A STEREO ECHO CANCELER WITH PRE-PROCESSING FOR CORRECT ECHO-PATH IDENTIFICATION

Yann Joncour and Akihiko Sugiyama

C&C Media Research Laboratories, NEC Corporation
1-1, Miyazaki 4-Chome, Miyamae-ku, Kawasaki, Kanagawa 216, Japan

ABSTRACT

A new stereo echo canceler with pre-processing for correct echo-path identification is proposed. The pre-processing is accomplished by a two-tap time-varying filter which delays the input signal periodically by one sample in one of the two channels. Aliasing components and audible clicks by pre-processing are made inaudible by selecting appropriate parameters for the filter. Simulations with the NLMS algorithm and a white Gaussian signal confirm the correct echo-path identification. For speech signals, the convergence speed of the proposed echo canceler is more than three times faster than that of an echo canceler with nonlinear transformations. Results of a subjective listening test demonstrate that quality of the pre-processed signals is 4.38 using the CCIR five-grade impairment scale. This is acceptable for general teleconference applications.

1. INTRODUCTION

Teleconferencing allows people at remote places to communicate as if they were in the same room. To improve the realism of this electronic meeting, a good localization of participants by means of stereo systems is required. Such systems include a stereo echo canceler to remove undesired acoustic echoes.

In the conventional structure, the four echo paths existing between the two loudspeakers and the two microphones are estimated by four adaptive filters which are linearly combined [1]. However the echo paths are not correctly identified for strongly cross-correlated input signals, like stereo speech signals [2]. The condition for which the echo is canceled leads to an infinite number of solutions to the filter coefficients. These coefficients misconverge to values which depend on the acoustic environment in the room where the talkers are located [3]. Consequently, any kind of acoustic changes in the remote room, such as talker movements, seriously degrade ERLE (Echo Return Loss Enhancement).

In practical situations, the input signals have low-level components which are not cross-correlated [4], and there are slight variations in the interchannel correlation [5]. Considering these two facts, the echo paths can be identified providing that the adaptation algorithm can deal with uncorrelated low-level components or small variations in the interchannel correlation. A strong interchannel correlation necessitates a fast-convergence algorithm. Such an algorithm requires much computation and its implementation is not easy.

Another solution consists in a partial decorrelation of input signals by introducing a non-linearity in each channel [6]. In spite of this decorrelation, the convergence of filter coefficients is still slow unless a fast-convergence algorithm is used.

This paper proposes a stereo echo canceler with pre-processing for correct echo-path identification. Section 2 reviews the conventional structure. The proposed structure is introduced in Section 3. In Section 4, simulation results show the convergence of filter coefficients. The audio quality of the pre-processed signals is evaluated by a subjective listening test in the last section.

2. CONVENTIONAL STRUCTURE

The shaded area labeled “conventional structure” in Fig. 1 represents the conventional stereo echo canceler. The symmetry allows us to consider only the echo received by the left microphone, $m_L(k)$. k is the time index for discrete signals. The structure related to this echo is shown with bold lines. The discussions remain the same for the other echo, $m_R(k)$.

In the discussions in this section, the adjustment of adaptive filters is not considered. Therefore, the impulse responses of the left and the right adaptive filters $w_{LL}(k)$ and $w_{RL}(k)$, respectively, are assumed to be time invariant. Let us define $h_{LL}(k)$ and $h_{RL}(k)$ as the left and the right echo-path impulse responses, respectively. The residual echo $e_L(k)$, is related to the left input signal $x_L(k)$, and the right input signal $x_R(k)$, by

$$\begin{aligned} e_L(k) &= [h_{LL}(k) - w_{LL}(k)] * x_L(k) \\ &\quad +[h_{RL}(k) - w_{RL}(k)] * x_R(k), \end{aligned} \tag{1}$$

