Acoust. Sci. & Tech. **22**, 5 (2001)

REVIEW

Stereophonic acoustic echo cancellation: An overview and recent solutions

Shoji Makino

NTT Communication Science Laboratories,
2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0237 Japan

(*Received 22 August 2000, Accepted for publication 8 January 2001*)

Keywords: Echo cancellation, Stereo system, Adaptive filter, Adaptive algorithm, Teleconferencing system

PACS number: 43.60.−c, 43.60.Lq

1. INTRODUCTION

A stereo teleconferencing system provides a more realistic presence compared to monaural systems. It helps listeners distinguish who is talking at the other end by means of spatial information. In such hands-free systems, stereophonic acoustic echo cancellers are absolutely necessary for full-duplex communication. The most significant problem with stereo echo cancellation using the conventional linear combiner structure is that the adaptive filter often misconverges or, even when it converges, the convergence is very slow because of the strong crosscorrelation between the stereo signals [1]. As a result, the conventional stereo echo canceller suffers from variation in both the near-end echo path and the far-end transmission path. The adaptive algorithm must track variations in not only the receiving room but also the transmission room. Accordingly, the performance of the stereo echo canceller degrades at the instant of any abrupt change in the environment in the transmission room. The difficult but important challenge is to devise a stereophonic acoustic echo canceller that converges independently of variations in the transmission room. For this aim, it is necessary for a stereo echo canceller to identify the true echo path impulse response quickly with low computational complexity. In this paper, this fundamental problem of stereophonic acoustic echo cancellation is discussed and recent solutions are reviewed.

2. FUNDAMENTAL PROBLEM OF STEREO ECHO CANCELLATION

2.1. Stereo Echo Cancellation

Stereo echo cancellation is achieved by linearly combining stereo signals (Fig. 1). Input signal vectors $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ and filter coefficient vectors $\hat{\boldsymbol{h}}_1(k)$ and $\hat{\boldsymbol{h}}_2(k)$ are combined as $\boldsymbol{x}(k) = [\boldsymbol{x}_1^T(k), \boldsymbol{x}_2^T(k)]^T$ and $\hat{\boldsymbol{h}}(k) = [\hat{\boldsymbol{h}}_1^T(k), \hat{\boldsymbol{h}}_2^T(k)]^T$. The combined filter coefficient vector $\hat{\boldsymbol{h}}(k)$ is updated by an adaptive algorithm. Thus, stereo echo cancellation is achieved by linearly combining two monaural echo cancellers. If the two input signals are obtained by filtering from a fixed common source, input signal vectors $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ have a strong and fixed crosscorrelation.

2.2. Non-uniqueness Problem

Unlike monaural echo cancellation, stereo echo cancellation has the specific problem of non-uniqueness. That is, the filter coefficient does not converge to the true echo path impulse responses.

Minimization of the weighted least squares criterion,

$$J(k) = \sum_{l=1}^{k} \lambda^{k-l} e^2(l), \tag{1}$$

leads to the normal equation

$$\boldsymbol{R}(k)\hat{\boldsymbol{h}}(k) = \boldsymbol{r}(k), \tag{2}$$

where

$$\boldsymbol{R}(k) = \sum_{l=1}^{k} \lambda^{k-l} \boldsymbol{x}(l)\boldsymbol{x}^T(l) = \begin{bmatrix} \boldsymbol{R}_{11}(k) & \boldsymbol{R}_{12}(k) \\ \boldsymbol{R}_{12}(k) & \boldsymbol{R}_{22}(k) \end{bmatrix} \tag{3}$$

$$\boldsymbol{r}(k) = \sum_{l=1}^{k} \lambda^{k-l} y(l)\boldsymbol{x}(l). \tag{4}$$

If input signals $x_1(k)$ and $x_2(k)$ are denoted as

$$x_1(k) = g_1(k) * s(k) \quad \text{and} \quad x_2(k) = g_2(k) * s(k), \tag{5}$$

where $g_1(k)$ and $g_2(k)$ are impulse responses between a talker and two microphones in the far-end transmission room, $s(k)$ is talker's speech, and $*$ denotes convolution, and if there is no noise and $g_1(k)$ and $g_2(k)$ are time-

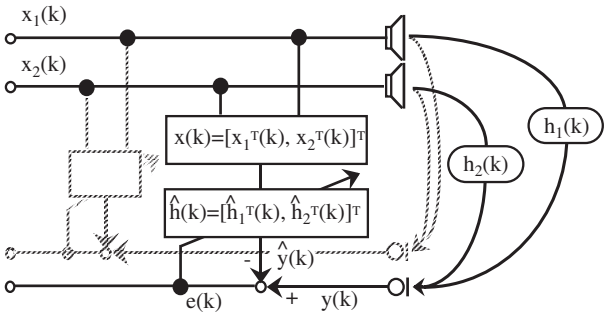


Fig. 1 Stereo echo canceller configuration.

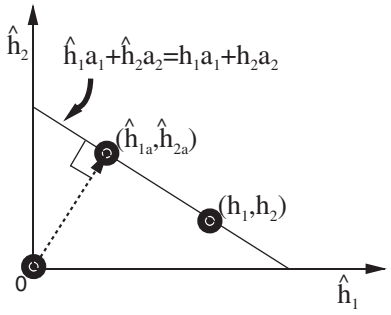


Fig. 2 Effect of crosscorrelation on the steady-state solution.

invariant, the input signal covariance matrix (3) is not full rank. Then, (2) has an infinite number of solutions. However, $\hat{\boldsymbol{h}}(k)$ is ''uniquely'' determined in the sense of a minimum-norm solution; by minimizing (1), the steady-state solution $[\hat{\boldsymbol{h}}_1^T(k), \hat{\boldsymbol{h}}_2^T(k)]$ converges to the point in subspace $\boldsymbol{H}_x$ nearest the initial point, where $\boldsymbol{H}_x$ is uniquely determined by the crosscorrelation between $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$. This does not mean that $\hat{\boldsymbol{h}}(k)$ equals $\boldsymbol{h}(k)$.

The problem of bad misalignment rarely appears in the single-channel case, although the autocorrelation is very high. This is because the autocorrelation of speech is time-variant with the consonants and vowels of words. (The bad misalignment only appears in the single-channel case for a long vowel period, where the autocorrelation is time-invariant. In this period, the error can be small even if the coefficient error is large and the adaptive filter fails to converge to the true echo path.) In the two-channel case, on the other hand, the crosscorrelation of the stereo signals is fixed when $g_1(k)$ and $g_2(k)$ are time-invariant. Therefore, the covariance matrix is severely ill-conditioned and the misalignment is much worse in the two-channel case.

