A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation

Jacob Benesty, Dennis R. Morgan, *Senior Member, IEEE,* and Man Mohan Sondhi

***Abstract*—Teleconferencing systems employ acoustic echo cancelers to reduce echoes that result from coupling between the loudspeaker and microphone. To enhance the sound realism, two-channel audio is necessary. However, in this case (stereophonic sound) the acoustic echo cancellation problem is more difficult to solve because of the necessity to uniquely identify two acoustic paths. In this paper, we explain these problems in detail and give an interesting solution which is much better than previously known solutions. The basic idea is to introduce a small nonlinearity into each channel that has the effect of reducing the interchannel coherence while not being noticeable for speech due to self masking.**

I. Introduction

ACOUSTIC echo cancelers (AEC's) are necessary for communication systems such as teleconferencing to reduce echos that result from coupling between the loudspeaker and microphone. With conventional single-channel (monophonic) systems, such AEC's simultaneously reduce the echo and identify the acoustic path so that the echo remains cancelled no matter what happens at the remote transmission room.

A stereo teleconferencing system provides a more realistic presence than a monophonic system, because listeners can use spatial information to help distinguish who is speaking. This is especially important for video teleconferencing involving many different talkers. However, there are now two acoustic paths to identify, which as we will explain, raises some fundamental problems.

Stereophonic acoustic echo cancellation can be viewed as a straightforward generalization of the single-channel acoustic echo cancellation principle [1], as illustrated in Fig. 1. The similarity between the single-channel and stereophonic AEC's, however, is deceptive. Stereophonic AEC's present problems that are fundamentally different from those of single-channel AEC's [2]. Note that, for simplicity, the acoustic paths to only one microphone are shown in the receiving room on the left; it is understood that similar analysis will apply to the other microphone signal. Clearly, according to this scheme, stereophonic acoustic echo cancellation consists of direct identification of a multi-input, unknown linear system, consisting of the parallel combination of two acoustic paths (h_1, h_2) extending through the receiving room from the loudspeakers to the microphone. The stereophonic AEC tries to model this unknown system by a pair of adaptive filters $(\hat{h}_1, \hat{h}_2)$. The same model applies to the other microphone with the acoustic paths replaced by the ones appropriate to that microphone. Also, a similar canceler system would typically be employed for the transmission room on the right.

In the following, we explain the main problems encountered due to the strong cross-correlation between the two input signals (x_1, x_2), and we propose a new solution based on nonlinear transformations to overcome these problems.

Manuscript received May 30, 1996; revised May 29, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James H. Snyder.

The authors are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974-0636 USA (e-mail: jbenesty@bell-labs.com).

Publisher Item Identifier S 1063-6676(98)01737-4.

II. The Nonuniqueness Problem

In this section we show that the solution of the normal equation is not as obvious as in the single-channel case. Indeed, since the two input signals are obtained by filtering from a common source, a problem of nonuniqueness is expected [2]. In the following discussion, we distinguish between the length (M) of the impulse responses in the transmission room, the length (L) of the modeling filters, and the length (N) of the impulse responses in the receiving room.

We assume that the system (transmission room) is linear and time invariant; therefore, we have the following relation [3]:

$$\mathbf{x}_{1,M}^T(n)\mathbf{g}_{2,M} = \mathbf{x}_{2,M}^T(n)\mathbf{g}_{1,M} \tag{1}$$

where

$$\mathbf{x}_{i,M}(n) = [x_i(n) \quad x_i(n-1) \quad \cdots \quad x_i(n-M+1)]^T, \quad i = 1, 2$$

are vectors of signal samples at the microphone outputs in the transmission room, ^{T} denotes the transpose of a vector or a matrix, and the impulse response vectors are defined as

$$\mathbf{g}_{i,M} = [g_{i,0} \quad g_{i,1} \quad \cdots \quad g_{i,M-1}]^T, \qquad i = 1, 2.$$

