A Stereophonic Echo Canceler Using Single Adaptive Filter

Shigenobu MINAMI
Information and Communication Systems Laboratory TOSHIBA Corp.
70 Yanagi-cho Saiwai-ku Kawasaki 252 JAPAN minami@cns.clab.toshiba.co.jp

Abstract

This paper proposes a new stereophonic echo canceler which can be realized by using single adaptive filter. Since most of conversations in telecommunication are single talking, Pseudo-Stereophonic Echo Canceler (PST-EC), which is a monaural echo canceler and has only been applied to pseudo-stereophonic speech sound produced by attenuating and delaying monaural sound, is applicable to stereophonic communication. To cope with double talking, stereophonic speech sound is classified into strongly correlated component (single talking) and less correlated component (double talking). Therefore, PST-EC and echo suppressors can be applied to the former and the latter component, respectively.

1. Introduction

In teleconference or multi-media terminals, stereophonic hands free communication contributes to enhance sound localization and fidelity. One of the issues to introduce it to telecommunication systems is echo control, which is a technique to avoid howling and speech quality degradation caused by acoustic coupling between microphone and speaker[1].

Echo canceler is well-known as a most promising approach to cope with these problems. The problem of stereophonic hands free applications, however, is that we need four times as many adaptive filters as monaural applications, since there are four echo paths between two speakers and two microphones. Slow convergence due to correlation between right and left speaker outputs [2] is another serious disadvantage of the conventional stereophonic echo canceler. In this paper, the author proposes a new approach to cope with the above problems.

2. Stereophonic Echo Canceler

Among the many researches carried out for stereophonic echo canceler[2-4], an echo canceler proposed by Hirano et. al. [4] is a practical approach. In [4] two adaptive filters were applied to two microphone outputs and the input of each filter was dynamically chosen between the two speaker outputs based on their power ratio. However, the echo canceler had a problem that cancellation performance deteriorated when far-end talker changed and double talking (more than two far-end talkers spoke at the same time) occurred . Another approach is the "Pseudo-Stereophonic Echo Canceler(PST-EC)[5]" proposed by the author. Though the canceler was designed only for pseudo-stereophonic speech sound [6] which was produced from monaural sound, it has a merit of monaural echo canceler, fast convergence and single adaptive filter implementation. The key of this paper is an extension of the PST-EC application to a real stereo environment without losing the inherent advantage of the PST-EC.

3. Pseudo-Stereophonic (PST) Speech Sound

A generic stereophonic sound production model for N talkers is shown in Fig.1 , where ith sound Si(z) is picked up by right and left microphones through transfer functions $F_{Ri}(z)$ and $F_{Li}(z)$ respectively . Microphone outputs denoted $X_R(z)$ and $X_L(z)$ are expressed by

$$X_R(z)=\sum_{i=0}^{N-1} F_{Ri}(z)S_i(z)+NS_R(z)$$
$$X_L(z)=\sum_{i=0}^{N-1} F_{Li}(z)S_i(z)+NS_L(z) \quad (1)$$

respectively, where $NS_R(z)$ and $NS_L(z)$ are room acoustic noise. By assuming the microphone outputs as input / output of a transfer function, we obtain cross-channel transfer function G(z) as,

$$G(z)=\frac{\sum_{i=0}^{N-1} F_{Li}(z)S_i(z)+NS_L(z)}{\sum_{i=0}^{N-1} F_{Ri}(z)S_i(z)+NS_R(z)} \quad (2)$$

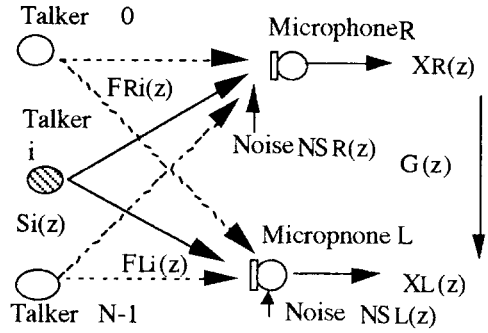

Fig.1 Stereophonic speech sound production model

0-7803-2431-5/95 $4.00

In usual conversations, the portion of double talking is relatively small when compared with single talking. Assuming the room acoustic noise is negligibly small, the cross-channel transfer function G(z) during single talking can be written as:

$$G(z)=\frac{F_L(z)}{F_R(z)} \qquad (3)$$

where $F_R(z)=F_{Ri}(z)$ and $F_L(z)=F_{Li}(z)$.

