DOUBLE-TALK ROBUST FREQUENCY DOMAIN ECHO CANCELLATION ALGORITHM WITH SCALABLE NONLINEAR REFERENCE AND ERROR FUNCTIONS

Suehiro Shimauchi, Yoichi Haneda, and Akitoshi Kataoka

NTT Cyber Space Laboratories, NTT Corporation
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180-8585 Japan
Email: shimauchi.suehiro@lab.ntt.co.jp

ABSTRACT

Several robust algorithms for echo cancellation use nonlinear reference and/or error functions. Most of them require time-variant threshold estimators, e.g., noise level estimators or double-talk detectors, since their nonlinearities have to be adjusted with changes in near-end signal levels. We propose a new frequency domain adaptive algorithm: the gradient-limited FLMS (GL-FLMS), in which the coefficients are updated with a nonlinear function of the error scaled by the reference magnitude, i.e., the error-to-reference ratio (ERR). When the acoustic coupling level between loudspeaker and microphone is bounded, the ERR is also bounded during single-talk, but may increase during double-talk. The GL-FLMS limits unexpected increases in the ERR to prevent divergence of the coefficients, while not neglecting updates to adjust when a large reference signal introduces a large error during single-talk.

1. INTRODUCTION

Adaptive filters have been successfully applied to acoustic echo cancellation. In the field, the robustness of the adaptation against double-talk is strongly required. Several approaches fulfilling this requirement have been proposed. In two echo path models [2] and duo-filter control systems [9], an adaptive background filter is used with a fixed foreground filter. These approaches perform well, but the dual filters increase the computational cost by about 50 % and control for the transfer of coefficients is necessary. Some adaptive algorithms in which the coefficients are updated with nonlinear functions of the reference input and/or error signals [5, 7] have robust properties. For example, error nonlinearities are used to achieve double-talk robustness for certain algorithms [1, 10, 11]. In the first [1], an infinite clipper is applied to the error. Other algorithms [10, 11] introduce a variable scale factor to the error nonlinearity, though a double-talk detector (DTD) is required to control it. The variable step-size control approach [8] can also be regarded as a nonlinear approach. In this variant of the NLMS algorithm, the step-size is controlled by using nonlinear function of the reference power level. A noise level estimator controls the nonlinearity threshold.

In this paper, we focus on the frequency domain echo cancellers which have a single-tap adaptive filter in each frequency bin. Of the conventional algorithms [1, 8, 10, 11], only one [10] is addressed to the frequency domain, though some of the others can be made applicable. However, none of them inherently has signal level scalability, so their robustness is only achieved at the cost of peripheral estimator or detector. We propose a new robust algorithm with a nonlinear function that provides scaling of the reference and error signal levels. This algorithm provides good performance without time-variant threshold estimators.

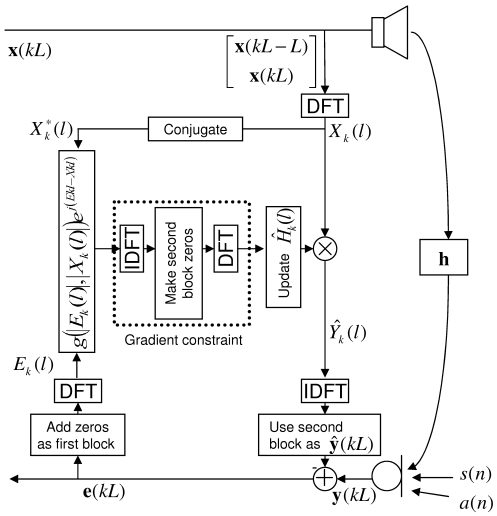

Fig. 1. Frequency domain echo canceller.