For the simplest example, if input stereo signals $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ are denoted as

$$\boldsymbol{x}_1(k) = a_1\boldsymbol{s}(k) \quad \text{and} \quad \boldsymbol{x}_2(k) = a_2\boldsymbol{s}(k), \tag{6}$$

where a_1 and a_2 are scalar constant values and $\boldsymbol{s}(k)$ is a source signal vector, and if the initial value $\hat{\boldsymbol{h}}(0)$ is set to the zero vector, subspace $\boldsymbol{H}_x$ corresponds to a line for convenience, as shown in Fig. 2, and filter coefficients $\hat{\boldsymbol{h}}_1(k)$ and $\hat{\boldsymbol{h}}_2(k)$ converge to

$$\hat{\boldsymbol{h}}_{1a}(k) = \frac{{a_1}^2}{{a_1}^2 + {a_2}^2}\left[\boldsymbol{h}_1(k) + \frac{a_2}{a_1}\boldsymbol{h}_2(k)\right] \neq \boldsymbol{h}_1(k) \tag{7}$$

$$\hat{\boldsymbol{h}}_{2a}(k) = \frac{{a_2}^2}{{a_1}^2 + {a_2}^2}\left[\frac{a_1}{a_2}\boldsymbol{h}_1(k) + \boldsymbol{h}_2(k)\right] \neq \boldsymbol{h}_2(k). \tag{8}$$

2.3. Performance of Conventional Stereo Echo Canceller

Figure 3 shows the performance of a conventional stereo NLMS echo canceller for input stereo signals with a fixed crosscorrelation. The signals were made from a monaural speech signal $s(k)$ by convolving two different impulse responses g_1 and g_2 in the computer.

It is possible to have good echo cancellation even when misalignment is large. With the fixed-crosscorrelation input signals, the echo return loss enhancement (ERLE) reaches at least 30 dB [Fig. 3(a)]. On the other hand, as shown in Fig. 3(b), the coefficient error convergence levels off after a few dB modification. This is due to incorrect estimation of

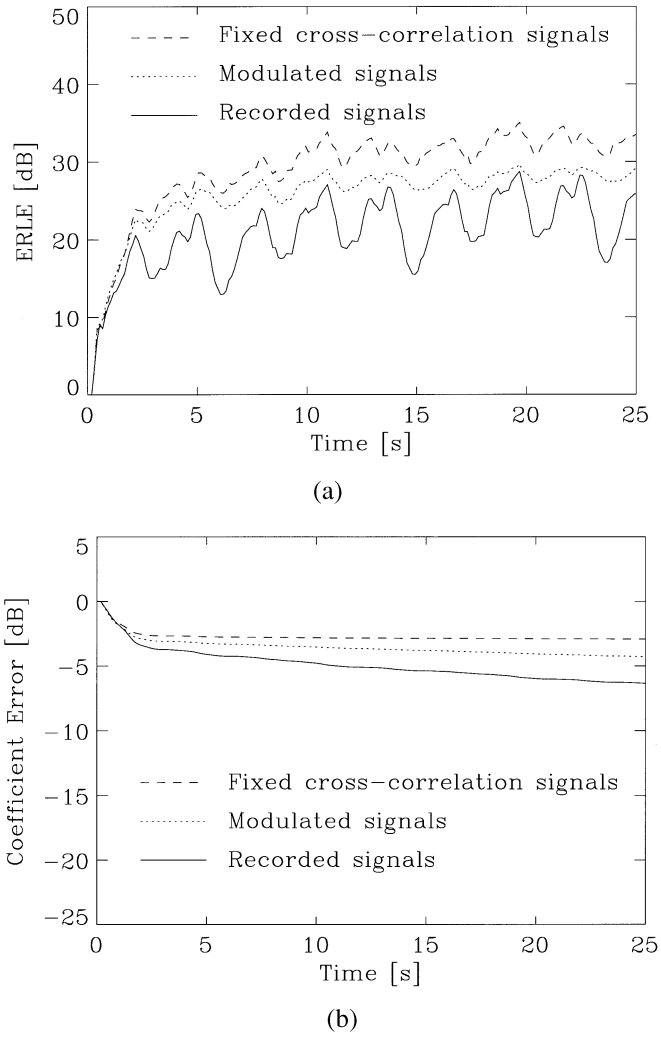


Fig. 3 Convergence of a conventional stereo NLMS echo canceller for speech input.

the true echo path impulse responses. In such a case, the cancellation will degrade if $g_1(k)$ and $g_2(k)$ change.

Note that with the noise-free fixed-crosscorrelation input signals, no adaptive algorithm, including the RLS algorithm, can achieve true echo path estimation.

3. CLUE FOR TRUE ECHO PATH ESTIMATION

A clue to solving the non-uniqueness problem can be found in practical teleconferencing situations.

Figure 3 also shows the performance of a conventional stereo NLMS algorithm with speech signals recorded with two microphones in a conference room while the speaker remained in one place. Note that the coefficient errors with the recorded signals decreases slightly in Fig. 3(b).

There are at least three clues to solving the non-uniqueness problem.

3.1. Independent Noise

In the real world, stereo signals $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ contain independent noise. These noise signals adapt the filters toward convergence.

3.2. Impulse Response Tail

If the length of impulse responses $g_1(k)$ and $g_2(k)$ in the transmission room is longer than that of the adaptive filters $\hat{\boldsymbol{h}}_1(k)$ and $\hat{\boldsymbol{h}}_2(k)$, the impulse response tail in the transmission room (truncated components) acts as independent noise. These noise signals adapt the filters toward convergence [2].

3.3. Variations in the Crosscorrelation

The crosscorrelation between stereo signals $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ varies slightly even when the talker does not move his body or head while speaking.

One might think that the change in the variation would cause another misconvergence and hence would not suppress the non-uniqueness problem. Fortunately, however, a ''new'' convergence process starts from the ''old'' misconverged solution.