This is easy to see in the frequency domain, since for a source spectrum $S(f)$, we have

$$X_1(f) = G_1(f)S(f) \tag{2}$$

and

$$X_2(f) = G_2(f)S(f). \tag{3}$$

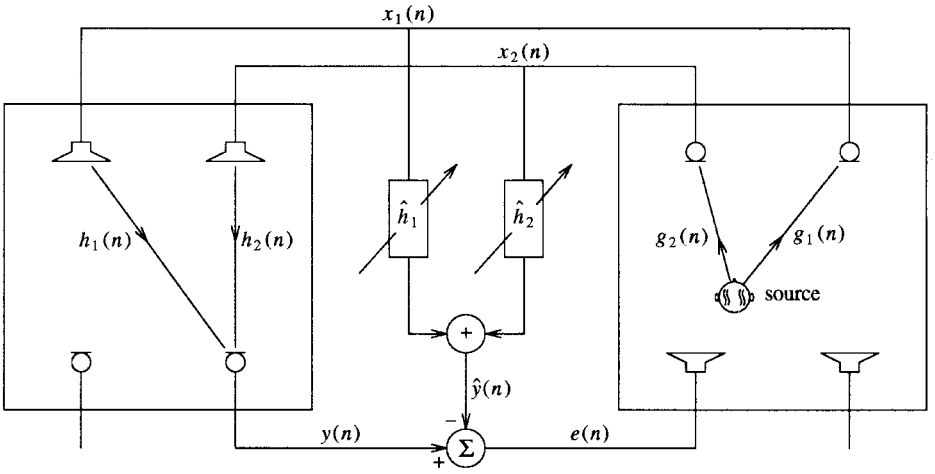


Fig. 1. Schematic diagram of stereophonic echo cancellation.

Hence

$$X_1(f)G_2(f) = X_2(f)G_1(f) \tag{4}$$

which is the Fourier transform of (1).

We could develop the theory in terms of Wiener filters using mathematical expectations. However, for concreteness, we choose here to work in terms of weighted least squares, which will lead to equivalent results and moreover is closer to the actual implementation. Thus, let us define the recursive least squares error criterion with respect to the modeling filters

$$J(n) = \sum_{p=1}^{n} \lambda^{n-p} e^2(p) \tag{5}$$

where λ $(0 < \lambda \leq 1)$ is an exponential forgetting factor,

$$e(n) = y(n) - \hat{\mathbf{h}}_{1,L}^T(n)\mathbf{x}_{1,L}(n) - \hat{\mathbf{h}}_{2,L}^T(n)\mathbf{x}_{2,L}(n) \tag{6}$$

is the error signal at time n between the microphone output

$$y(n) = \mathbf{h}_{1,N}^T\mathbf{x}_{1,N}(n) + \mathbf{h}_{2,N}^T\mathbf{x}_{2,N}(n) \tag{7}$$

and its estimate, and

$$\begin{aligned}
\mathbf{h}_{i,N} &= [h_{i,0} \quad h_{i,1} \quad \cdots \quad h_{i,N-1}]^T \\
\mathbf{x}_{i,N}(n) &= [x_i(n) \quad x_i(n-1) \quad \cdots \quad x_i(n-N+1)]^T \\
\hat{\mathbf{h}}_{i,L}(n) &= [\hat{h}_{i,0}(n) \quad \hat{h}_{i,1}(n) \quad \cdots \hat{h}_{i,L-1}(n)]^T \\
\mathbf{x}_{i,L}(n) &= [x_i(n) \quad x_i(n-1) \quad \cdots \quad x_i(n-L+1)]^T, \\
& \qquad i = 1, 2.
\end{aligned}$$