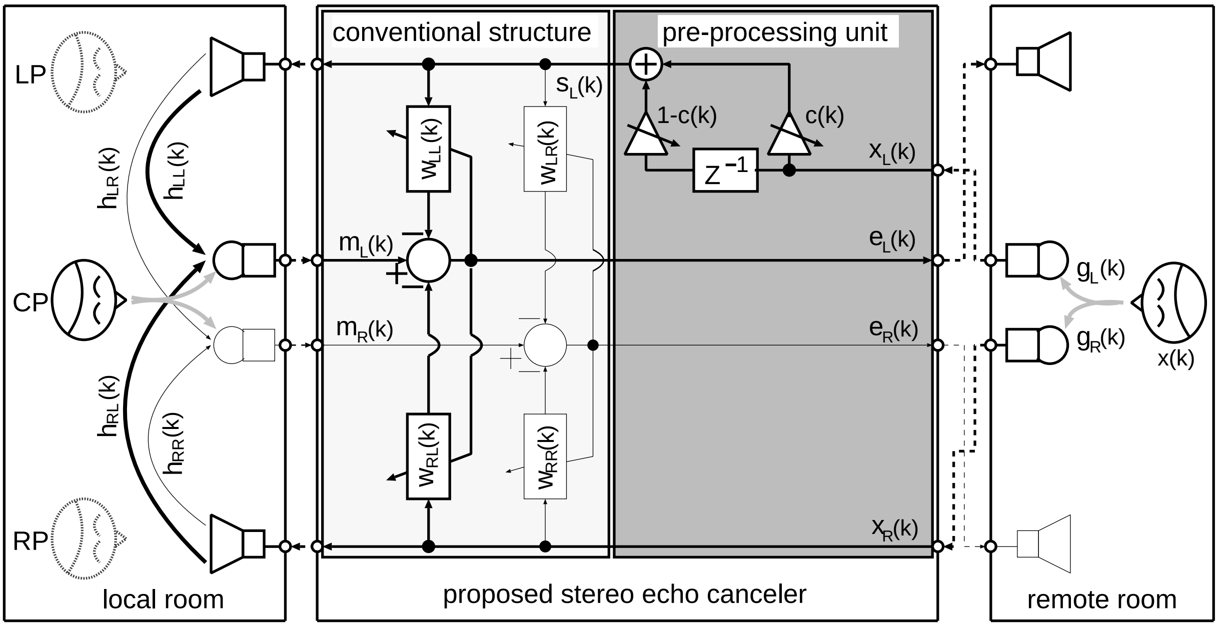

Fig. 1: The proposed stereo echo canceler for teleconferencing.

where the operator $*$ denotes the convolution between two discrete signals. The left and the right acoustic paths in the remote room are characterized by the impulse responses $g_L(k)$ and $g_R(k)$, respectively. For a single-talker configuration, a source signal $x(k)$, is convoluted with $g_L(k)$ and $g_R(k)$, to form the input signals $x_L(k)$ and $x_R(k)$, as follows:

$$x_L(k) = g_L(k) * x(k) \tag{2}$$

$$x_R(k) = g_R(k) * x(k). \tag{3}$$

Substituting (2) and (3) in (1) gives

$$\begin{aligned} e_L(k) &= [h_{LL}(k) - w_{LL}(k)] * g_L(k) * x(k) \\ &\quad +[h_{RL}(k) - w_{RL}(k)] * g_R(k) * x(k). \end{aligned} \tag{4}$$

The condition for which the echo is canceled, ie. $e_L(k) = 0$, leads to the following equation.

$$\begin{aligned} &[h_{LL}(k) - w_{LL}(k)] * g_L(k) = \\ &\quad [h_{RL}(k) - w_{RL}(k)] * g_R(k). \end{aligned} \tag{5}$$

There is an infinite number of solutions to (5), and therefore it does not imply that $w_{LL}(k) = h_{LL}(k)$ and $w_{RL}(k) = h_{RL}(k)$. It means that the correct echo-path identification is not achieved.

3. PROPOSED STRUCTURE

The proposed stereo echo canceler differs from the conventional one by a pre-processing which delays the input signal periodically in one of the two channels. The pre-processing is represented in Fig. 1 by the shaded area labeled "pre-processing unit". It is assumed that the pre-processing takes place in the left channel. It may be equipped with in the right channel instead. The pre-processing consists of a two-tap filter whose time-varying coefficients are controlled by a periodic function $c(k)$ with a period Q. When $c(k) = 1$ for the first Q/2 iterations, the processed signal $s_L(k)$, is identical to the input signal $x_L(k)$. When $c(k) = 0$ for the following Q/2 iterations, the processed signal is a one-sample delayed version of the input signal. These operations are repeated every Q iterations.