2. FREQUENCY DOMAIN ECHO CANCELLER

2.1. Configuration

The frequency domain echo canceller that we are discussing is shown in Fig. 1. The reference signal $x(n)$ at a discrete time index n is picked up by a microphone, passing through a room echo path which has an impulse response modeled as $\mathbf{h}=[h_1,\cdots,h_L]^T$ where the length is L and T denotes the transpose. The total microphone signal $y(n)$ is expressed as

$$y(n) = \mathbf{h}^T\mathbf{R}\mathbf{x}(n) + s(n) + a(n), \qquad (1)$$

where $\mathbf{x}(n)=[x(n-L+1),\cdots,x(n)]^T$, $\mathbf{R}$ is the matrix that reverses the order of element of $\mathbf{x}(n)$, and $s(n)$ and $a(n)$ are, respectively, speech and ambient noise at the near-end. The transformed reference $X_k(l)$ at the k-th step and l-th frequency bin is an element of the discrete Fourier transform (DFT) of $[\mathbf{x}^T(kL-L),\mathbf{x}^T(kL)]^T$, where the index k is counted up every L increments of n, and $l=0,\cdots,2L-1$. The filter coefficient for k and l is $\hat{H}_k(l)$. The filter output is $\hat{Y}_k(l) = \hat{H}_k(l)X_k(l)$. The echo replica $\hat{\mathbf{y}}(kL)$ corresponds to the last L elements of the inverse DFT of $[\hat{Y}_k(0),\cdots,\hat{Y}_k(2L-1)]^T$. The error is $\mathbf{e}(kL) = \mathbf{y}(kL) - \hat{\mathbf{y}}(kL)$, where $\mathbf{y}(kL) = [y(kL-L+1),\cdots,y(kL)]^T$. The transformed error $E_k(l)$ is an element of the DFT of $[\mathbf{z}^T$,

$\mathbf{e}^T(kL)]^T$, where $\mathbf{z}$ is an $L\times 1$ zero vector. By omitting the gradient constraint in Fig. 1, we concentrate on the unconstrained case where we can formulate the equation for updating $\hat{H}_k(l)$ as

$$\hat{H}_{k+1}(l)=\hat{H}_k(l)+\mu\cdot g(|E_k(l)|,|X_k(l)|)\,e^{j(\theta_{Ekl}-\theta_{Xkl})}, \quad (2)$$

where θ_{Ekl} and θ_{Xkl} denote phases of $E_k(l)$ and $X_k(l)$ respectively, $g(|E_k(l)|,|X_k(l)|)$ is an arbitrary function of $|E_k(l)|$ and $|X_k(l)|$, which we simply denote as $g(\cdot)$ below, and μ is a step-size with value determined by what $g(\cdot)$ is.

In the unconstrained fast (or frequency domain) LMS (UFLMS) [4] case,

$$g(|E_k(l)|,|X_k(l)|)=|E_k(l)|\frac{|X_k(l)|}{P_k(l)}, \quad (3)$$

where $P_k(l)$ is the smoothed power of $X_k(l)$ as obtained by using a smoothing factor α: $P_k(n)=(1-\alpha)P_{k-1}(n)+\alpha|X_k(n)|^2$.

2.2. Frequency domain sign-sign algorithm

Our purpose in this paper is to find a desirable $g(\cdot)$ that provides both strong double-talk stability and fast convergence. As an example of a robust algorithm, we review the frequency domain sign-sign algorithm (FSSA) [12], which has

$$g(|E_k(l)|,|X_k(l)|)=1. \quad (4)$$

While the time domain sign-sign algorithm (SSA) is known to be computationally efficient, its convergence is very slow [5, 7]. The FSSA achieves relatively faster convergence, especially for the colored reference, since its adaptation does not depend on the reference signal's level: $|X_k(l)|$. Since the FSSA estimates only the phase difference $\theta_{Ekl}-\theta_{Xkl}$, it is robust against noise in the error as long as the phase difference is random. However, the FSSA has a similar property to the SSA, in that its region of convergence is a ball around the true solution with radius proportional to the step-size μ [6]. This imposes some limit on the accuracy of convergence.

