AN ADAPTIVE NOISE CANCELER WITH LOW SIGNAL-DISTORTION BASED ON VARIABLE STEPSIZE SUBFILTERS FOR HUMAN-ROBOT COMMUNICATION

Miki Sato, Akihiko Sugiyama, and Shin'ichi Ohnaka

Multimedia Research Laboratories
NEC Corporation
Kawasaki 216-8555, **JAPAN**

ABSTRACT

This paper proposes an adaptive noise canceller (ANC) with low signal-distortion for human-robot communication. The proposed ANC has two sets of adaptive filters for noise and crosstalk; namely, main filters (MFs) and subfilters (SFs) connected in parallel thereto. To reduce signal-distortion in the output, the stepsizes for coefficient adaptation in the MFs are controlled according to estimated signal-to-noise ratios (SNRs) of the input signals. This SNR estimation is carried out using SF output signals. The stepsizes in the SFs are determined based on the ratio of the primary and the reference input signals to cope with a wider range of SNRs. This ratio is used as a rough estimate of the input signal SNR at the primary input. Computer simulation results using TV sound and human voice recorded in a carpeted room show that the proposed ANC reduces both residual noise and signal-distortion by as much as 20dB compared to the conventional ANC. Evaluation in speech recognition with this ANC reveals that with a realistic TV sound level, as good recognition rate as in the noise-free condition is achieved.

1. INTRODUCTION

Speech recognition is essential in human-robot communication [1]. The recognition rate has reached a satisfactory level for on-microphone scenarios where the microphone is located close to the mouth so that undesirable influence by non-speech signals is negligible. However, it is still challenging to perform off-microphone speech recognition, where the microphone is placed at a distance from the speech source [2]. The speech at the microphone is contaminated by reverberation, ambient noise, and inteference, which degrade the speech recognition rate. Therefore, extracting the desired speech from a noise-contaminated signal plays a key role in human-robot communication.

A signle-microphone or multimicrophone system is used for eliminating undesirable signals contaminating the target speech according to the requirements specific to the application [3]. For human-robot communication, a two-microphone solution based on adaptive noise cancellation [4, 5] is a good compromise with respect to the number of microphones, noise cancellation capability, and the distortion. In an adaptive noise canceller (ANC), two microphones are employed: the primary microphone to obtain the noise-contaminated speech and the reference microphone to obtain only the correlated component of the noise detected by the primary microphone. The noise collected at the reference microphone is processed by an adaptive filter to generate a replica of the noise component in the primary input signal. When there is crosstalk leaking into the reference microphone from the speech source, the CTRANC structure [5] can be a good solution.

For achieving good noise reduction and small signal distortion at the ANC output, a dual-filter structure has been developed in combination with the CTRANC structure [6]. This ANC has two pairs of cross-coupled adaptive filters, each of which consists of the main filter and a subfilter. The subfilters serve as pilot filters whose output is used to estimate the signal-to-noise ratios (SNRs) of the primary and the reference signals. The stepsizes for coefficient adaptation in the main filters are controlled according to the estimated SNRs. Thanks to the stepsize control using the subfilters, good noise suppression and low signal-distortion at the output are simultaneously achieved. However, this ANC was developed for communication headsets in noisy environment. The fixed stepsize for each subfilter does not provide satisfactory performance for a wide range of SNRs that are encountered in human-robot communication.

This paper proposes a new adaptive noise canceller with low signal-distortion for human-robot communication. The subfilter stepsizes are adaptively controled based on an approximated SNR at the primary and the reference input. In the following section, the new noise canceller with variable stepsize subfilters is presented. In Section 3, performance of the proposed noise canceller is evaluated by computer simulations in comparison with the conventional ANC. Results of speech recognition are also presented in a human-robot communication scenario.