As evident from (3), the cross-channel transfer function during single talking is just a transfer function between sound source and microphone. Therefore, G(z) is stable as long as talker's location or room acoustic characteristics remain same. By multiplying G(z) by $X_R(z)(=F_R(z)S(z))$, $X_L(z)$ is given by

$$\begin{aligned} X_L(z)&=G(z)F_R(z)S(z) \\ &=F_L(z)S(z) \quad . \qquad (4) \end{aligned}$$

Further simplification is possible if the reverberation components of $F_R(z)$ and $F_L(z)$ are very small and are assumed to be direct wave responses, $D_R(z)$ and $D_L(z)$.

Then $F_R(z)$ and $F_L(z)$ are given by

$$\begin{aligned} F_R(z)&\cong D_{R(z)}=\phi(z,\delta_R)l_R Z^{-\tau_R} \\ F_L(z)&\cong D_{L(z)}=\phi(z,\delta_L)l_L Z^{-\tau_L} \end{aligned} \qquad (5)$$

where l_R and l_L are attenuation, τ_R and τ_L are integer delay values expressed by sample numbers, $\phi(z,\delta)$ is an anti-aliasing filter characteristics when the fraction part of d/T (d:delay, T: sampling period) is δ. If the delay d is completely divided by T, $\phi(z,0)$ becomes 1. Based on the above assumptions, the cross-channel transfer function for the direct wave, $\tilde{G}(z)$, is expressed by

$$\tilde{G}(z)=\varphi(z)lZ^{-\tau_d} \qquad (6)$$

where

$\varphi(z)=\phi_L(z,\delta_R)/\phi_R(z,\delta_L)$ $l=l_L/l_R$ $\tau_d=\tau_L-\tau_R$.

Similarly, predicted left signal $\hat{X}_L(z)$ can be obtained and is given by

$$\hat{X}_L(z)=\tilde{G}(z)X_R(z) \quad . \qquad (7)$$

This equation means that we predict one channel sound from another channel monaural sound and cross-channel transfer function. In Fig.2, the dotted line of EL(z) does not exist in this single talking case. Contrary to single talking, the prediction error denoted $E_L(z)$ has an actual value during double talking and is used to produce perfect stereophonic sound as

$$X_L(z)=\hat{X}_L(z)+E_L(z) \quad . \qquad (8)$$

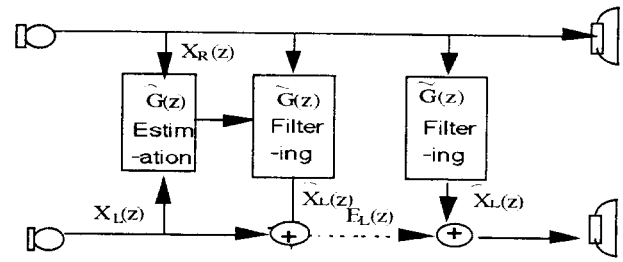

Fig. 2 Pseudo-Stereophonic Sound Principle

4. Decomposition and Composition Of Stereophonic Speech Sound

By applying the Pseudo-Stereophonic principle, we can decompose stereophonic sound into strongly correlated sound (single talk speech sound) and others. The process is shown in Fig.3, where one of the two transfer functions for the ith frame, $\hat{G}_{Ri}(z)$ and $\hat{G}_{Li}(z)$, is chosen to endorse causality .

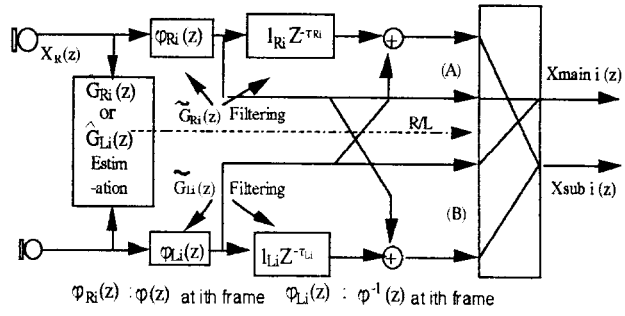

Fig.3 Decomposition of stereophonic sound

Switching is used so that outputs are independent from the causality selection R/L as

$$X_{\text{main i}}(z)=\begin{cases} \varphi_{Ri}X_R(z) \cdots \text{if } R/L=0 \\ \varphi_{Li}X_L(z) \cdots \text{if } R/L=1 \end{cases}$$

$$X_{\text{sub i}}(z)=\begin{cases} E_R(z) \cdots \text{if } R/L=0 \\ E_L(z) \cdots \text{if } R/L=1 \end{cases} \qquad (9)$$

where φ_{Ri} and φ_{Li} are effects of anti-aliasing filter.