3. NEW ROBUST ALGORITHM

3.1. Scalable nonlinearities

For robust adaptation, the error estimation is important, i.e, to what extent are the near-end signals included in the error? The error-to-reference ratio (ERR) is useful as a means for estimating the influence of the near-end signals in the error. In the single-talk case of the far-end talker, the ERR does not exceed the acoustic coupling level (ACL) between the loudspeaker and the microphone, unless the adaptive filter diverges. In the double-talk case, the range of the ERR is widely distributed with the ratio between the averaged levels of the far- and near-end signals as its mean. So we can expect different distributions of the ERR for the single- and double-talk cases. It is reasonable to limit the ERR to the level expected during single-talk. Taking this into account, we define $g(\cdot)$ for a new robust version of the UFLMS as

$$g(|E_k(l)|,|X_k(l)|)=\psi_{S_1}\left(\frac{|E_k(l)|}{|X_k(l)|}\right)\frac{|X_k(l)|^2}{P_k(l)}, \quad (5)$$

where $\psi_b(a)=\min\{a,b\}$, which is known as the Huber function [3], and S_1 is a (limiter) threshold. S_1 can be determined from the ACL and the averaged level balance between the near- and far-end signals. We can choose a fixed S_1 according to the specifications of the real system in which the algorithm is to be implemented.

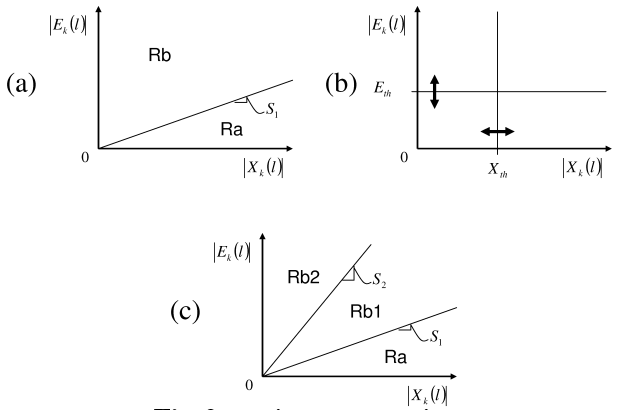

Fig. 2. Region segmentations.

3.2. Gradient-limited FLMS algorithm

An alternative to (5) is given below

$$g(|E_k(l)|,|X_k(l)|)=\psi_{S_1}\left(|E_k(l)|\frac{|X_k(l)|}{P_k(l)}\right). \quad (6)$$

Equation (6) is a gradient-limited version of (3). We call the algorithm with (6) the gradient-limited FLMS (GL-FLMS). If $|E_k(l)|\cdot|X_k(l)|/P_k(l)>S_1$, it corresponds to the FSSA in (4). The stability of the GL-FLMS is sufficiently ensured if μ is chosen from within the stability bounds of the UFLMS. Since there is no essential difference between (5) and (6), in the discussion below, we take (6) as the simpler example.

3.3. Interpretation and improvement

As shown in Fig. 2 (a), S_1 corresponds to the slope of the region boundary in the reference-error plane. The region Ra is for when far-end single-talk is expected and the filter is thus updated by the UFLMS. The region Rb is for when double-talk is expected and the filter is thus updated by the FSSA. Though echo path changes may be covered by the region Rb, the FSSA can still adapt to them without serious loss of convergence rate. On the other hand, Fig. 2 (b) shows the form of segmentation used in several conventional approaches [8, 10, 11]. The regions are separated by time-variant boundaries that are perpendicular to the reference or error axis. Separating the double-talk region requires adequate and frequent boundary control.

If the maximum of the ACL is known or we can expect it to be bounded by a maximum level S_2, the region Rb in Fig. 2 (a) can be separated into Rb1 and Rb2 (Fig. 2 (c)). In the region Rb2, near-end single-talk is expected, while double-talk is expected in the region Rb1. Thus, by more strongly limiting the amount of updating in the region Rb2, the accuracy of convergence can be improved. An improved variant of the GL-FLMS is thus

$$g(|E_k(l)|,|X_k(l)|) =\psi_{S_1}\left(|E_k(l)|\frac{|X_k(l)|}{P_k(l)}\right)\psi_{\frac{1}{S_2}}\left(\frac{P_k(l)}{|E_k(l)||X_k(l)|}\right)S_2, \quad (7)$$