Consider the case where the crosscorrelation between input signals $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ varies, e.g., a_1 and a_2 in Eq. (6) vary respectively to b_1 and b_2 (Fig. 4). First, $[\hat{\boldsymbol{h}}_1^T(k), \hat{\boldsymbol{h}}_2^T(k)]$ converges to $[\hat{\boldsymbol{h}}_{1a}^T(k), \hat{\boldsymbol{h}}_{2a}^T(k)]$, as shown in Eqs. (7) and (8). Then it converges to $[\hat{\boldsymbol{h}}_{1b}^T(k), \hat{\boldsymbol{h}}_{2b}^T(k)]$, the point nearest the ''initial'' point $[\hat{\boldsymbol{h}}_{1a}^T(k), \hat{\boldsymbol{h}}_{2a}^T(k)]$. Consequently, the norm of filter coefficient error vector $\boldsymbol{e}_b$ becomes smaller than the norm of $\boldsymbol{e}_a$. The coefficient error between $\boldsymbol{h}(k)$ and $\hat{\boldsymbol{h}}(k)$ becomes smaller with every variation in the crosscorrelation between the stereo signals.

Thus, after many variations in the crosscorrelation, the stereo echo canceller can converge to the ''true'' solution [3].

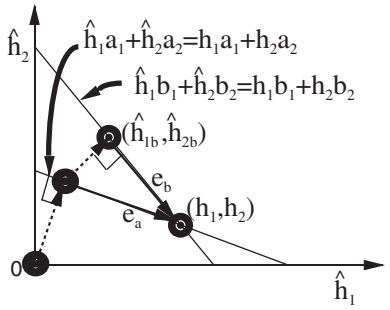


Fig. 4 Effect of variation in the crosscorrelation between stereo signals.

4. RECENT SOLUTIONS

4.1. Addition of Independent Noise and Variations in Crosscorrelation

Recently, several methods for overcoming the non-uniqueness problem have been proposed (Fig. 5). The function block in Fig. 5 is for adding the independent noise and variations in the crosscorrelation to stereo signals [3]. Some functions were successfully applied to create an independent component in the stereo input signals [4–7] while others were successfully applied to create an effective variations in the crosscorrelation between stereo input signals [8,9]. These functions are especially necessary when receiving fixed crosscorrelation stereo signals, e.g., those generated by a mixing machine. The important point is that the noise and the variations generated should not be audible and should not degrade stereo perception.

The frequency characteristics of the human auditory system are not flat, and speech signals mask the distortion. For this reason, simultaneous masking and temporal masking in the human auditory system are utilized in the function block. In simultaneous masking, a large frequency component will mask smaller ones in a nearby frequency band, whereas in temporal masking, components just before or right after a large audio component are masked. Therefore, if we properly adjust the additional noise and variations, we can make the processed distortion less audible and make the stereo echo canceller converge faster. Consequently, convergence speed to the true echo path

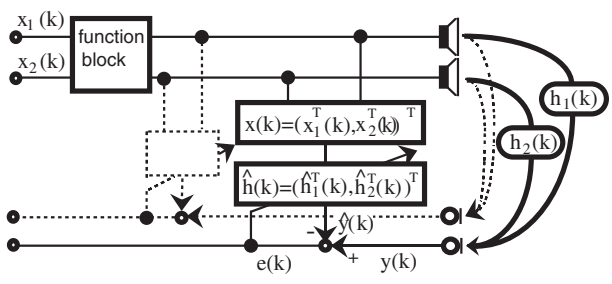


Fig. 5 Configuration of a stereo echo canceller with a function block.

impulse response is significantly improved.

4.2. Nonlinear Processing

Nonlinear processing is used to produce an independent noise [4], such that

$$\boldsymbol{x}'(k) = \boldsymbol{x}(k) + \alpha f[\boldsymbol{x}(k)], \tag{9}$$

where f is a nonlinear function, such as a simple half-wave rectifier. Since this noise synchronizes with the original speech, it is easily masked by the original speech. Such a transformation reduces the condition number of the covariance matrix, thereby reducing the misalignment. With a reasonable value of α, this distortion is hardly audible in typical listening situations and does not affect stereo perception. The authors have tried this method and confirmed the results with subjective listening tests.

The nonlinearly processed signals are input to exclusive adaptive filters in Ref. [5]. The exclusive adaptive filters converge to the true solution without suffering from crosscorrelation between the original stereo signals.

4.3. Noise Shaping

The effect of noise masking by the human auditory system in perceptual audio coding (MPEG audio coder) has been used.

Taking advantage of human auditory simultaneous masking properties, Ref. [6] proposed adding to each channel a random noise spectrally shaped so as to be masked by the presence of stereo signals. To achieve proper masking, the levels of each additive auxiliary noise is controlled carefully so that the noises are inaudible and the stereo perception is unchanged. Reference [7] proposed utilizing a perceptual audio coder that introduces in frequency and time a quantization noise that is below the hearing threshold. It is also possible to add inaudible noise, if the coder-introduced quantization noise is below the global perceptual masking level.

Due to these inaudibly shaped quantization noises, convergence to the true echo path is ensured.

4.4. Decorrelation Filters

If signals $\boldsymbol{x}_1(k)$ and $\boldsymbol{x}_2(k)$ are uncorrelated, then complete alignment is achieved. Thus, decorrelating filters have been proposed to decorrelate the two signals such that

$$\boldsymbol{x}'_1(k) = \boldsymbol{x}_1(k) - f_2(k)\boldsymbol{x}_2(k) \tag{10}$$

$$\boldsymbol{x}'_2(k) = \boldsymbol{x}_2(k) - f_1(k)\boldsymbol{x}_1(k). \tag{11}$$

Early trials of the decorrelation filters [1,10] failed since decorrelated signals $\boldsymbol{x}'_1(k)$ and $\boldsymbol{x}'_2(k)$ are themselves filtered versions of the same signal $s(k)$. Therefore, if there is no noise and crosscorrelation is fixed, ''perfect'' decrosscorrelation results in $\boldsymbol{x}'_1(k) = 0$ and $\boldsymbol{x}'_2(k) = 0$. With these signals, the adaptive filters cannot continue stable adaptation. (Perfect de-crosscorrelation cannot be easily achieved since we would need the inverse of $g_1(k)$ and $g_2(k)$ to calculate $f_1(k)$ and $f_2(k)$. However, $g_1(k)$ and $g_2(k)$ have no stable inverse in the real world.)