2. ANC WITH VARIABLE STEPSIZE SUBFILTERS

Figure 1 shows a blockdiagram of the proposed ANC with variable stepsize subfilters. Four adaptive filters, namely, the main adaptive filters (MF1, MF2) and the sub adaptive filter (SF1, SF2) generate noise and crosstalk replicas. Average powers $R_S(k)$ and $R_N(k)$ of the primary signal $x_P(k)$ and the reference signal $x_R(k)$ are first calculated. A ratio of $R_S(k)$ to $R_N(k)$, representing a rough estimate of the SNR at the primary input, is used for controling the stepsize of SF1 and SF2.

The SF1 output $y_1(k)$ and the subtraction result $e_1(k)$ are used to estimate a more precise SNR at the primary input. $e_1(k)$ serves as an approximation to the target speech, and $y_1(k)$ is used as that to the noise. The stepsize for MF1 is controlled based on the estimated SNR calculated from SF1 output signals. SF2 works for crosstalk instead of noise in a similar way to that of SF1. The resulting SNR estimate from SF2 output signals is used to control the MF2 stepsize.

2.1. Principle of Stepsize Control

The stepsize for coefficient adaptation should be kept smaller when there is more interference in the error. In the structure in Fig. 1, the stepsize, $\mu_{SF1}(k)$, for SF1 and the stepsize, $\mu_{MF1}(k)$, for MF1

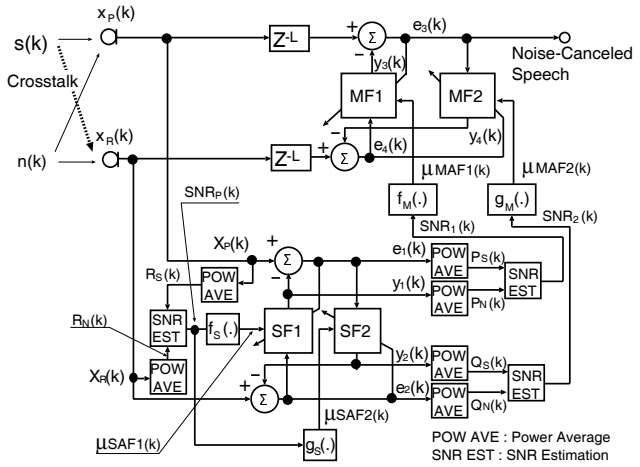

Figure 1: Blockdiagram of the Proposed ANC.

should be set to a small value when the SNR at the primary input is high to avoid distortion at the ANC output. On the other hand, $\mu_{SF1}(k)$ and $\mu_{MF1}(k)$ can be set large when this SNR is low for fast convergence and rapid tracking of the noise-path change. A similar rule applies to $\mu_{SF2}(k)$ and $\mu_{MF2}(k)$ for coefficient adaptation in SF2 and MF2 with respect to the SNR at the reference input. All these stepsizes can be controlled appropriately once the SNRs for the adaptive filters become available.

2.2. Stepsize Control in SF1 and SF2

The SNR for the primary signal, $SNR_P(k)$, is approximated by

$$SNR_P(k) = 10 \log_{10} \{R_S(k)/R_N(k)\} \text{ [dB]} \quad (1)$$

$$R_S(k) = \sum_{j=0}^{M-1} x_P^2(k-j) \quad (2)$$

$$R_N(k) = \sum_{j=0}^{M-1} x_R^2(k-j) \quad (3)$$

where $x_P(k)$ and $x_R(k)$ are the primary and the reference signals. They are summed up over M samples for an averaged SNR.

$\mu_{SF1}(k)$ and $\mu_{SF2}(k)$ are controlled by the estimated SNR, $SNR_P(k)$, as in the following equations.

$$\mu_{SF1}(k) = \begin{cases} \mu_{S_{min}} & SNR_P(k) > SNR_{P_{max}} \\ \mu_{S_{max}} & SNR_P(k) < SNR_{P_{min}} \\ f_S(SNR_P(k)) & otherwise \end{cases} \quad (4)$$

$$\mu_{SF2}(k) = \begin{cases} \mu_{S_{min}} & SNR_P(k) < SNR_{P_{min}} \\ \mu_{S_{max}} & SNR_P(k) > SNR_{P_{max}} \\ g_S(SNR_P(k)) & otherwise \end{cases} \quad (5)$$