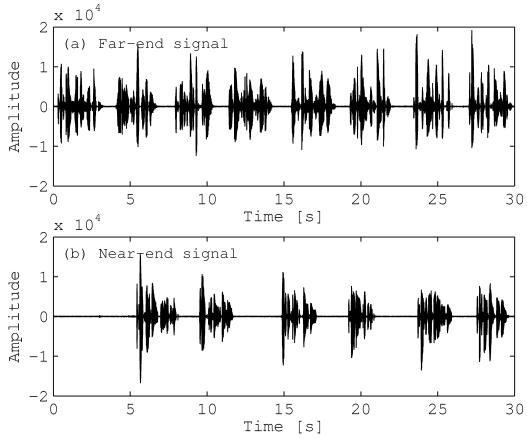

Fig. 3. Signals: (a) far-end and (b) near-end.

$$= \begin{cases} |E_k(l)|\dfrac{|X_k(l)|}{P_k(l)} & \text{(in the region Ra),} \\ S_1 & \text{(in the region Rb1),} \\ S_1 S_2 \dfrac{P_k(n)}{|E_k(l)||X_k(l)|} & \text{(in the region Rb2).} \end{cases} \quad (8)$$

4. SIMULATION

4.1. Comparisons with conventional algorithms

We compared the GL-FLMS based on (7) with the UFLMS based on (3) and some robust algorithms shown below.

4.1.1. Conventional method I

When we apply one conventional approach [8] to (3), we obtain

$$g(|E_k(l)|,|X_k(l)|) = |E_k(l)|\frac{P_k(l)|X_k(l)|}{P_k^2(l)+P_{th}^2(l)}, \quad (9)$$

where $P_{th}(l)$ is a time-variant threshold based on the estimated noise level. The noise level estimate is updated if the error level is above the filter's output level.

4.1.2. Conventional method II

The approaches [10, 11] correspond to

$$g(|E_k(l)|,|X_k(l)|) = \psi_{k_0(l)}\left(\frac{|E_k(l)|}{s(l)}\right)\frac{s(l)|X_k(l)|}{P_k(l)}, \quad (10)$$

where $k_0(l)$ is a constant, and the scale factor $s(l)$ is controlled with a DTD.

4.1.3. Results

The conditions were as follows. The desired echo was made by using a 1024-tap echo path with the average ACL of 0 dB. The sampling rate was 16 kHz. The smoothing factor α was 0.8 in all cases. For the GL-FLMS, $S_1 = 0.5$ and $S_2 = 2$. For conventional methods I and II, the parameters were basically those given in [8] and [11] respectively, although some of them, such as noise level estimating parameters, were rescaled to take the sampling rate, block size, or tap length into account. We chose the step-sizes to achieve the same steady-state echo return loss enhancement (ERLE) for a stationary signal input with averaged speech spectrum. White Gaussian noise was added to the echo as ambient noise with an overall echo-to-noise ratio (ENR) of 20 dB. As a result, $\mu = 0.2$ (UFLMS), $\mu = 0.23$ (conventional I), $\mu = 0.2$ (conventional II), and $\mu = 0.32$ (GL-FLMS), respectively.

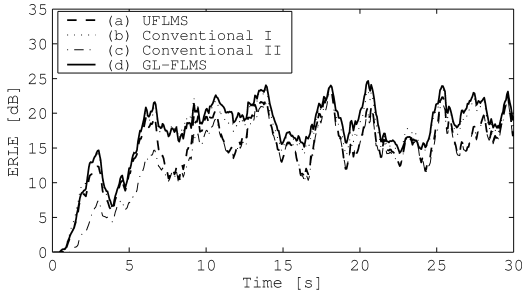

Fig. 4. Comparisons with different methods (single-talk).