If there is independent noise and variations in crosscorrelation, decorrelated signals $\boldsymbol{x}'_1(k)$ and $\boldsymbol{x}'_2(k)$ become orthogonal to $\boldsymbol{x}_2(k)$ and $\boldsymbol{x}_1(k)$, respectively. However, $\boldsymbol{x}'_1(k)$ is not orthogonal to $\boldsymbol{x}'_2(k)$.

$$\boldsymbol{x}'_1(k) = \boldsymbol{x}_1(k) - \frac{\boldsymbol{x}_2^T(k)\boldsymbol{x}_1(k)}{\boldsymbol{x}_2^T(k)\boldsymbol{x}_2(k)}\boldsymbol{x}_2(k) \tag{12}$$

$$\boldsymbol{x}'_2(k) = \boldsymbol{x}_2(k) - \frac{\boldsymbol{x}_1^T(k)\boldsymbol{x}_2(k)}{\boldsymbol{x}_1^T(k)\boldsymbol{x}_1(k)}\boldsymbol{x}_1(k) \tag{13}$$

The adjustment vectors $\boldsymbol{x}'_1(k)$ and $\boldsymbol{x}'_2(k)$ should be orthogonal to input vectors $\boldsymbol{x}_2(k)$ and $\boldsymbol{x}_1(k)$, respectively, especially ''after'' convergence. Also, the decorrelation filter removes the correlated speech that disrupts the adaptation. Therefore, this method is ''effective'' in emphasizing the noise and variations. Note that this decorrelation tends to suffer from instability at the beginning of convergence.

On the other hand, in the de-autocorrelation filter for a single-channel case (even for a stereo case), the adjustment vector $\boldsymbol{x}'(k)$ should be orthogonal to the input vector $\boldsymbol{x}(k-1)$, (as will be shown in section 5.3):

$$\boldsymbol{x}'(k) = \boldsymbol{x}(k) - \frac{\boldsymbol{x}^T(k-1)\boldsymbol{x}(k)}{\boldsymbol{x}^T(k-1)\boldsymbol{x}(k-1)}\boldsymbol{x}(k-1). \tag{14}$$

4.5. Time-varying All-pass Filtering

Time-varying single-pole all-pass filtering of each stereo signal was proposed in Ref. [8]. The amount of time-variation is restricted within the just-noticeable interaural delay in psychoacoustics to maintain spatial perception. Due to this variation in the crosscorrelation, it is possible to achieve perfect alignment between the adaptive filters and the true echo path.

4.6. Time-varying One-sample Delay Filtering

A two-tap time-varying filter was proposed in Ref. [9]. This filter delays the input signal periodically by one sample in one of two channels. Aliasing components and audible clicks produced by the time-varying filter are made inaudible by selecting appropriate parameters for the filter. Due to the variation in the crosscorrelation, the correct echo path identification is achieved. Subjective listening tests showed that the center listening position is the most sensitive and listeners at this position perceive acceptable impairment between the original signals and the processed signals. We tried this method and confirmed the results with subjective listening tests.

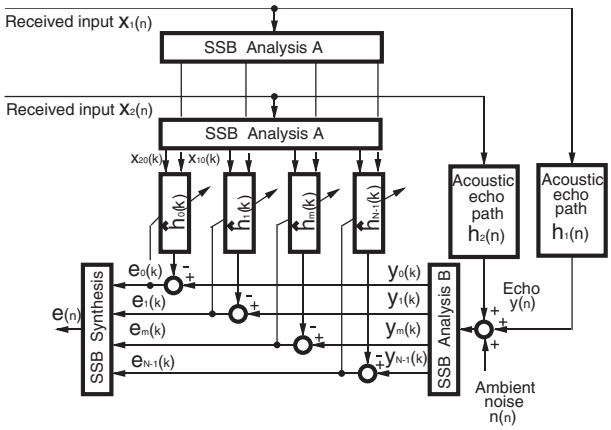


Fig. 6 Subband stereo echo canceller.

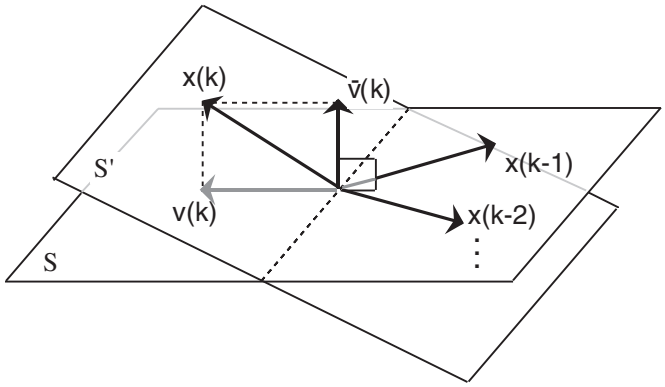


Fig. 7 Geometric interpretation of subspace determined by crosscorrelation.

4.7. Comparisons

The additive noise causes degradation of speech quality while additive variations cause degradation of stereo perception. Also, there are trade-offs between the convergence speed and subjective impairment.

The frequency response of the variations should be so designed as not to reduce any frequency components which $g_1(k)$ and $g_2(k)$ pass [11].

4.8. Subband Processing

The configuration of the subband stereo echo canceller is shown in Fig. 6. In the subband structure, signals are divided into N smaller frequency subbands and downsampled by downsampling rate R. As a result, the sampling interval becomes longer than that of the fullband. This procedure emphasizes the variation of crosscorrelation in the stereo signals. Consequently, convergence speed to the true echo path impulse response can be improved [12].

5. ADAPTIVE ALGORITHMS

5.1. Emphasis of Independent Noise and Variation in Crosscorrelation

The independent noise and the variation in the crosscorrelation between stereo signals can be used to estimate the true echo path impulse responses. The input signal covariance matrix is now full rank, but very ill-conditioned because of the inaudibility restriction. Convergence to the true solution depends on a relatively small term. The independent noise and the variations in the crosscorrelation adapt the filters toward convergence, while the speech signal disrupts the adaptation. The next step is how to emphasize the noise and variations and use them to accelerate the filter coefficient error convergence.