In the composition process, right and left speaker outputs are obtained by using main signal Xmain i (z) and prediction error Xsub i (z) as shown in Fig.4 .

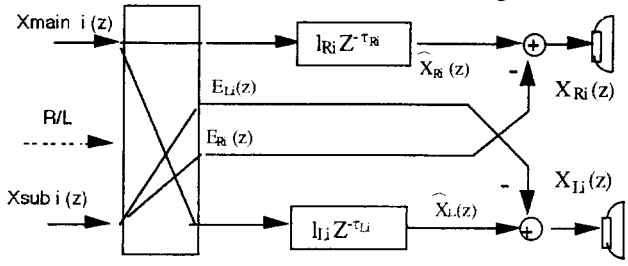

Fig. 4 Composition of stereophonic sound

5 Echo Canceler Using Single Adaptive Filter

5.1 Proposed System

Based on the decomposition of stereophonic speech sound, it becomes possible to realize stereophonic echo canceler by applying monaural echo canceler as shown in Fig. 5.

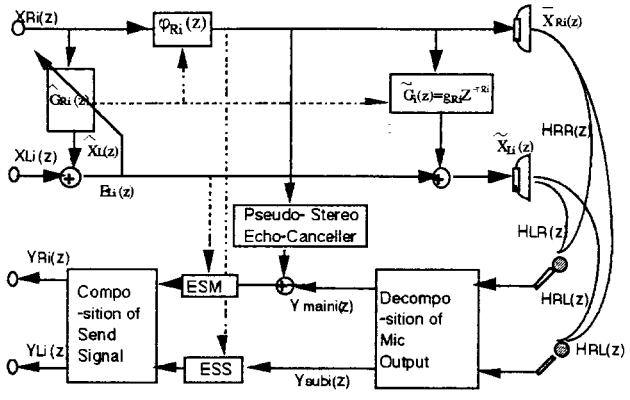

Fig. 5 Proposed Echo Canceler

In this system, PST-EC is inserted between the received signal Xmaini(z) and the transmitted main signal Ymaini(z) . Echo suppressor, ESM, in Fig.5, is adopted so as to reduce echo signals during double talking. Since cross-channel correlation cancel error , E_{Li} (z), is almost zero during single talking, the inserted attenuation of the echo suppressor ESS can be small. On the other hand, during double talking , the inserted loss is increased to cut the echo produced by the increased correlation cancel error $E_{Li}(z)$. Another echo suppressor, ESS, is applied to the uncorrelated sound $Y_{subi}(z)$ of the microphone outputs which is obtained by using the following operation.

$$\begin{bmatrix} Y_{main\,i}(z) \\ Y_{sub\,i}(z) \end{bmatrix} = \begin{bmatrix} g_{MRi} & g_{MLi} \\ g_{SRi} & -g_{SRi} \end{bmatrix} \begin{bmatrix} Y_{R\,i}(z) \\ Y_{L\,i}(z) \end{bmatrix}. \quad (10)$$

Here, weighting factors, g_{MRi} and G_{MLi} , are chosen so that the signal to noise ratio during single talking can be maxmized, and vice versa for g_{SRi} and G_{SLi} .

5.2 Echo Cancellation During Single Talking

During single talking, PST-EC can be applied to the proposed system (Fig.6). The echo canceler is designed to prevent cancellation ability degradation at change in echo path characteristics originated by the change of the delay and attenuation in, $\bar{G}_{Ri}(z)$, $\bar{G}_{Li}(z)$. PST-EC principle is briefly described as follows:

(1)Four transfer functions between speakers and microphones are calculated from the estimated transfer functions by the monaural echo canceler . This is possible since the echo path, including inserted loss and delay at speakers and microphone, is a linear combination of four echo path response between two speakers and two microphones with four sets of delay and loss control functions. The transfer function of the echo path is

$$\bar{H}_i(z)=g_{MRi}(\bar{G}_{Ri}(z)H_{RR}(z)+\bar{G}_{Li}(z)H_{LR}(z)) + g_{MLi}(\bar{G}_{Ri}(z)H_{RL}(z)+\bar{G}_{Li}(z)H_{LL}(z)). \quad (11)$$