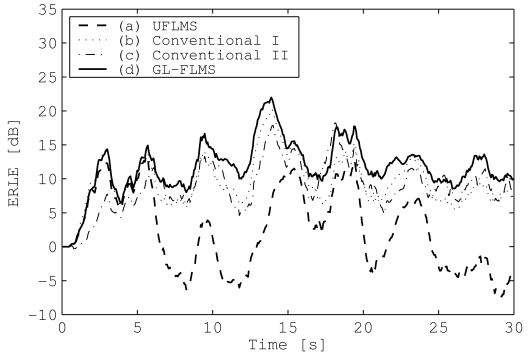

Fig. 5. Comparisons with different methods (double-talk).

The results for the speech signals were obtained in the following way. The far- and near-end signals (Fig. 3) were male and female speech, respectively. The average ENR was 20 dB. To ensure minimum stability for all algorithms, the adaptation was frozen when the far-end signal level was smaller than 12 dB below an average signal level. The echo path was changed at 3 seconds. Figure 4 shows the results for ERLE performance in the single-talk case where there was no near-end speech. The near-end signals were excluded in calculating the ERLEs. We can see that conventional method I and the GL-FLMS were more robust against ambient noise than the UFLMS and conventional method II. Figure 5 shows the results for ERLE performance in the double-talk case. The GL-FLMS showed the best robustness. We applied an 'ideal' DTD with advance detection of the near-end speech to conventional II only.

Although there might be other parameter choices with which the conventional methods would have performed better, note that the GL-FLMS displayed its robust performance even when the parameters were fixed and chosen in a simple and reasonable way.

4.2. Dependence on environmental conditions

The dependences of performance on the near- and far-end level balance and the ACL were examined. The conditions were the same as in 4.1, except as specified below.

To evaluate the dependence on level balance, three kinds of level combinations were tested by scaling the signals in Fig. 3 as follows: (a) the far-end signal by +6 dB and the near-end signal by -6 dB, (b) the far-end signal by -6 dB and the near-end signal by +6 dB, and (c) the original far- and near-end signals. Since the ambient noise level was unchanged, the ERLEs behaved differently according to the

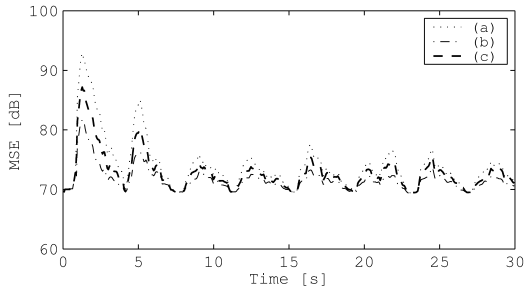

Fig. 6. Dependence on far-end signal level (single-talk).

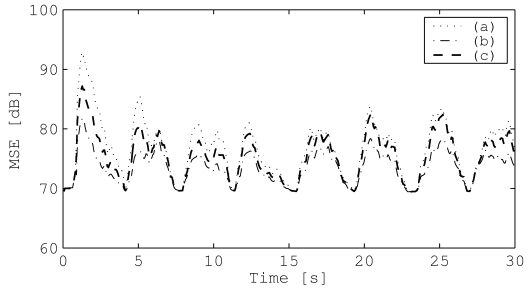

Fig. 7. Dependence on balance between far- and near-end signal levels.

ENRs. So the mean square error (MSE), as calculated with the near-end speech excluded, was used to compare the absolute residual echo levels. Figure 6 shows results obtained without the near-end signal. This indicates the dependence on reference input level during single-talk. The residual echoes converged to a similar steady-state level in all cases. Figure 7 shows results obtained in the double-talk case. In cases (a) and (b), despite the 12 dB of level difference, no notable degradation from case (c) was observed.

The dependence on ACL was evaluated by comparing the performance in these three cases: (a) average ACL = +10 dB, (b) average ACL = -10 dB, and (c) average ACL = 0 dB. The far- and near-end signals of Fig. 3 were used again. Figure 8 shows the MSEs obtained for the single-talk case. The differences between Fig. 6 and Fig. 8, especially in the steady-state, were due to the fixed reference threshold used to freeze adaptation for all of the simulations that involved speech. Figure 9 shows MSEs obtained in the double-talk case. Once the residual echo converged to the steady-state level, the ACL differences did not affect the robustness against double-talk.