If the crosscorrelation is constant until time $k-1$, combined stereo signal vectors exist in the subspace determined by the crosscorrelation, i.e., $\boldsymbol{x}(k-1)$, $\boldsymbol{x}(k-2)$, $\ldots \in \boldsymbol{S}$, as shown in Fig. 7. If the crosscorrelation then varies at time k, the combined stereo signal vector $\boldsymbol{x}(k)$ exists in a subspace different from $\boldsymbol{S}$, i.e., $\boldsymbol{x}(k) \in \boldsymbol{S}'$. We can thus treat $\boldsymbol{x}(k)$ as the sum of two orthogonal components, that is,

$$\boldsymbol{x}(k) = \boldsymbol{v}(k) + \bar{\boldsymbol{v}}(k) \quad [\boldsymbol{v}(k) \in \boldsymbol{S}, \bar{\boldsymbol{v}}(k) \perp \boldsymbol{S}], \tag{15}$$

where $\bar{\boldsymbol{v}}(k)$ is the new component used for convergence and $\boldsymbol{v}(k)$ is the redundant component. The independent noise component is contained in $\bar{\boldsymbol{v}}(k)$. Thus, if the direction of adjustment vector $\Delta\hat{\boldsymbol{h}}(k)$ is made the same as that of $\bar{\boldsymbol{v}}(k)$ by removing $\boldsymbol{v}(k)$ from $\boldsymbol{x}(k)$, the adjustment of $\hat{\boldsymbol{h}}(k)$ is achieved with no redundancy; the noise and the variation emphasis is achieved by deriving $\bar{\boldsymbol{v}}(k)$, and filter coefficient vector $\hat{\boldsymbol{h}}(k)$ is adjusted in the direction of the ''emphasized'' vector $\bar{\boldsymbol{v}}(k)$. This discussion also applies to the case where the crosscorrelation varies slightly until time $k-1$. In this case, subspace $\boldsymbol{S}$ becomes slightly wider; however, the vector $\bar{\boldsymbol{v}}(k)$ still becomes an ''emphasized'' vector orthogonal to ''wide'' subspace $\boldsymbol{S}$.

Since $\boldsymbol{v}(k)$ is correlated with $\boldsymbol{x}(k-1)$, $\boldsymbol{x}(k-2)$, $\ldots$, it can be removed from $\boldsymbol{x}(k)$ by the decorrelation; for example, by using the recursive least squares (RLS) algorithm or the projection algorithm. The RLS algorithm removes the redundant component completely, but its computational cost is very high. On the other hand, a projection algorithm, or affine projection algorithm, of order p removes the major redundant p components at small computational cost [13–17].

5.2. Stereo RLS Algorithm

The two-channel RLS algorithm de-autocorrelates completely; however, its computational cost is very high.

5.3. Stereo Projection Algorithm

The stereo projection algorithm has been shown to be effective in stereo echo cancellation [3,18]. This algorithm emphasizes the noise and the variation in crosscorrelation in the stereo signals and de-autocorrelates the two input signals to improve the convergence speed to the true echo

path impulse response.

A p-th order projection algorithm updates the filter coefficient vector $\hat{\boldsymbol{h}}(k)$ as

$$\hat{\boldsymbol{h}}(k+1) = \hat{\boldsymbol{h}}(k) + \mu \Delta \hat{\boldsymbol{h}}(k). \tag{16}$$

$$\Delta \hat{\boldsymbol{h}}(k) = \beta_1(k)\boldsymbol{x}(k) + \beta_2(k)\boldsymbol{x}(k-1) + \cdots + \beta_p(k)\boldsymbol{x}(k-p+1), \tag{17}$$

where μ is scalar step size ($0 < \mu < 2$) and $\beta_1(k)$, $\beta_2(k)$, $\ldots$, $\beta_p(k)$ are determined so that $\hat{\boldsymbol{h}}^T(k+1)$ satisfies the following equations when $\mu = 1$ [19].

$$\begin{aligned} \hat{\boldsymbol{h}}^T(k+1)\boldsymbol{x}(k) &= y(k) \\ \hat{\boldsymbol{h}}^T(k+1)\boldsymbol{x}(k-1) &= y(k-1) \\ &\vdots \\ \hat{\boldsymbol{h}}^T(k+1)\boldsymbol{x}(k-p+1) &= y(k-p+1). \end{aligned} \tag{18}$$

Consequently, $\Delta \hat{\boldsymbol{h}}(k)$ becomes a decorrelated component of $\boldsymbol{x}(k)$ by subtracting the correlated components of $\boldsymbol{x}(k-1)$, $\boldsymbol{x}(k-2)$, $\ldots$, $\boldsymbol{x}(k-p+1)$ from $\boldsymbol{x}(k)$. In Fig. 7, $\boldsymbol{v}(k)$ is the only component of $\boldsymbol{x}(k)$ from which the correlated components of $\boldsymbol{x}(k-1)$, $\boldsymbol{x}(k-2)$, $\ldots$, $\boldsymbol{x}(k-p+1)$ are removed because $\bar{\boldsymbol{v}}(k)$ is orthogonal to $\boldsymbol{x}(k-1)$, $\boldsymbol{x}(k-2)$, $\ldots$, $\boldsymbol{x}(k-p+1)$. If order p is adequately determined, almost all the components are removed from $\boldsymbol{v}(k)$ and the direction of the adjustment vector $\Delta \hat{\boldsymbol{h}}(k)$ becomes the same as that of $\bar{\boldsymbol{v}}(k)$.

Unlike the ''extended'' algorithms described below, the stereo projection algorithm does not take the de-crosscorrelation into account explicitly. However, de-autocorrelating each channel automatically results in de-crosscorrelation. Thus, the stereo projection algorithm does not suffer from unstable convergence.

5.4. Extended LMS Algorithm and Extended Projection Algorithm

The extended LMS algorithm [2] and the extended projection algorithm [18] try to de-crosscorrelate the two signals in the algorithm formula. De-crosscorrelating the stereo signal that has strong crosscorrelation tends to cause instability, as explained in section 4.4. Nevertheless, the extended LMS algorithm and the extended projection algorithm are effective, since they introduce some constants for controlling the stability.

Using the input-output relationships for reversed stereo signals, Ref. [20] gave a basis to the extended LMS algorithm and the extended projection algorithm

6. PERFORMANCE EXAMPLES

6.1. Performance of the Stereo Projection Algorithm

Figure 8 shows the performance of the stereo projection algorithm for the projection order $p = 1$, 2, 4, 8. Setting $p = 1$ is equivalent to using a conventional NLMS algorithm.

