GAIN RELAXATION: A USEFUL TECHNIQUE FOR SIGNAL ENHANCEMENT WITH AN UNAWARE LOCAL NOISE SOURCE TARGETED AT SPEECH RECOGNITION

Ryoji Miyahara† *and Akihiko Sugiyama*

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

ABSTRACT

This paper proposes gain relaxation in signal enhancement designed for speech recognition with an unaware local noise source. An attention is drawn to a new performance degradation problem in signal enhancement combined with automatic speech recognition (ASR), which is encountered in real products with an unaware noise source. Gain relaxation, as a solution, selectively applies softer enhancement of a target signal to eliminate potential degradation in speech recognition caused by small undesirable distortion in the target signal components. Evaluation of directional interference suppression with signals recorded by a commercial PC (personal computer) demonstrates that signal enhancement over the input is achieved without sacrificing the performance for clean speech.

Index Terms— Signal enhancement, Speech recognition, Noise suppressor, Beamformer, Phase difference

1. INTRODUCTION

Audio signals are captured by microphones placed in various different environment. A target signal which user tries to capture is often contaminated by different types of noise and interference. Stationary noise at a relatively high signal-to-noise ratio (SNR) can be well suppressed by a noise suppressor with a single microphone [1]-[5]. A lower SNR and/or nonstationary noise require more complex dual microphone solutions [6]-[19]. For point sources of noise or interference, acoustic beamformers, also known as microphone arrays (MAs), are more effective [20]–[24]. Phase-based time-frequency (T-F) masking [26]-[29] is also a simple but effective technique for point sources.

These signal enhancement techniques are effective for telecommunication or recording purposes as well as preprocessing for automatic speech recognition (ASR). In case of telecommunication and recording applications, the performance of the employed signal enhancement is mostly evaluated subjectively by users. For ASR, however, the performance is expressed by a successful recognition rate or an error rate. Because such rates are objective measures, the result is more definite than a subjective measure. This fact sometimes causes a problem in real products.

Products with signal enhancement capability are often evaluated in ASR scenarios. Recognition/error rates are used as a measure to decide if the product is sufficiently good or not. In general, if a signal enhancement technique is applied to the raw signal collected in a noisy environment, users expect better performance than the raw signal. It is more true in a high SNR when nobody expects even a small degradation. However, it is not always guaranteed. An example is a personal computer (PC) equipped with a cooling fan. The fan noise interferes the input signal although the resulting SNR is relatively high. It is an unconscious noise for the user, because the fan noise has such a small power and the distance between the fan and the user ear is much longer than that between the fan and the microphone. If there is any degradation in such a quiet environment, it is easy to notice. A similar environment can be found in voice recorders used with a projector with a cooling fan.

This kind of performance degradation has not been dealt with in literatures. Those degradations in ASR performance is usually found in some types of nonlinear signal enhancement techniques such as noise suppressors (NSs) and phase-based T-F masking where only the magnitude of the input signal is manipulated for signal enhancement. Although these techniques are simple but effective, degradation of ASR performance in a high SNR environment often gives fatal impression to users or those who evaluate the performance for business purposes.

This paper proposes gain relaxation in signal enhancement designed for speech recognition with an unaware local noise source. The following section discusses the performance degradation problem in signal enhancement for ASR with an example. Section 3 presents gain relaxation as a solution to the aforementioned problem. In Section 4, evaluation results with phase-based T-F masking in a commercial PC scenario are presented to show that word error rates are reduced without sacrificing the performance for the clean speech.

2. DEGRADATION IN SIGNAL ENHANCEMENT FOR ASR

Let us first look at Fig. 1 as an example of performance degradation in signal enhancement when it is combined with ASR. NoSE and dNS stand for no signal enhancement and directional NS, respectively. A directional NS [28] applies a predetermined directional gain in a frequency domain depending on the input signal DOA represented by the phase difference between signals from adjacent microphones.