$\mu_{S_{max}}$ and $\mu_{S_{min}}$ are the maximum and the minimum stepsizes for $\mu_{SF1}(k)$ and $\mu_{SF2}(k)$. $f_S(\cdot)$ and $g_S(\cdot)$ are functions of $SNR_P(k)$. $f_S(\cdot)$ should be a decreasing function because a small stepsize is suitable for a large SNR. On the other hand, it is desirable that $g_S(k)$ is an increasing function.

2.3. Stepsize Control in MF1 and MF2

Estimated SNR for MF1, $SNR_1(k)$, and that for MF2, $SNR_2(k)$, are calculated based on the output signals of SF1 and SF2. $SNR_1(k)$ and $SNR_2(k)$ are given by

$$SNR_1(k) = 10 \log_{10} \{P_S(k)/P_N(k)\} \text{ [dB]} \quad (6)$$

$$SNR_2(k) = 10 \log_{10} \{Q_S(k)/Q_N(k)\} \text{ [dB]}. \quad (7)$$

$P_S(k)$ and $P_N(k)$ are estimated speech and noise in the primary signal. Similarly, $Q_S(k)$ and $Q_N(k)$ represent estimates of the crosstalk and the noise in the reference signal. They can be calculated by SF1 and SF2 output signals by

$$P_N(k) = \sum_{j=0}^{M-1} y_1^2(k-j) \quad (8)$$

$$P_S(k) = \sum_{j=0}^{M-1} e_1^2(k-j) \quad (9)$$

$$Q_S(k) = \sum_{j=0}^{M-1} y_2^2(k-j) \quad (10)$$

$$Q_N(k) = \sum_{j=0}^{M-1} e_2^2(k-j). \quad (11)$$

$y_1(k)$ and $e_1(k)$ are the noise replica generated by SF1 and the subtractor output signal approximating the desired speech. $y_2(k)$ and $e_2(k)$ are respectively the crosstalk replica that is the output of SF2 and the subtractor output approximating the noise. $P_S(k)$, $P_N(k)$, $Q_S(k)$, and $Q_N(k)$ all contain summing operation over M samples as $R_S(k)$ and $R_N(k)$.

The following equations determine $\mu_{MF1}(k)$ and $\mu_{MF2}(k)$ based on $SNR_1(k)$ and $SNR_2(k)$, respectively.

$$\mu_{MF1}(k) = \begin{cases} \mu_{M_{min}} & SNR_1(k) > SNR_{1_{max}} \\ \mu_{M_{max}} & SNR_1(k) < SNR_{1_{min}} \\ f_M(SNR_1(k)) & otherwise \end{cases} \quad (12)$$

$$\mu_{MF2}(k) = \begin{cases} \mu_{M_{min}} & SNR_2(k) < SNR_{2_{min}} \\ \mu_{M_{max}} & SNR_2(k) > SNR_{2_{max}} \\ g_M(SNR_2(k)) & otherwise \end{cases} \quad (13)$$

$\mu_{M_{max}}$ and $\mu_{M_{min}}$ are the maximum and the minimum stepsizes for $\mu_{MF1}(k)$ and $\mu_{MF2}(k)$. $f_M(\cdot)$ and $g_M(\cdot)$ are a decreasing function of $SNR_1(k)$ and an increasing function of $SNR_2(k)$, respectively.

The estimated $SNR_1(k)$ and $SNR_2(k)$ generate time delays depending on M. To compensate for these delays, L-sample delay units Z^{-L} are incorporated into the input paths of the primary and the reference signals of MF1 and MF2.

