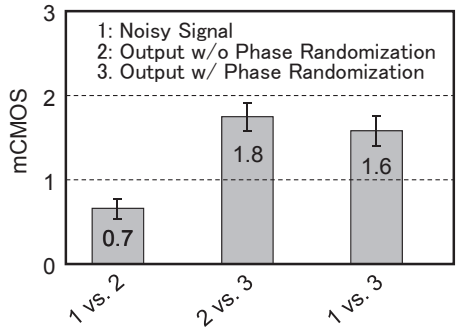

Fig. 8. Subjective evaluation result. 1: Input, 2 and 3: Output without and with phase randomization.

Table 1. Parameter values.

M_L	5	σ_L	12dB	Q	16	γ	0.98
M_H	5	σ_H	12dB	α	3dB		

Table 2. Speech and environmental signal used for subjective evaluations.

Speech	Male and female speech
Env Sig 1	Street noise with crow caws
Env Sig 2	Street noise with bike-brake creaks and car honks
Env Sig 3	Office noise with telephone rings
Env Sig 4	Street noise with a car back-up alarm
SNR	$-6.7 \leq SNR\,(or\ TNR)^3 \leq -1.1$

Figure 8 depicts the results with and without phase randomization. The left and the right vertical bars represent the AF-NS output quality compared to the noisy signal. The effect of power compensation is demonstrated by the center one, which compares the AF-NS outputs with and without power compensation. A higher score means a higher quality of the AF-NS output signal than the noisy speech. In the case of the center bar, a higher score demonstrates that power compensation is more effective. Two horizontal lines connected by a vertical line represent the 95% confidence level.

Because the lower limit of the 95% confidence interval lies in the positive region, the proposed AF-NS with or without power compensation has statistically higher quality with an average score of 1.7 or 0.7 than the noisy signal as depicted in the left or the right bar. Similarly, the AF-NS output with power compensation is better in subjective quality with an average score of 1.0 than the AF-NS output without it. This is confirmed by the center bar with its lower limit of the 95% confidence interval in the positive region in Fig. 8.

5. CONCLUSION

An auto-focusing noise suppressor (AF-NS) for cellphone movies has been proposed. A simple algorithm has been developed with peak preservation and suppression to the environmental-signal level. To reduce the residual AF-noise clicks, phase randomization have been introduced. Subjective evaluation results have demonstrated that the proposed AF-NS achieves score of 1.6 and 1.8 in the 7-grade CMOS compared to the input noisy signal and an AF-NS with no phase randomization.

[3]S is speech plus environmental signal in Tab. 2 and N is the AF noise. Mixtures without speech are included.

6. REFERENCES

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

[2] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. ICASSP'79, pp. 208–211, Apr. 1979.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of speech," Proc. of IEEE, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.

[4] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft- decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, no. 2, pp.137–145, Apr. 1980.

[5] 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.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[7] T. V. Ramabadran, J. P. Ashley and M. J. McLaughlin, "Background noise suppression for speech enhancement and coding," IEEE Workshop on Speech Coding and Tel., pp.43–44, Sep. 1997.

[8] M. Kato, A. Sugiyama and M. Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA," Proc. IWAENC2001, pp.183–186, Sep. 2001.

[9] J. Benesty, S. Makino, and J. Chen, Eds., "Speech Enhancement," Springer, Berlin, Mar. 2005.

[10] A. Subramanya, M. L. Seltzer, A. Acero, "Automatic removal of typed keystrokes from speech signals," Sig. Proc. Letters, Vol. 14, No. 5, pp.363–366, May 2007.

[11] R. C. Nongpiur, "Impulse noise removal in speech using wavelets," Proc. ICASSP2008, pp.1593–1596, Mar. 2008.

[12] R. Talmon, I. Cohen, S. Gannot, "Speech enhancement in transient noise environment using diffusion filtering," Proc. ICASSP2010, 4782–4785, Mar. 2010.

[13] R. Talmon, I. Cohen, S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," Proc. ICASSP2011, pp.5084–5087, May 2011.

[14] R. Talmon, I. Cohen, S. Gannot, "Transient noise reduction useing nonlocal diffusion filters," Trans. ASLP, Vol. 19, No. 6, pp.1584–1599, Jun. 2011.

[15] H. Chen, C. Bao, F. Deng, D. Zhang, M. Jia, "A MDCT-based click noise reduction method for MPEG-4 AAC codec," Proc. WCSP2011, pp. 1–5, Nov. 2011.

[16] T. Gulzow, "Spectral-subtraction-based speech enhancement using a new estimation technique for non-stationary noise," Proc. IWAENC'99, pp. 76–79, Sep. 1999.

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

[18] A. Kundu, S. K. Mitra, "A computationally efficient approach to the removal of impulse noise from digitized speech," Trans. ASSP, Vol. 35, No. 4, pp.571–574, Apr. 1987.

[19] S. J. Godsill, P. J. W. Rayner, "A Bayesian approach to the restoration of degraded audio signals," Trans. SAP, Vol. 3, No. 4, pp.267–278, Apr. 1995.

[20] S. J. Godsill, C. H. Tan, "Removal of low frequency transient noise from old recordings using model-based signal separation techniques," Proc. WASPAA97, CD-ROM, Oct. 1997.

[21] R. Rajagopalan, B. Subramanian, "Removal of impulse noise from audio and speech signals," Proc. SCS2003, pp.161–163, Jul. 2003.

[22] A. Abramson, I. Cohen, "Enhancement of Speech Signals Under Multiple Hypotheses using an Indicator for Transient Noise Presence," Proc. ICASSP2007, pp.553–556, Apr. 2007.

[23] N. Kyoya, K. Arakawa, "A method for impact noise reduction from speech using a stationary-nonstationary separating filter," Proc. ISCIT2009, pp.33–37, Sep. 2009.

[24] A. Sugiyama, "Single-Channel Impact-Noise Suppression with No Auxiliary Information for Its Detection,h Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics (WASPAA), pp.127-130, Oct. 2007.

[25] "Minimum performance requirements for noise suppresser application to the AMR speech encoder," 3GPP TS 06.77 V8.1.1, Apr. 2001.