Command error rates (CERs) do not change by application of dNS. However, a word error rate (WER) for clean speech is increased by 2% by dNS. This may be caused by inappropriate suppression of small power components of the target signal. A small error may be introduced in the phase difference by various imperfections such as the target DOA, microphone-gain mismatch, or their combinations. A slightly changed phase difference may fall in out-of-passband, leading to wrong and possibly fatal suppression of a

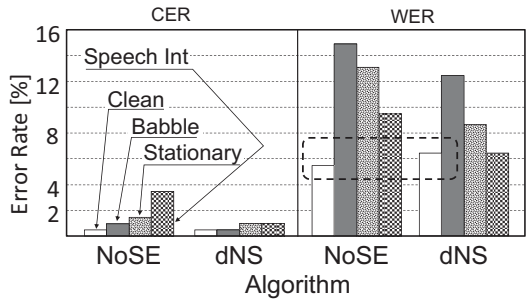

Fig. 1. CER and WER with (dNS) and without (NoSE) a directional NS (dNS) [29]. Word error rate for clean speech is degraded by directional NS. Evaluation conditions are equal to those in Sec. 4.

target component. Influence is more serious for high SNRs which otherwise needs little suppression. This degradation is investigated using Fig. 2.

Figure 2 is an explanatory diagram for relationship of an error rate with no enhancement (P), improvement by enhancement (Q), degradation by distortion (R), and an error rate with enhancement. The error rate with no enhancement increases as the SNR degrades. A high SNR provides a small error and a low SNR results in a large error rate. The SNR vs. error rate may not be proportional to the SNR. It has different characteristics for different ASR engines. Let us assume the error rate with no enhancement can be represented by a straight line as a gray straight line in Fig. 2. When a signal enhancement technique is employed, improvement in the ASR error rate looks like a dashed line Q in Fig. 2. It is because no or negligible error-rate reduction is obtained in a high SNR environment, while the improvement is saturated as the SNR is decreased. Therefore, the dashed line Q does not change its value with the SNR in high and low SNR environments. Please note that Q mostly takes negative values because its contribution reduces the error rate. On the other hand, there is an undesirable increase in the error rate by distortion in the enhanced signal. The increase or degradation in the error rate starts from a high SNR and is almost the same until a low SNR. For very low SNR values, this increase/degradation is progressively increased as the SNR. This undesirable increase is represented by a dashed line R. Because it is an increase, it takes positive values. The resulting error rate with signal enhancement is an integration of P, Q, and R and is represented by a solid line S. Equivalently, S=P+Q+R.

For easy comparison of the error rate P without signal enhancement and the error rate S with signal enhancement, P and S are copied to Fig. 3. The solid line has a hump, as shown by gray circle, in a high SNR region where the error rate is higher than the original value of P with no enhancement. Therefore, the error rate is degraded for some SNR values. The position T of the hump varies depending on the ASR engine and is not predictable.

3. SOLUTION: GAIN RELAXATION

A solution to this degradation in ASR error rate is to often or disable signal enhancement in the hump region in Fig. 3 by applying a larger gain than is actually calculated. Because a higher gain for softer suppression is applied, it is called gain relaxation. When suppression by the originally calculated gain is not significant, the gain is relaxed with a large value to avoid fatal suppression of small-power components which often play an important role in speech recognition. The actual decision for relaxation is performed based on a ratio of averaged input and output powers for more stable result. Gain relaxation has an effect of flattening out the hump and eliminates degraded error rate in ASR.

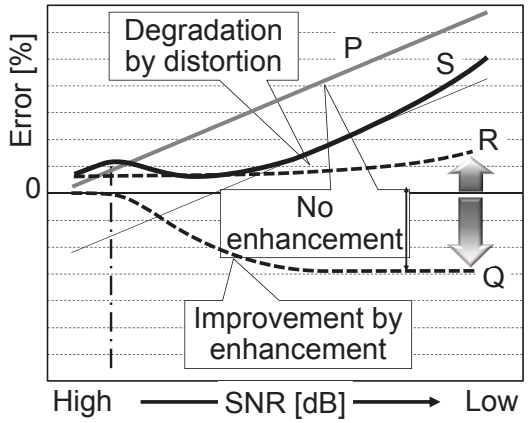

Fig. 2. Error rate with no enhancement (P), improvement by enhancement (Q), degradation by distortion (R), and error rate with enhancement.