3. EVALUATION BY RECORDED SIGNALS

3.1. Noise Reduction and Distortion

Performance of the proposed ANC was evaluated by computer simulations from the viewpoints of noise reduction and speech distortion in comparison with the conventional algorithm [6]. TV sound and male voice was recorded in a carpeted room with a dimension of 5.5 m (Width) × 5.0 m (Depth) × 2.4 m (Height) in a human-robot communication scenario. The primary microphone was mounted on the forehead and the reference microphone was attached to upper back of a robot, whose height is approximately

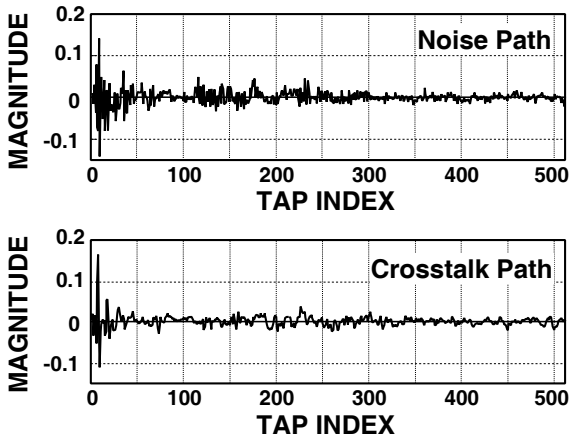

Figure 2: Impulse Response of the Noise and the Crosstalk Path.

Table 1: Parameter Settings.

Parameter	Selected Value	Parameter	Selected Value
N	512	M	128
L	64		
μ_{SF1}	0.02	μ_{SF2}	0.002
$SNR_{P_{max}}$	5 dB	$SNR_{P_{min}}$	-7 dB
$SNR_{1_{min}}$	0 dB	$SNR_{1_{max}}$	10 dB
$\mu_{S_{min}}$	0.002	$\mu_{S_{max}}$	0.1
$SNR_{2_{max}}$	0 dB	$SNR_{2_{min}}$	-10 dB
$\mu_{M_{max}}$	0.1	$\mu_{M_{min}}$	0.002

0.4 m. The impulse responses of the noise path and the crosstalk path were measured with this set-up. An example with a speaker distance of 0.5m is depicted in Fig. 2

The noise component, which is the convolution of the noise source with the noise path, was added to the speech signal to create a noise-contaminated signal. The reference signal was generated by adding the noise to the crosstalk generated by convolution of the speech signal with the crosstalk path. SNRs of the primary and the reference signals in the utterance were set to [35, 15] and [10, -10] dB, which correspond to low and high noise levels with a speaker distance of 0.5 and 1.5 m, respectively. The sampling frequency was 11.025 kHz. Other parameters are shown in Table 1.

Figure 3 shows the stepsize for MF1 (upper) and that for MF2 (lower) in the case of [35, 15] dB SNRs. Dips in the upper figure and peaks in the lower figure both correspond to speech sections. The stepsizes of the proposed ANC represented by the solid line show better match with speech sections than those of the conventional ANC expressed in a dotted line.

The output signal (upper) and speech distortion (lower) at the output are illustrated in Figs. 4 and 5 for SNR settings of [35, 15] dB and [10, -10] dB, respectively. The speech distortion, $D(k)$, was calculated by

$$D(k) = 10\ \log_{10}\{\frac{\sum_{j=0}^{Q-1}\{e_3(k-j)-s(k-j)\}^2}{\sum_{j=0}^{Q-1}s^2(k-j)}\}\ [\text{dB}]. \quad (14)$$

Peaks in the output signal and dips in the distortion correspond to speech sections. Both noise reduction and distortion are improved by as much as 20dB in Fig. 4. In the case of Fig. 5, noise reduction