(2) Similarly, the transfer function of, $\hat{H}_{i-3}(z)$, $\hat{H}_{i-2}(z)$, $\hat{H}_{i-1}(z)$, $\hat{H}_i(z)$, which are estimated by the monaural echo canceler at four consecutive frames, can be expressed in terms of four speaker and microphone estimated transfer functions, $\hat{H}_{RR}(z)$, $\hat{H}_{RL}(z)$, $\hat{H}_{LR}(z)$, $\hat{H}_{LL}(z)$, as :

$$\hat{\mathbf{H}}_i(z)=\mathbf{L}_i(z)\bar{\mathbf{H}}(z) \quad (12)$$

where $\hat{\mathbf{H}}_i(z)=(\hat{H}_i(z),\hat{H}_{i-1}(z),\hat{H}_{i-2}(z),\hat{H}_{i-3}(z))^T$

$$\bar{\mathbf{H}}(z)=(\hat{H}_{RR}(z),\hat{H}_{RL}(z),\hat{H}_{LR}(z),\hat{H}_{LL}(z))^T$$

$$\mathbf{L}_i(z)=\begin{bmatrix} A_i(z) & B_i(z) & C_i(z) & D_i(z) \\ A_{i-1}(z) & B_{i-1} & C_{i-1}(z) & D_{i-1}(z) \\ A_{i-2}(z) & B_{i-2} & C_{i-2}(z) & D_{i-2}(z) \\ A_{i-3}(z) & B_{i-3} & C_{i-3}(z) & D_{i-3}(z) \end{bmatrix}$$

$A_i(z)=g_{MRi}\bar{G}_{Ri}(z)$, $B_i(z)=g_{MRi}\bar{G}_{Li}(z)$

$C_i(z)=g_{MLi}\bar{G}_{Ri}(z)$, $D_i(z)=g_{MLi}\bar{G}_{Li}(z)$.

(3) Four estimated speaker-microphone transfer functions are obtained by

$$\bar{\mathbf{H}}(Z)=\mathbf{L}_i^{-1}\ \hat{\mathbf{H}}_i(z). \quad (13)$$

(4) From (13), initial value of the tap coefficients at delay and loss functions change due to far-end talker's change is

$$\hat{\mathbf{H}}_{i+1}(z)=\mathbf{L}_{i+1}(z)\ \bar{\mathbf{H}}(z). \quad (14)$$

where $\mathbf{L}_{i+1}(z)=(A_{i+1}(z),B_{i+1}(z),C_{i+1}(z),D_{i+1}(z))^T$.

6. Simulation Results

Computer simulations were carried out to evaluate the basic performance of the proposed method. In the simulation system shown in Fig. 6, the following conditions were used.

(1) Room acoustic characteristics were simulated assuming that talker A and B spoke alternatively every 50 frames (300 samples for each frame) without any double talking. Reverberation time of the room was assumed as 100msec and direct wave and reverberation wave ratio was assumed as 30dB at the talker's location.

(2) $\hat{G}_{Ri}(z)$ and $\hat{G}_{Li}(z)$ were estimated by 64 tap LMS adaptive filters. One of the filters was used according to talker's location. Main tap coefficient of the adaptive

filter was copied to the correlation cancellation filter and correlation add filter.

(3) 256 tap PST-EC was applied to stereophonic echo paths with 10 msec reverberation time.

(4)ERLE is an echo and cancellation error power ratio (dB) and LRES is a residual echo power(dB).

(5)Gaussian noise and speech were used as signal source.

As can be seen from Fig.8, with the proposed method, degradation in conventional echo canceler (see Fig. 7) at the change in echo path due to far-end talker's movement was prevented in the case of white Gaussian noise source. Same results were attained in the case of speech (Fig.9 and Fig.10). It should be noted that convergence speed of E-PST-EC is as fast as monaural echo canceler as long as far-end talker remains at the same position.