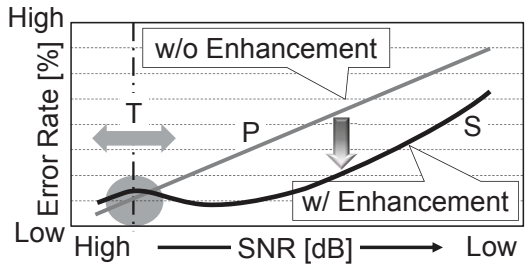

Fig. 3. Input-signal SNR vs. error rate in ASR. Signal enhancement helps reduce error rate except in a high SNR range.

Figure 4 illustrates an NS structure with gain relaxation. Spectral gain is relaxed in the shaded area. Floor gain controller (FGC) calculates a floor gain $G_{flr}(l,k)$, representing a minimum value for the spectral gain, based on the spectral gain $G_s(l,k)$ and a noisy signal power $|X(l,k)|^2$. $G_{flr}(l,k)$ and $G_s(l,k)$ are compared to take whichever is bigger as a relaxed gain $G_R(l,k)$ as

$$G_R(l,k) = \max\{G_{flr}(l,k), G_s(l,k)\} \quad (1)$$

This is a gain relaxation process where a floor gain $G_{flr}(l,k)$ with a maximum, which may be larger than $G_s(l,k)$, pulls up the substantial spectral gain for small distortion.

Shown in Fig. 5 are details of FGC. VAD calculates output to input power ratios $R_L(l)$ and $R_H(l)$ in low and high frequency subbands based on the current value of an input-output power ratio R_0 between the corresponding frequencies k_{S1} and k_{S2} as

$$R_S(l) = \alpha R_S(l-1) + (1-\alpha)R_0, \quad (2)$$

$$R_0 = \sum_{k=k_{S1}}^{k_{S2}} |X(l,k)|^2 / \sum_{k=k_{S1}}^{k_{S2}} G_s(l,k)|X(l,k)|^2, \quad (3)$$

$$\alpha = \begin{cases} \alpha_q & R_0 > R_S(l-1) \\ \alpha_s & otherwise \end{cases}, \quad (4)$$

where S represents L or H and $\alpha_q > \alpha_s$. α is a parameter to control.

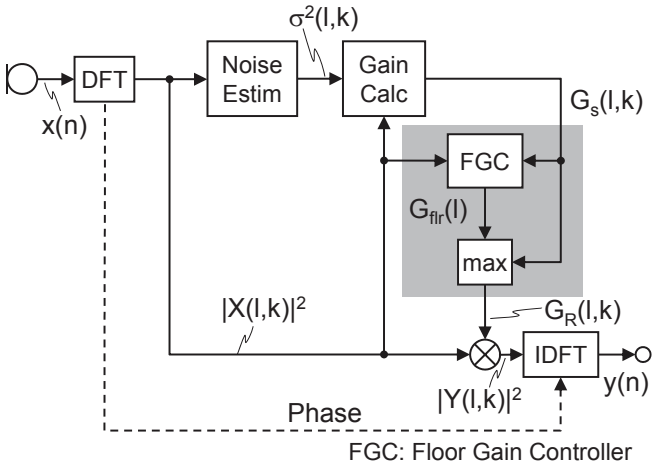

Fig. 4. ASR oriented noise suppressor with gain relaxation. Spectral gain is relaxed in the shaded area.