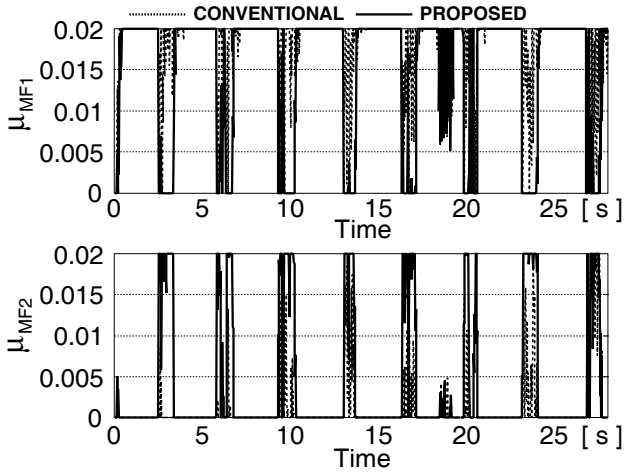

Figure 3: Step Size of MF1 and MF2.

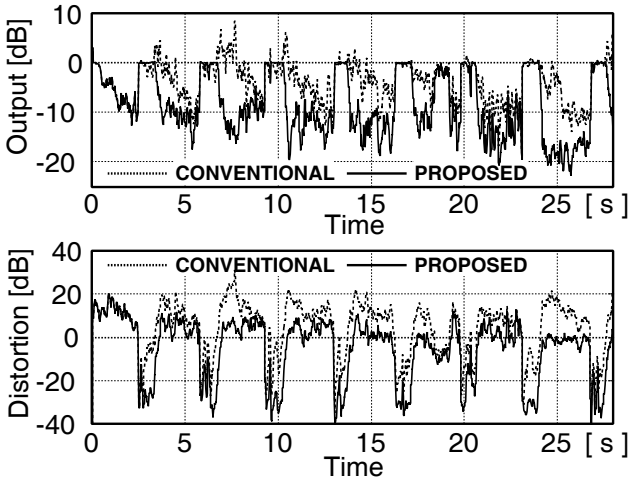

Figure 4: Output Signal and Distortion, SNR=[35, 15] dB.

is improved by as much as 8dB. The improvement in distortion in Fig. 5 is not as evident as that in Fig. 4. However, improvement close to 10dB can be observed in circled areas.

3.2. Speech Recognition

Speech recognition was performed with noise-cancelled speech. The experimental set-up is illustrated in Fig. 6. 1500 utterances by 30 defferent male, female, and child speakers were presented at a distance of 0.5 and 1.5 m. The noise source was placed at a distance of 1.0 m in a direction of 30, 60, 90, 135, or 180 degrees. Speaker independent speech recognition based on demi-syllable hidden Markov model [8] was used with a dictionary of 600 robot commands.

Figures 7 and 8 depict the speech recognition rate for a commercial and a news TV-sound. Shaded columns represent improvements, *i.e.* the difference in the recognition rate with and without the noise canceler. For the commercial program, the recognition rate is equivalent to that in the noise-free condition when the speaker distance is 0.5 m with a 57dB noise in directions of 90 to 180 degrees. The maximum improvement reaches 65%. The

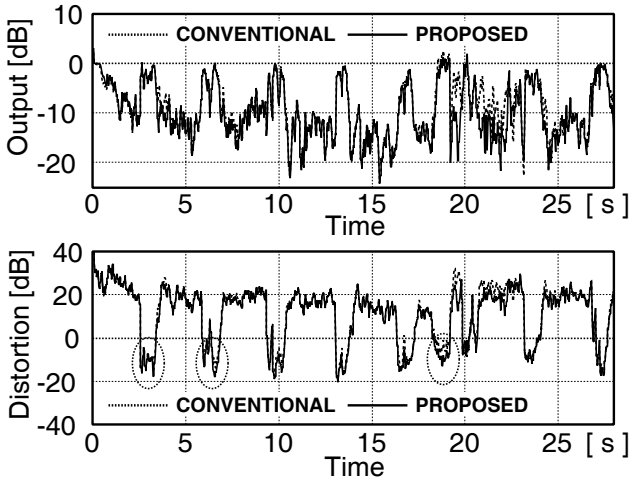

Figure 5: Output Signal and Distortion, SNR=[10, -10] dB.

recognition rate is degraded accordingly for off-direction noise placement, longer distance of the speech source, and/or a higher noise level of 67dB. For the news program, the recognition rate is slightly degraded compared to that in Fig. 7 for noise directions of 135 and 180 degrees. However, with a noise directions of 30, 60, and 90 degrees, the recognition rate is significantly lower. The improvement is also degraded accordingly. This degradation is caused by similar spectral components in the the speech to be recognized and the news program.