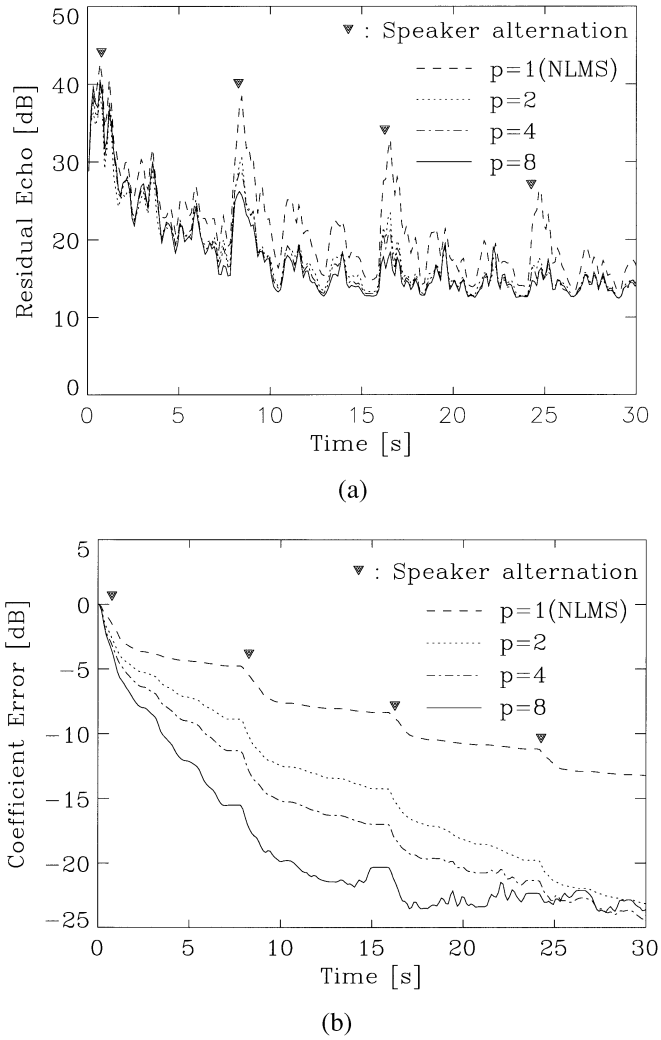


Fig. 8 Convergence for speech input with stereo projection algorithm.

The input stereo signals were speech signals made by two people speaking alternately and recorded with two microphones in a conference room. The speakers were in different parts of the room and did not move their bodies or heads. The number of taps was 512 in each filter of $\hat{\boldsymbol{h}}_1(k)$ and $\hat{\boldsymbol{h}}_2(k)$. The sampling frequency was 8 kHz. The true echo path impulse responses were measured in the conference room with a reverberation time of 150 ms. Ambient noise with a fixed SNR of 35 dB was added.

As the projection order p increased, the canceller became more effective in preventing the residual echo increase caused by speaker alternation, as showed in Fig. 8(a), and filter coefficient error convergence was achieved more quickly, as shown in Fig. 8(b).

6.2. Effect of Subband Processing

Figure 9 shows coefficient error convergence for $p = 1$ (NLMS). The downsampling rate is given by $R = N/4$. The convergence improved considerably with the number

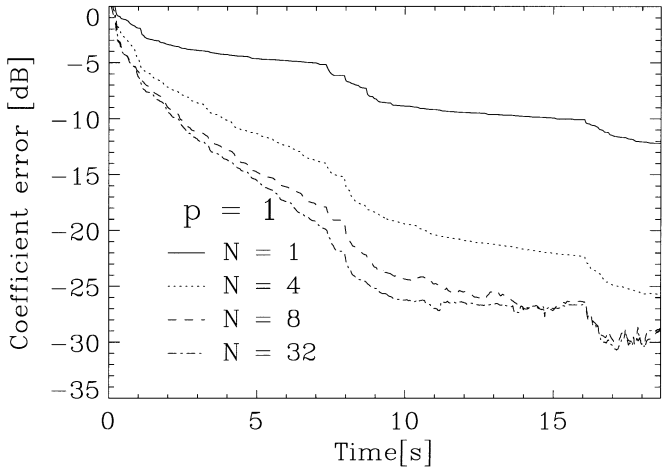


Fig. 9 Coefficient error convergence for speech input when $p = 1$ NLMS algorithm.

of subbands [12].

It can be seen that Figs. 8(b) and 9 roughly look the same. The subband structure has almost the same effect as the stereo projection algorithm, as explained below.

The p-th order projection algorithm updates filter coefficient $\hat{\boldsymbol{h}}(k+1)$, which satisfies Eq. (18), where $p < L$. Equation (18) shows that if $\boldsymbol{x}(k-i+1)$ is input, then filter $\hat{\boldsymbol{h}}(k+1)$ outputs the correct value $y(k-i+1)$. On the other hand, in subband processing with downsampling rate R, $R-1$ samples are redundant and Eq. (18) is automatically satisfied for $i = 1, 2, \cdots, R$. Thus, subband processing with downsampling rate R, filter length L/R, and update occasion $1/R$ is equivalent to the *R-th* order projection algorithm with original sampling, filter length L, and update occasion $1/R$.

6.3. Subband Stereo Projection Algorithm

Using the projection algorithm in the subband structure further improves convergence speed [12].

Figure 10(a) shows residual echo convergence for $N = 32$ subbands. This figure shows that the residual echo is not affected by speaker alternation in the transmission room.

Figure 10(b) shows coefficient error convergence for $N = 32$ subbands. Convergence speed can be improved by increasing the projection order. The convergence with a projection order of 32 is more than doubled compared to that with $p = 1$ (NLMS).

Since the number of taps needed in each subband is reduced by downsampling by a factor of R, the projection algorithm can decorrelate the received input of a small-tap adaptive filter with a relatively small projection order [21]. Therefore, instead of using the fullband with a high projection order, which has a high computational load, we can use subbands with a low projection order, which yields a low computational load.