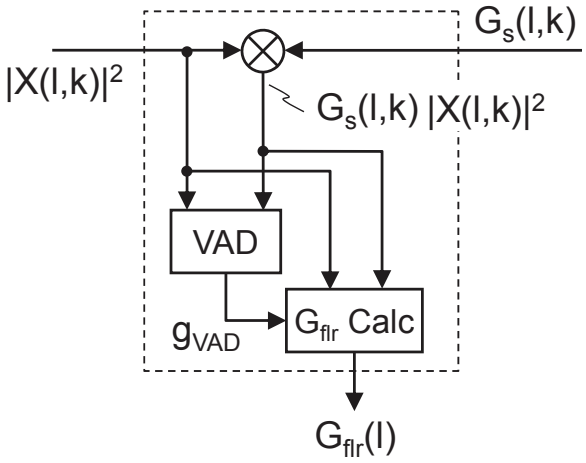

Fig. 5. Floor gain controller (FGC).

the tracking speed and accuracy. $g_{\rm VAD}$ is determined as

$$g_{\rm VAD} = \begin{cases} 1.0 & R_L(l) > g_{th} \quad or \quad R_H(l) > g_{th} \\ 0 & otherwise \end{cases} . \quad (5)$$

For $G_{flr}(l,k)$ calculation, signal and interference power, $P_S(l)$ and $P_N(l)$, are estimated by

$$P_B(l) = \beta P_B(l-1) + (1-\beta)P_0, \quad (6)$$

$$P_0 = \sum_{k=0}^{N-1} |X(l,k)|^2 / \sum_{k=0}^{N-1} G_s(l,k)|X(l,k)|^2, \quad (7)$$

$$B = \begin{cases} S & g_{\rm VAD} = 1.0 \\ N & otherwise \end{cases} , \quad (8)$$

based on the current input-to-output power ratio P_0 in the fullband. Finally, $G_{flr}(l,k)$ is determined by

$$G_{flr}(l,k) = \begin{cases} G_{\max} & R_{flr}(l) \geq R_{\max} \\ G_{slope} R_{flr}(l,k) + G_{\min} & otherwise \\ G_{\min} & R_{flr}(l) \leq R_{\min} \end{cases} , \quad (9)$$

$$R_{flr}(l) = P_S(l)/P_N(l), \quad (10)$$

$$G_{slope} = (G_{\max} - G_{\min})/(R_{\max} - R_{\min}). \quad (11)$$

The final enhanced signal power in each frequency is obtained by multiplying the noisy signal power $|X(l,k)|^2$ by the relaxed gain $G_R(l,k)$ as

$$|Y(l,k)|^2 = G_R(l,k)|X(l,k)|^2. \quad (12)$$

Because $G_{flr}(l,k)$ is proportional to $R_{flr}(l)$ which roughly represents SNR, gain relaxation is achieved.

Figure 6 illustrates a structure of the proposed directional NS. Spectral gain calculation and multiplication are omitted for simplicity, however, may be combined whenever it is needed. For gain relaxation, the same explanation as that for Fig. 4 applies by replacing $G_s(l,k)$ with $G_d(l,k)$, $|X(l,k)|$ with $|X_s(l,k)|$, and $|Y(l,k)|$ with $|Y_s(l,k)|$.

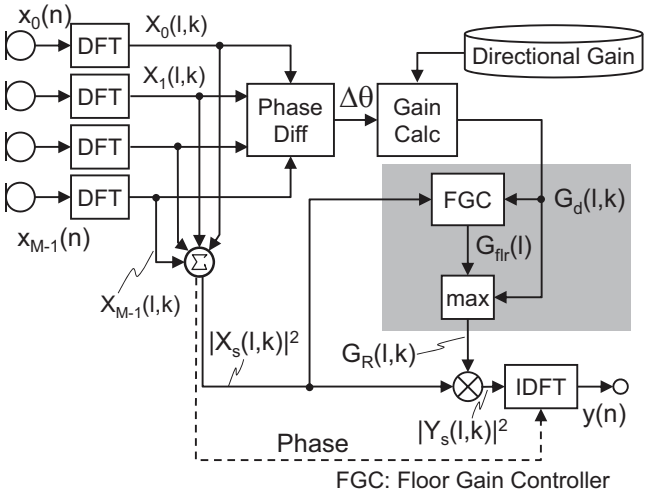

Fig. 6. ASR oriented directional noise suppressor with gain relaxation. Directional gain is relaxed in the shaded area.