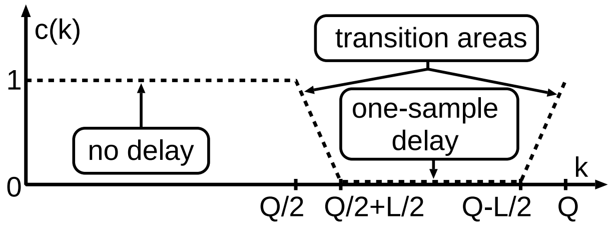

Fig. 2: c(k) with a period Q.

By this pre-processing, the condition for which the echo is canceled leads to two equations. For $c(k) = 1$, the equation derived from $e_L(k) = 0$ is expressed by (5), like for the conventional method. For $c(k) = 0$, a supplementary equation is derived from $e_L(k) = 0$, since the input signal is delayed by one sample. The common solution to these two equations is unique and corresponds to the true echo-path impulse responses [7]. The correct echo-path identification is achieved.

Naturally, listeners should not perceive any degradation of the processed signals. However, two kinds of degradations exist [8]. First, the processed signals contain aliasing

Tab. 1: Parameters for simulations.

Parameter	N	ENR	Q	L
White signal	64	40 dB	60	6
Speech signal	1000	40 dB	2000	200

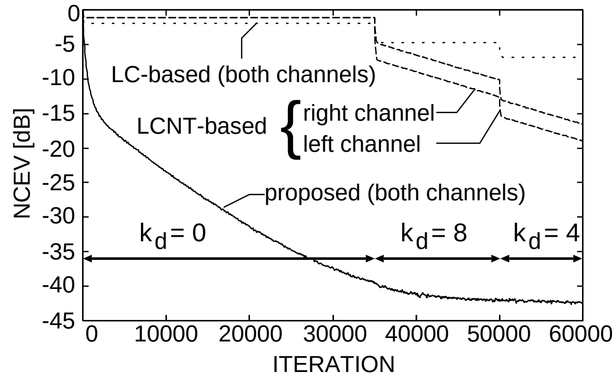

Fig. 3: Coefficient convergence. Average over 25 white Gaussian signals.

components. Secondly, “clicks” are heard when the input signal is suddenly delayed by one sample, or equivalently, when $c(k)$ moves from one to zero. The same phenomenon occurs when $c(k)$ varies from zero to one. The aliasing components can be made inaudible by selecting a long period for $c(k)$, controlled by Q [8]. To avoid the audible clicks, $c(k)$ varies smoothly between zero and one over L iterations, as depicted in Fig. 2, for a smooth transition.

4. SIMULATION RESULTS

Simulations were carried out for the structure based on linear combination (LC) [1], the one based on linear combination with nonlinear transformations (LCNT) [6], and the proposed structure. The parameters for simulations are given in Tab. 1. N stands for the length of adaptive filters. ENR (Echo-to-Noise Ratio) is defined as the ratio of the echo power before cancellation to the power of an independent white Gaussian noise added to the echo. Adaptation was performed with the normalized LMS (NLMS) algorithm with a step size μ of 0.5. For the LCNT-based structure, $\alpha = 0.5$ was selected as a reasonable trade-off between the convergence speed and the audio distortion [6].

A white Gaussian signal with zero mean and a unit variance was used for the input signals, $x_L(k)$ and $x_R(k)$. As an example of strong cross-correlation, $x_L(k)$ was a k_d-sample delayed version of $x_R(k)$. Modifications of k_d after 35000 and 50000 iterations simulated acoustic changes in the remote room. Q and L were selected to obtain a complete convergence within 60000 iterations. Another pair should be used from a viewpoint of audio quality. To show the convergence of filter coefficients, the norm of the coefficient-error vector (NCEV) was calculated in both channels.