4. CONCLUSION

An adaptive noise canceller (ANC) with low signal-distortion for human-robot communication has been proposed. The new ANC employs a dual-filter structure for both noise and crosstalk paths, where each subfilter estimates the input SNR for controlling the main-filter stepsize. The ratio of the primary and the reference input signals are used to control the subfilter stepsizes appropriately so that good noise cancellation and low signal distortion are simultaneously obtained for a wide range of SNRs. Computer simulation results using TV sound and human voice recorded in a carpeted room has shown that both residual noise and signal-distortion are reduced by as much as 20dB compared to the conventional ANC. Evaluation in speech recognition with a real TV sound has revealed that the recognition rate is improved by 20-65% when the TV is located in a close vicinity of the reference microphone direction.

5. REFERENCES

[1] Y. Fujita, “Personal Robot PaPeRo," J. of Robotics and Mechatronics, vol.14, No.1, Jan. 2002.

[2] H.-G. Hirsch and D. Pearce, “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions,” Proc. ISCA ITRW ASR 2000, Sep. 2000.

[3] M. Brandstein and D. Ward, “Microphone Arrays,” Springer Verlag, Berlin, 2001.

[4] B.Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, Jr., and R. C. Goodlin, “Adaptive Noise Cancelling : Principles and Applications,” Proc. IEEE, Vol. 63, No.12, pp. 1692-1716, 1975.

[5] G. Mirchandani, R. L. Zinser and J. B. Evans, “A New Adaptive Noise Cancellation Scheme in the Presence of Crosstalk,” IEEE Trans. Circuits and Systems, Vol. 39, pp. 681-694, 1992.

[6] S. Ikeda and A. Sugiyama, “An Adaptive Noise Canceller with Low Signal-Distortion in the Presence of Crosstalk,” IEICE Trans. Fund, pp.1517-1525, Aug. 1999.

[7] G. C. Goodwin and K. S. Sin, “Adaptive filtering, prediction and control,” Englewood Cliffs, NJ: Prentice-Hall Info. Syst. Sci., 1985.

[8] T. Watanabe, “Problems in the design of a speech recognition system and their solution," Trans., vol.J.79-D-II, no.12, pp.2022-2031, Dec. 1996 (in Japanese).

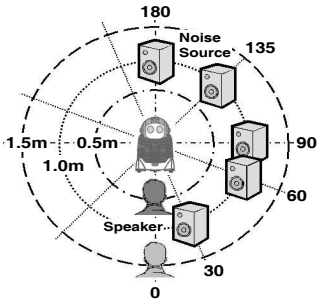

Figure 6: Experimental Set-Up for Speech Recognition.

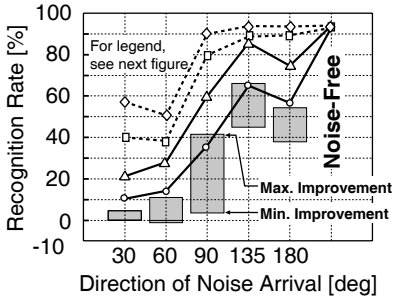

Figure 7: Speech Recognition Result (Commercial Program).

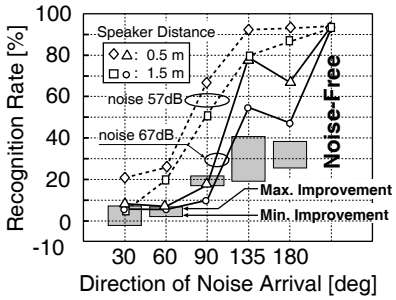

Figure 8: Speech Recognition Result (News Program).