4. EVALUATIONS

A laptop PC equipped with two built-in microphones was placed on a table in a $5 \times 5 \times 2.5$ m room. The microphone spacing was 4.5 cm. The screen face was fixed with an angle of 110 degrees to its keyboard and the distance from the center of its screen hinges to a loudspeaker for target-speech radiation was set to 609.6 mm. Four loudspeakers were arranged for noise sources as illustrated in Fig. 7. An interfering speech signal that was located 914.4 mm away from the center of the screen hinges with an angle of 60 degrees to the PC-target-speech line. The target signals consisted of 10 male and 10 female native English speakers. The power ratio of the target signal to the noise was adjusted to 16 dB and that to the interfering speaker to 5 dB. A commercially available speech recognition engine was used.

The recorded 2-channel signals were processed by a directional NS with and without directional gain relaxation. The directional gain $G_d(k,l)$ was designed with a constant beamwidth along fre-

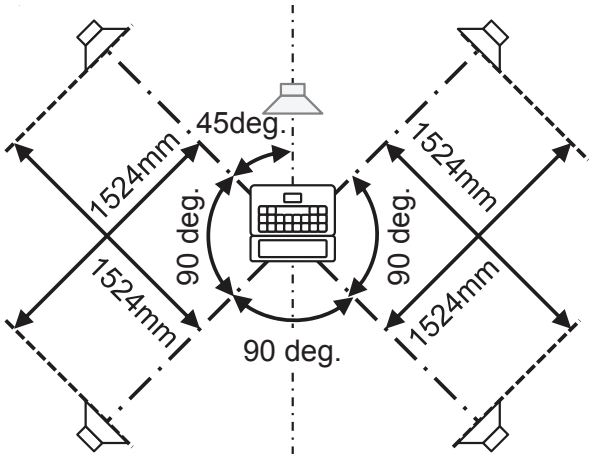

Fig. 7. Layout of four loudspeakers for noise source.

Table 1. Parameters

Referece freq. k_0	1 kHz	k_{L1}	1kHz	$G_{\max}$	1.0
Passband @ k_0	± 30 deg	k_{L2}	4kHz	$G_{\min}$	0.3
Stopband @ k_0	± 45 deg	k_{H1}	4kHz	α_q	0.8
Passband gain	$G_{\max}$	k_{H2}	6kHz	α_s	0.2
Stopband gain	$G_{\min}$	$R_{\max}$	24dB	β	0.98
g_{th}	-0.2dB	$R_{\min}$	18dB		

quency [28]. Evaluations were performed without a spectral gain for four different conditions, namely, clean speech, babble noise, stationary noise, and speech interference. They are to model an ideal environment, a party environment, a car environment, and an interfering-talker environment. The DOA of the interfering talker was set to 45 degrees. Parameters are shown in Tab. 1.

Figures 8 and 9 show command error rate (CER) and word error rate (WER) by no speech enhancement (NoSE), a directional NS without gain relaxation (dNS) [28], and the directional NS with gain relaxation (dNS-GR). A short bar exhibits a low error rate and good performance. Figure 8 indicates that both dNS and dNS-GR achieve error rates comparable to or lower than NoSE with respect to CER. For speech interference, dNS-GR exhibits 0.5

In the case of WER in Fig. 9, the error rate by dNS is 1.7% higher than that by NoSE for clean speech. It is 20 words of 1185 words and may not be significant in number. However, such an error is more seriously recognized at a high SNR because there should be little error by nature. This increase in error does not exist in case of dNS-GR due to directional gain relaxation. dNS-GR has 0.8It is not recognized as serious as the increase by dNS for clean speech. Because of a lower SNR than clean speech, some degradation is anticipated and small one is not noticeable by itself.

These characteristics are better demonstrated in Fig. 10 which shows error-rate improvements for dNS and dNS-GR in cases of CER and WER. Because this metric means a difference from the error by NoSE, there is no score for NoSE itself. It should be noted that a negative value represents degradation from NoSE. It is easy to understand that dNS-GR achieves all positive scores. It demonstrates that dNS-GR always provides some improvement over NoSE in both CER and WER. As a trade-off, some compromise compared to dNS needs to be accepted.