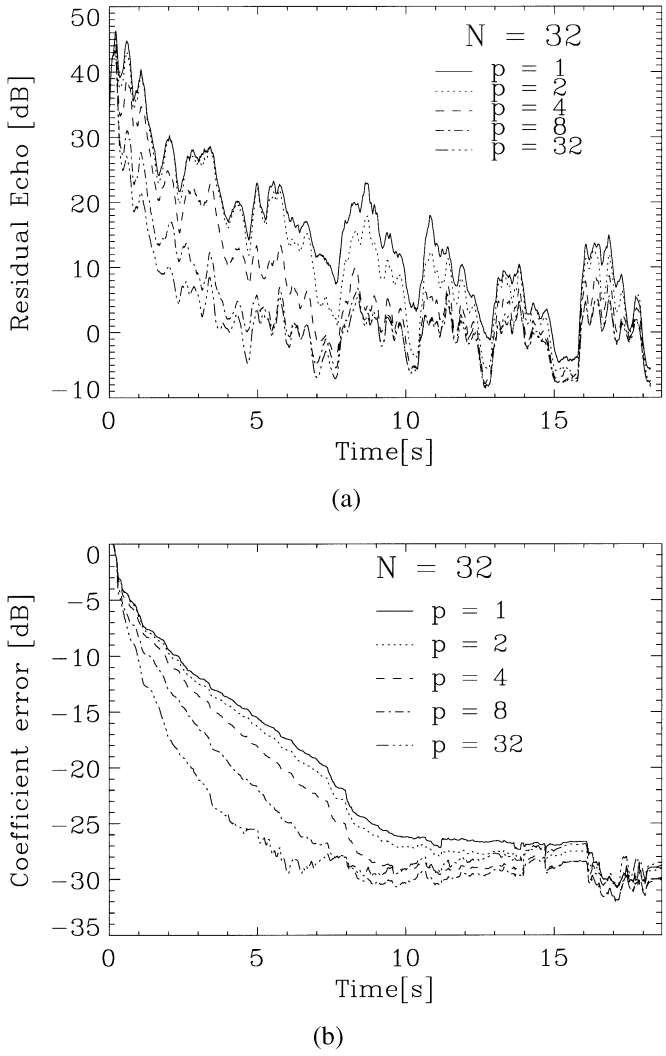


Fig. 10 Convergence for speech input when $N = 32$ subbands.

7. IMPLEMENTATION EXAMPLE

Figure 11 shows an implementation example of the stereo echo canceller. A system was implemented with a frequency range of 100 Hz to 20 kHz on DSPs. The number of taps in the filter are 1,200 (0.1–4 kHz), 800 (4–

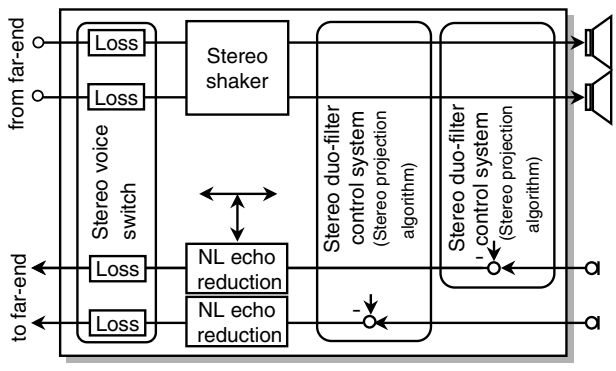


Fig. 11 Stereo echo canceller implementation example.

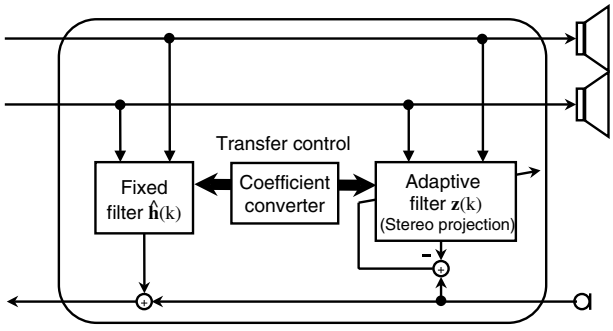


Fig. 12 Stereo duo-filter structure.

8 kHz). In the range of 8 to 20 kHz, the stereo voice switch alone was used [22].

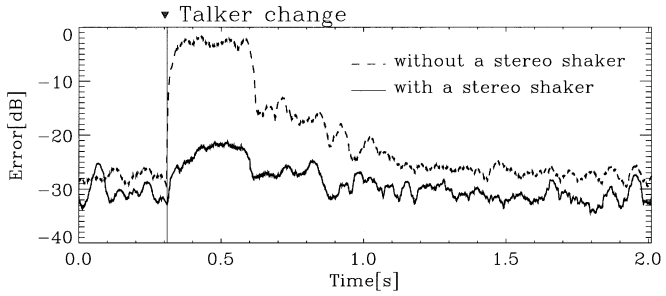
To achieve true echo path estimation, a stereo shaker (function block) was introduced in eight frequency bands and adjusted subjectively so as to be inaudible and not affect stereo localization in two-way conversations in teleconferencing rooms. A duo-filter control system [23] including a continually running adaptive filter and a fixed filter is used for double-talk control (Fig. 12). A second-order stereo projection algorithm is used in the adaptive filter, and a stereo voice switch was also implemented.

7.1. Convergence at Change of Far-end Talker Position

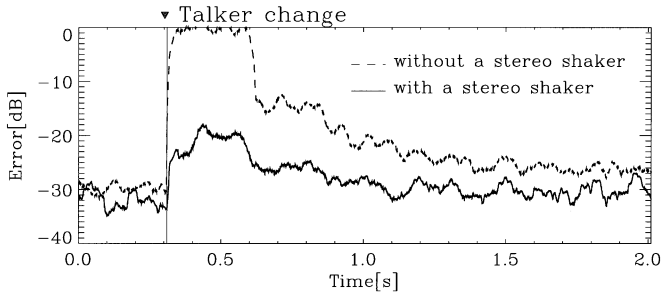
Real-time experiments were performed with the hardware in two teleconferencing rooms. Each room had a volume of 150 m^3 and a reverberation time of 300 ms. Ambient noise was about 25 dB in SNR. The stepsize μ of the projection algorithm was set to be 0.5.

Figure 13 shows the error level of (a) right channel S_{out} and (b) left channel S_{out}. Only the echo canceller was active. The far-end stereo signals were white Gaussian noise with fixed crosscorrelation. After the far-end right ''talker'' talked 90 s, the ''talker'' changed positions in the transmission room from right to left. Without the stereo shaker, the error was increased 20 dB when the talker changed positions, since the echo path was not correctly identified. With the stereo shaker, the error was not increased by the change in talker positions. This error difference of 20 dB indicates that the misalignment of the echo canceller was about 20 dB.

The stereo echo canceller has been used daily in a teleconferencing system. The echo canceller gained a sound level of 10 dB compared to a conventional system. Howling and echo were eliminated and speech quality was improved. Combined with a 2.4-m (W)×1.3-m (H) screen, the sound localization brought high presence to teleconferencing. More than 500 guests have used it [22].