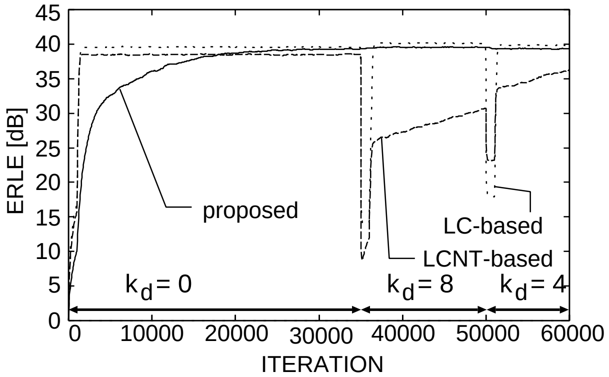

Fig. 4: ERLE. Average over 25 white Gaussian signals.

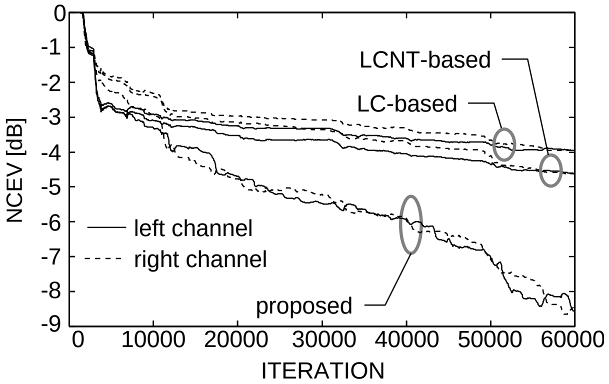

Fig. 5: Coefficient convergence for speech signals.

Fig. 3 shows that NCEV for the LC-based structure misconverges to a final value which depends on the acoustic environment in the remote room. The LCNT-based structure is not effective for $k_d = 0$, corresponding to the ultimate interchannel correlation. For the proposed structure, NCEV is not degraded by any acoustic change in the remote room, and it reaches the optimum value for the given ENR. The correct echo-path identification has been achieved.

Fig. 4 exhibits the behavior of ERLE. For the LC-based and LCNT-based structures, ERLE is degraded each time an acoustic change occurs, since the adaptive filters do not identify the echo paths correctly. For the proposed structure, ERLE is not degraded at all and reaches the optimum value for the given ENR.

NCEV for recorded speech signals sampled at 8 kHz and real echo-path impulse responses measured in a conference room are shown in Fig. 5. $Q = 2000$ and $L = 200$ were selected to obtain a good audio quality for the processed signals. NCEV is almost saturated at -4 dB for the LC-based structure and -4.6 dB for the LCNT-based structure. For the proposed structure, NCEV reaches around -8.5 dB within 60000 iterations and it keeps decreasing. After the 10000-th iteration, NCEV convergence for the proposed structure is three and half, and five times faster compared to the LCNT-based and LC-based structures, respectively.

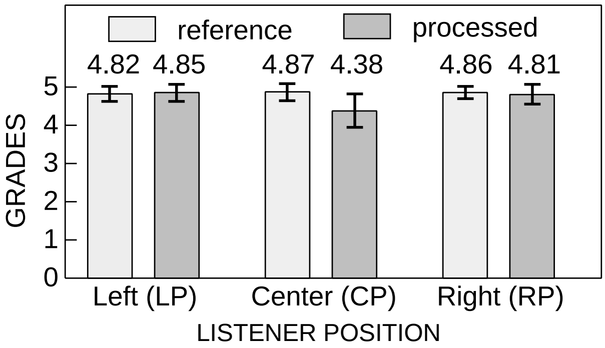

Fig. 6: Listening test results. Average over 6 excerpts.

5. LISTENING TEST RESULTS

The "triple stimulus/hidden reference/double blind approach" was used for the subjective listening test [9]. Listeners were asked to rate the perceived difference between the known reference (original signals) and two other versions corresponding to the hidden reference and the processed signals, using the CCIR[1] five-grade impairment scale [10]. Participants did not know which version was under test. 12 listeners occupied successively a left position (LP), a center position (CP), and a right position (RP) as depicted in Fig. 1. The test was carried out with 6 excerpts of female and male narrations, and vocal musics. $Q = 2000$ and $L = 200$ were selected to remove audible aliasing and clicks.