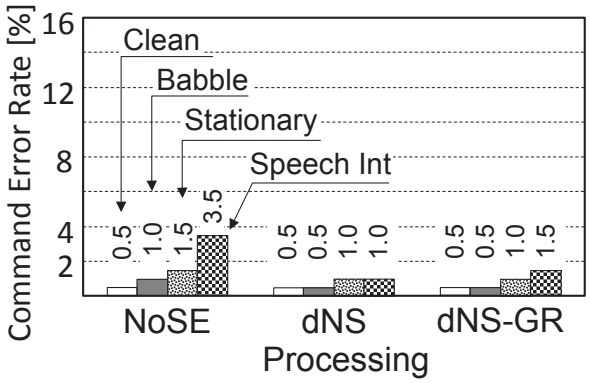

Fig. 8. Command error rate (CER) for 200 commands.

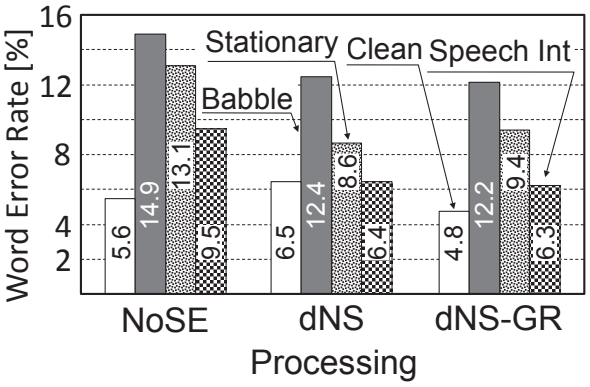

Fig. 9. Word error rate (WER) for dictation with 1185 words.

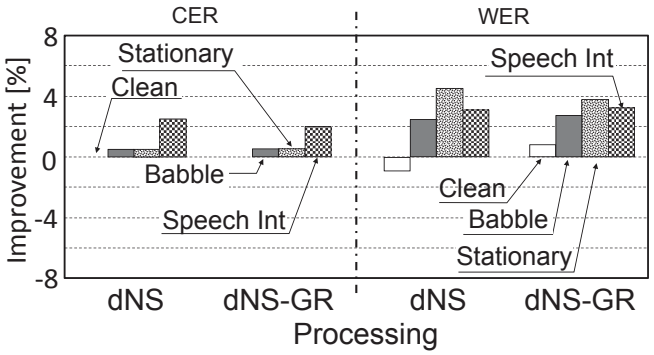

Fig. 10. CER and WER error-rate improvements.

5. CONCLUSION

Gain relaxation in signal enhancement designed for speech recognition with an unaware local noise source has been proposed. Gain relaxation, which takes an SNR based gain floor or the original directional gain whichever has a larger value, has been introduced to make softer suppression when there is little interference and no need for suppression. It has been demonstrated by evaluation with recorded signals that this relaxation eliminates potential degradation in speech recognition caused by small undesirable distortion in the target signal components. Improvements over the output with no signal enhancement have been achieved without sacrificing the performance for clean speech.

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] 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.

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

[4] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," IEEE Trans. Speech and Audio Processing, Vol. 11, No. 5, pp. 466–475, Sep. 2003.

[5] 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.

[6] M.-S. Choi and H.-G. Kang, "A two-channel minimum mean-square error log-spectral amplitude estimator for speech enhancement," Proc. HSCMA2008, pp.152–155, May 2008.

[7] J. Freudenberger, S. Stenzel and B. Venditti, "A noise PSD and cross-PSD estimation for two-microphone speech ehancement systems," Proc. SSP2009, pp.709–712, Aug. 2009.

[8] S. -Y. Jeong, K. Kim, J. -H. Jeong, K. -C. Oh, and J. Kim, "Adaptive noise power spectrum estimation for compact dual channel speech enhancement," Proc. ICASSP2010, pp.1630–1633, Apr. 2010.