(a)



(b)

Fig. 13 Error level of (a) right channel S_{out}, (b) left channel S_{out} for white Gaussian input with fixed crosscorrelation. After 90 s, the far-end ''talker'' changed positions in the transmission room from right to left.

8. CONCLUSIONS

The fundamental problem of stereo echo cancellation has been discussed and recent solutions have been reviewed. While using our hardware daily for over a year, we have noticed that there is plenty of noise and variations in the real world. (Sometimes, there is too much noise and we even need noise reduction. The noise and variations are sometimes larger than the processed ones.) If two or more independent and spatially separated sources are active in the transmission room, then the non-uniqueness problem essentially disappears. The next step is to create a sophisticated combined control of double-talk, voice switching, and nonlinear echo reduction.

ACKNOWLEDGEMENTS

We would like to thank Dr. Yutaka Kaneda, Mr. Masashi Tanaka, and Mr. Kuniyasu Suzuki for many fruitful discussions. We also thank Dr. Jacob Benesty for detailed discussions.

REFERENCES

[1] M. Sondhi, D. Morgan and J. Hall, ''Stereophonic acoustic echo cancellation—An overview of the fundamental problem'', *IEEE Signal Process. Lett.*, **2**, 148 (1995).

[2] J. Benesty, F. Amand, A. Gilloire and Y. Grenier, ''Adaptive filtering algorithm for stereophonic acoustic echo cancellation'', *Proc. ICASSP*, 95, 3099 (1995).

[3] S. Shimauchi and S. Makino, ''Stereo projection echo canceller with true echo path estimation'', *Proc. ICASSP*, 95, 3059 (1995).
[4] J. Benesty, D. Morgan and M. Sondhi, ''A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation'', *Proc. ICASSP*, 97, 303 (1997).
[5] S. Shimauchi, Y. Haneda, S. Makino and Y. Kaneda, ''New configuration for a stereo echo canceller with nonlinear preprocessing'', *Proc. ICASSP*, 98, 3685 (1998).
[6] A. Gilloire and V. Turbin, ''Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers'', *Proc. ICASSP*, 98, 3681 (1998).
[7] T. Gansler and P. Eneroth, ''Influence of audio coding on stereophonic acoustic echo cancellation'', *Proc. ICASSP*, 98, 3649 (1998).
[8] M. Ali, ''Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation'', *Proc. ICASSP*, 98, 3689 (1998).
[9] Y. Joncour and A. Sugiyama, ''A Stereo echo canceller with pre-processing for correct echo-path identification'', *Proc. ICASSP*, 98, 3677 (1998).
[10] Y. Mahieux, A. Gilloire and F. Khalil, ''Annulation D'echo en teleconference stereophonique'', *Proc. Quatorzieme Colloque GRETSI*, 515 (1993).
[11] A. Hirano, K. Nakayama and K. Watanabe, ''Convergence analysis of stereophonic echo canceller with pre-processing—Relation between pre-processing and convergence—'', *Proc. ICASSP*, 99, 861 (1999).
[12] S. Makino, K. Strauss, S. Shimauchi, Y. Haneda and A. Nakagawa, ''Subband stereo echo canceller using the projection algorithm with fast convergence to the true echo path'', *Proc. ICASSP*, 97, 299 (1997).
[13] K. Ozeki and T. Umeda, ''An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties'', *Trans. IEICE Jpn.*, **J67-A**, 126 (1984).
[14] M. Tanaka, Y. Kaneda, S. Makino and J. Kojima, ''Fast projection algorithm and its step size control'', *Proc. ICASSP*, 95, 945 (1995).
[15] S. Gay and S. Tavathia, ''The fast affine projection algorithm'', *Proc. ICASSP*, 95, 3023 (1995).
[16] M. Tanaka, Y. Kaneda, S. Makino and J. Kojima, ''A fast projection algorithm for adaptive filtering'', *Trans. IEICE Jpn.*, **E78-A**, 1355 (1995).
[17] F. Amand, J. Benesty, A. Gilloire and Y. Grenier, ''A fast two-channel projection algorithm for stereophonic acoustic echo cancellation'', *Proc. ICASSP*, 96, 949 (1996).
[18] J. Benesty, P. Duhamel and Y. Grenier, ''A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation'', *IEEE Signal Process. Lett.*, **3**, 35 (1996).
[19] S. Makino and Y. Kaneda, ''Exponentially weighted stepsize projection algorithm for acoustic echo cancellers'', *Trans. IEICE Jpn.*, **E75-A**, 1500 (1992).
[20] S. Shimauchi and S. Makino, ''Stereo echo cancellation algorithm using imaginary input-output relationships'', *Proc. ICASSP*, 96, 941 (1996).
[21] S. Makino, J. Noebauer, Y. Haneda and A. Nakagawa, ''SSB subband echo canceller using low-order projection algorithm'', *Proc. ICASSP*, 96, 945 (1996).
[22] S. Shimauchi, S. Makino, Y. Haneda, A. Nakagawa and S. Sakauchi, ''A stereo echo canceller implemented using a stereo shaker and a duo-filter control system'', *Proc. ICASSP*, 99, 857 (1999).
[23] Y. Haneda, S. Makino, J. Kojima and S. Shimauchi, ''Implementation and evaluation of an acoustic echo canceller using duo-filter control system'', *Proc. EUSIPCO*, 96, 1115 (1996).

Shoji Makino was born in Nikko, Japan, on June 4, 1956. He received the B. E., M. E., and Ph. D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively. He joined the Electrical Communication Laboratory of Nippon Telegraph and Telephone Corporation (NTT) in 1981. Since then, he has been engaged in research on electroacoustic transducers and acoustic echo cancellers. He is now a Senior Research Scientist, Supervisor, Group Leader at the Speech Open Laboratory of the NTT Communication Science Laboratories. His research interests include acoustic signal processing, and adaptive filtering and its applications. Dr. Makino received the Outstanding Technological Development Award of the Acoustical Society of Japan in 1995, and the Achievement Award of the Institute of Electronics, Information, and Communication Engineers of Japan in 1997. He is the author or co-author of more than 100 articles in journals and conference proceedings, and more than 130 patents. He is a member of the Audio and Electroacoustics Technical Committee of the IEEE Signal Processing Society. He served on the Organizing Committee of the 1999 IEEE Workshop on Acoustic Echo and Noise Control. He is a Senior Member of the IEEE, a member of the Acoustical Society of Japan, and the Institute of Electronics, Information, and Communication Engineers of Japan.