Fig. 6 shows the test results. The mean values of the grades are represented with a vertical bar. The vertical line centered around a mean value is the confidence interval performed on the 95% confidence level using a normal distribution. The impairment between the original signals and the pre-processed signals was perceived only at the center position, as depicted in Fig. 6. It is explained by a slight to-and-fro movement of the stereo sound image. However, the pre-processed signals have been scored 4.38 which is 0.5 lower than the original signals. This impairment is not annoying, and therefore, it is acceptable for teleconferencing. A DSP implementation and performance evaluation of the proposed stereo echo canceler also supports these results [11].

6. CONCLUSION

A new stereo echo canceler with pre-processing for correct echo-path identification has been proposed. A two-tap time-varying filter located in one of the two channels periodically delays the input signal by one sample. Aliasing components and audible clicks by pre-processing have been made inaudible by selecting appropriate parameters for the filter. Simulations with the NLMS algorithm have shown that the convergence speed of the proposed echo canceler is more than three times faster than that of an echo canceler with nonlinear transformations. A subjective listening test has shown that the processed signals were scored 4.38 using the CCIR five-grade impairment scale, which is satisfactory for general teleconferencing.

[1] CCIR has been changed into ITU-R.

ACKNOWLEDGMENTS

The authors would like to thank fellows of the Digital Signal Processing Technology Group, C&C Media Research Laboratories, NEC Corporation, for participating in the subjective listening test. They are indebted to Dr. Takao Nishitani, Deputy General Manager of Silicon Systems Research Laboratories, NEC Corporation, for his guidance and encouragement. They also express their thanks to Dr. Ozawa, a Senior Manager of C&C Media Research Laboratories, NEC Corporation, for his fruitful comments.

REFERENCES

[1] T. Fujii and S. Shimada, "A note on Multi-Channel Echo Cancellers," Technical Report of IEICE, CS84-178, pp. 7-14, Jan. 1985 (in Japanese).

[2] A. Hirano and A. Sugiyama, "Convergence Characteristics of a Multichannel Echo Canceller with Strongly Cross-Correlated Input Signals - Analytical Results-," Proc. of 6th DSP Symposium of IEICEJ, pp. 144-149, Nov. 1991.

[3] A. Hirano and S. Koike, "Convergence Analysis of a Stereophonic Acoustic Echo Canceller Part I: Convergence Characteristics of Tap Weights," Proc. of 11th DSP Symposium of IEICEJ, pp. 569-574, Nov. 1996.

[4] J. Benesty, F. Amand, A. Gilloire and Y. Grenier, "Adaptive Filtering Algorithms for Stereophonic Echo Cancellation," Proc. of ICASSP'95, pp. 3027-3030, May 1995.

[5] S. Shimauchi and S. Makino, "Stereo Projection Echo Canceller with True Echo Path Estimation," Proc. of ICASSP'95, pp. 3059-3062, May 1995.

[6] J. Benesty, D.R. Morgan and M.M. Sondhi, "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation," Proc. of ICASSP'97, pp. 303-306, Apr. 1997.

[7] Y. Joncour and A. Sugiyama, "A Unique and Strict Identification of the Echo Path Impulse Response in Stereo Echo Cancellation," Technical Report of IEICE, DSP96-100, pp. 17-24, Dec. 1996.

[8] Y. Joncour and A. Sugiyama, "A Stereo Echo Canceler with Correct Echo-Path Identification," Technical Report of IEICE, DSP97-1, IE97-1, pp. 1-8, Apr. 1997.

[9] S. Bergman, C. Grewin and T. Rydén, "The SR Report on The MPEG/Audio Subjective Listening Test Stockholm April/May 1991," ISO/IEC JTC1/SC2/WG11 MPEG 91/010, May 1991.

[10] CCIR Recommendation 562.

[11] Y. Joncour, A. Sugiyama, and A. Hirano "DSP Implementation and Performance Evaluation of a Stereo Echo Canceler with Pre-Processing," submitted to EUSIPCO-98.