[9] K. Kim, S. -Y. Jeong, J. -H. Jeong, K. -C. Oh, and J. Kim, "Dual channel noise reduction method using phase difference-based spectral amplitude estimation," Proc. ICASSP2010, pp.217–220, Apr. 2010.

[10] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," IEEE Trans. ASLP, Vol. 20, No. 2, pp.599–609, Feb. 2012.

[11] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," Proc. ICASSP2012, pp.217–220, Mar. 2012.

[12] J. Zhang, R. Xia, Z. Fu, J. Li, and Y. Yan, "A fast two-microphone noise reduction algorithm based on power level ratio for mobile phone," Proc. ICSLP2012, pp.206–209, Dec. 2012.

[13] Z.-H. Fu, F. Fan and J. -D. Huang, "Dual-microphone noise reduction for mobile phone application," Proc. ICASSP2013, pp.7239–7243, May 2013.

[14] J. Taghia, R. Martin, J. Taghia and A. Leijon, "Dual-channel noise reduction based on a mixture of circular-symmetric complex Gaussians on unit hypersphere," Proc. ICASSP2013, pp.7289–7293, May 2013.

[15] B. Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, Jr., R. C. Goodlin: Adaptive noise cancelling: principles and applications, *Proc. IEEE*, 63, (12), pp.1692–1716, 1975.

[16] A. Sugiyama, "Low-distortion noise cancellers – Revival of a classical technique," *Speech and audio processing in adverse environment*, Chap. 7, Hänsler and Schmidt, ed. Springer, 2008.

[17] X. Zhang, H. Zeng, and A. Lunardhi, "Noise estimation based on an adaptive smoothing factor for improving speech quality in a dual-microphone noise suppression system," Proc. ICSPCS2011, pp.1–5, Dec. 2011.

[18] A. Sugiyama, M. Kato, and M. Serizawa, "A low-distortion noise canceller with an SNR-modified partitioned power-normalized PNLMS algorithm," Proc. APSIPA ASC 2009, pp.222–225, Oct. 2009.

[19] A. Sugiyama, and R. Miyahara, "A low-distortion noise canceller with a novel stepsize control and conditional cancellation," Proc. EUSIPCO2014, Sep. 2014.

[20] M. Brandstein and D. Ward (Eds), "Microphone arrays," Springer, 2001.

[21] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," *Adaptive signal processing, Applications to real-world problems*, Chap. 6, Benesty and Huang, ed. Springer, 2003.

[22] G. W. Elko and J. Meyer, "Microphone arrays," *Handbook of speech processing*, Chap. 50, Benesty, Sondhi, and Huang, ed. Springer, 2008.

[23] J. Benesty, Y. Huang, and J. Chen (Eds), "Microphone array signal processing," Springer, 2008.

[24] L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," IEEE Trans. Antennas and Propagations, vol. AP-30, no. 1, pp.27–34, Jan. 1982.

[25] A. Sugiyama and R. Miyahara, "A new generalized sidelobe canceller with a compact array of microphones suitable for mobile terminals," Proc. ICASSP2014, pp.820–824, May 2014.

[26] P. Aarabi and G. Shi, "Phase-based dual-microphone tobust speech enhancement," IEEE Trans. Syst., Man, and Cyber., Vol. 34, No. 4, pp.1763–1773, Aug. 2004.

[27] O. u. R. Qazi, B. v. Dijk, M. Moonen and J. Wouters, "Speech understanding performance of cochlear implant subjects using time-frequency masking-based noise reduction," IEEE Trans. Bio. Eng., Vol. 59, No. 5, pp.1364–1373, May 2012.

[28] A. Sugiyama and R. Miyahara, "A directional noise suppressor with a specified beamwidth," Proc. ICASSP2015, Apr. 2015.

[29] A. Sugiyama and R. Miyahara, "A directional noise suppressor with a constant beamwidth for multichannel signal enhancement," Proc. EUSIPCO2015, Sep. 2015.