5. CONCLUSIONS

We have proposed a double-talk robust frequency domain algorithm: the gradient-limited FLMS (GL-FLMS), which nonlinearly controls the sizes of updates according to the error-to-reference ratio. Unlike some conventional robust algorithms, the GL-FLMS achieves its robustness with fixed thresholds predetermined on the basis of the averaged level balance of the far- and near-end signals and the bounds on acoustic coupling level expected in real situations. The algorithm's effectiveness and reasonableness have been confirmed through simulation.

6. ACKNOWLEDGEMENTS

We would like to thank Dr. H. Ohara for fruitful discussions and helpful suggestions.

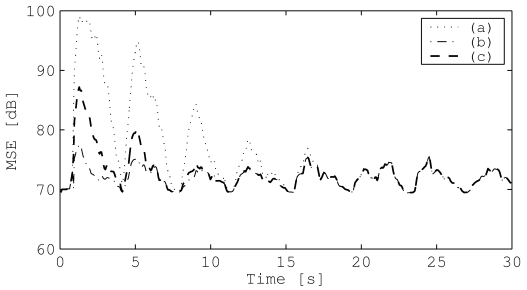

Fig. 8. Dependence on ACL (single-talk).

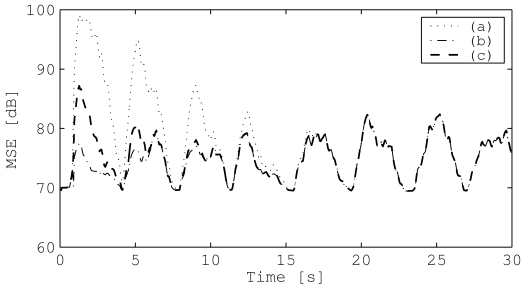

Fig. 9. Dependence on ACL (double-talk).

7. REFERENCES

[1] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. XLVI, no. 3 pp. 497-511, Mar. 1967.

[2] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceller with two echo path models," *IEEE Trans. Communications*, vol. COM-25, no. 6 pp. 589-585, Jun. 1977.

[3] P. J. Huber, *Robust statistics*. Wiley, New York, 1981.

[4] D. L. Duttweiler, "Adaptive filter performance with nonlinearities in the correlation multiplier," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 4, pp. 578-586, Aug. 1982.

[5] D. Mansour and A. H. Gray Jr., "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 5, pp. 726-734, Oct. 1982.

[6] A. Dasgupta, C. R. Johnson, Jr., and A. M. Baksho, "Sign-sign LMS convergence with independent stochastic inputs", *IEEE Trans. Information Theory*, vol. 36, no. 1, pp. 197-201, Jan. 1990.

[7] W. A. Sethares, "Adaptive algorithm with nonlinear data and error functions," *IEEE Trans. Signal Processing*, vol. 40, no.9, pp. 2199-2206, Sep. 1992.

[8] A. Hirano and A. Sugiyama, "A noise-robust stochastic gradient algorithm with an adaptive step size suitable for mobile hands-free telephones", *Proc. ICASSP95*, vol. 2, pp. 1392-1395, May 1995.

[9] Y. Haneda, S. Makino, J. Kojima, and S. Shimauchi, "Implementation and evaluation of an acoustic echo canceller using duo-filter control system," *Proc. EUSIPCO96*, vol. 2, pp. 1115-1118, Sep. 1996.

[10] T. Gänsler, "A robust frequency-domain echo canceller", *Proc. ICASSP97*, vol. 3, pp. 2317-2320, Apr. 1997.

[11] T. Gänsler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation", *IEEE Trans. Speech and Audio*, vol. 8, no. 6, pp. 656-663, Nov. 2000.

[12] S. Shimauchi, Y. Haneda, and A. Kataoka, "Study on frequency domain echo canceller based on higher order statistics", *Proc. Spring Mtg. Acoust. Soc. Japan*, pp. 615-616, Mar. 2003 (in Japanese).