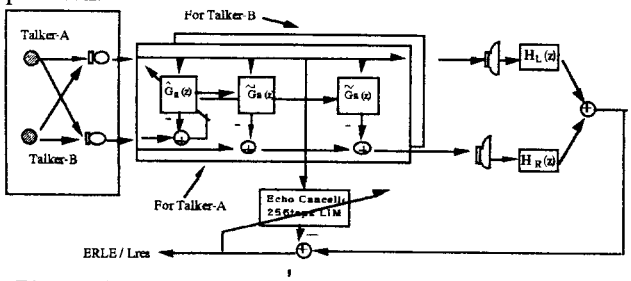

Fig.6 Simulation System Configuration

7. Conclusions

A new stereophonic echo canceler was proposed. Since single adaptive filter is used, cost increase for introducing stereophonic speech can be drastically reduced without suffering performance degradation due to correlation of the two microphone out put signals.

[References]

[1] H. Oikawa, S.Minami, " A new echo canceler realized by high performance DSP", IEEE, ISCAS'88 R66.1 June, 1988

[2] T. Fujii and S.Shimada, "A Note on Multi-Cannel Echo Cancelers," technical report of ICICE on CS, pp 7-14, Jan. 1984 (in Japanese)

[3] M.M.Dondhi and D.R. Morgan,"Acoustic Echo Cancellation for Stereophonic Teleconferencing", Proc. of Workshop on Applications of Signal Processing to Audio and Acoustics, May 1991

[4] A.Hirano and A.Sugiyama, "DSP Implementation and Performance Evaluation of a Compact Stereo Echo Canceler", Proc. of IEEE ICASSP'94, vol.2, pp II-245-248, Apr. 1994

[5] S.Minami,"An Acoustic Echo Canceler for Pseudo Stereophonic Voice", IEEE GLOBCOM'87 35.1 Nov. 1987

[6] S.Minami, " A stereophonic Voice Coding Method for teleconferencing", IEEE
ICCC. 86, 46.6, June 1986

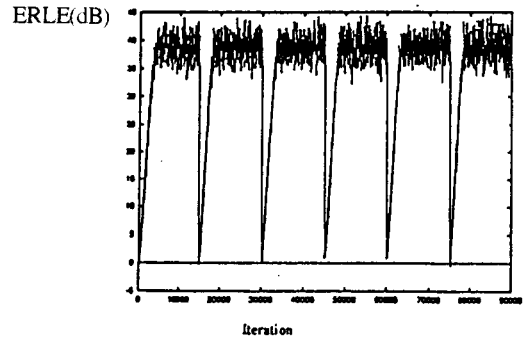

Fig.7 ERLE(dB) degrades at talker's change
(White Gaussian Noise, Conventional Echo Canceler)

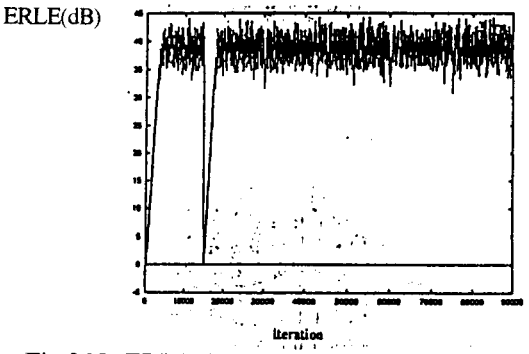

Fig.8 No ERLE(dB) degradation was observed at talker's change (White Gaussian Noise, Proposed Echo Canceler)

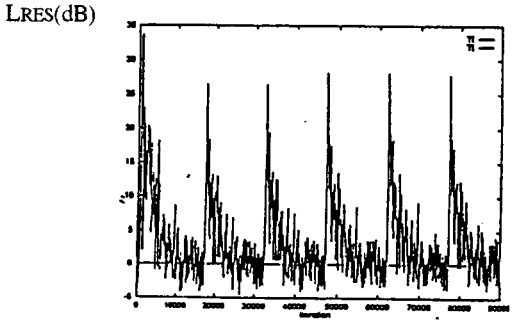

Fig.9 LRES (dB) increased at talker's change
(Speech, Conventional Echo Canceler)

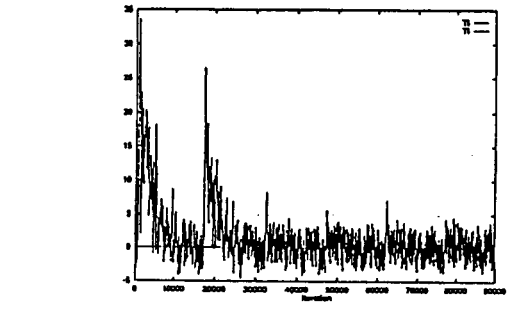

Fig.10 No LRES (dB) increase at talker's change
(Speech, Proposed Echo Canceler)