The minimization of (5) leads to the normal equation

$$\mathbf{R}(n)\begin{bmatrix} \hat{\mathbf{h}}_{1,L}(n) \\ \hat{\mathbf{h}}_{2,L}(n) \end{bmatrix} = \mathbf{r}(n) \tag{8}$$

where

$$\mathbf{R}(n) = \sum_{p=1}^{n} \lambda^{n-p} \begin{bmatrix} \mathbf{x}_{1,L}(p) \\ \mathbf{x}_{2,L}(p) \end{bmatrix} [\mathbf{x}_{1,L}^T(p) \quad \mathbf{x}_{2,L}^T(p)] \tag{9}$$

is an estimate of the input signal covariance matrix and

$$\mathbf{r}(n) = \sum_{p=1}^{n} \lambda^{n-p} y(p) \begin{bmatrix} \mathbf{x}_{1,L}(p) \\ \mathbf{x}_{2,L}(p) \end{bmatrix} \tag{10}$$

is an estimate of the cross-correlation vector between the input and output signals. In the following, we assume that the estimated autocorrelation matrices of the two input signals are invertible. Now, the important question is: is $\mathbf{R}(n)$ full rank or not? If it is not, then there is no unique solution to the problem and an adaptive algorithm will drive to any one of many possible solutions, which can be very different from the "true" desired solution $\hat{\mathbf{h}}_{1,L} = \mathbf{h}_{1,L}$ and $\hat{\mathbf{h}}_{2,L} = \mathbf{h}_{2,L}$, where

$$\mathbf{h}_{i,L} = [h_{i,0} \quad h_{i,1} \quad \cdots \quad h_{i,L-1}]^T, \qquad i = 1, 2.$$

Let us examine two possible cases according to the length of the modeling filters [3], as follows.

$L \geq M$: Consider the vector $\mathbf{u} = [\mathbf{g}_{2,M}^T \; 0 \; \cdots \; 0 - \mathbf{g}_{1,M}^T \; 0 \; \cdots \; 0]^T$, containing $2 \times (L - M)$ zero coefficients. We can verify using (1) that $\mathbf{R}(n)\mathbf{u} = \mathbf{0}_{2L\times 1}$, so $\mathbf{R}(n)$ is not invertible. From this result, we deduce for $L \geq N$ that

$$\hat{\mathbf{h}}_{1,L} = \mathbf{h}_{1,L} + \beta[\mathbf{g}_{2,M}^T \quad 0 \quad \cdots \quad 0]^T \tag{11}$$

and

$$\hat{\mathbf{h}}_{2,L} = \mathbf{h}_{2,L} + \beta[-\mathbf{g}_{1,M}^T \quad 0 \quad \cdots \quad 0]^T \tag{12}$$

where β is an arbitrary constant. These nonunique "solutions" are dependent on the impulse responses in the transmission room. This, of course, is intolerable because $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$ can change instantaneously, for example, as one person stops talking and another starts [2]. Equations (11) and (12) suppose that the covariance matrix $\mathbf{R}(n)$ has only one eigenvalue equal to zero and the corresponding eigenvector is $\mathbf{u}$, which is true if the two impulse responses $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$ have no common zeros and the autocorrelation matrix of the source signal is full rank. If $\mathbf{R}(n)$ has more than one zero eigenvalue, $\hat{\mathbf{h}}_{1,L}$ and $\hat{\mathbf{h}}_{2,L}$ will still depend explicitly on $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$.

$L < M$: This is the real case, since $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$ are actually of infinite length. Now, (1) can be expressed as

$$\mathbf{x}_{1,L}^T(n)\mathbf{g}_{2,L} + q_1(n-L) = \mathbf{x}_{2,L}^T(n)\mathbf{g}_{1,L} + q_2(n-L) \tag{13}$$

with

$$q_1(n-L) = \sum_{i=L}^{M-1} x_1(n-i)g_{2,i}$$

and

$$q_2(n-L)=\sum_{i=L}^{M-1}x_2(n-i)g_{1,i}.$$

From (1) we know that $\mathbf{x}_{1,M}(n)$ and $\mathbf{x}_{2,M}(n)$ are linearly related, but from (13) we can see that the same is not true for $\mathbf{x}_{1,L}(n)$ and $\mathbf{x}_{2,L}(n)$ [except if $q_1(n-L)=q_2(n-L)$, which happens only when $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$ have at least $M-L$ common zeros—an event that rarely occurs in practice]. Hence, in principle, the covariance matrix $\mathbf{R}(n)$ is full-rank, but is very ill-conditioned because $q_1(n-L)$ and $q_2(n-L)$ are in general very small.

Thus, for the practical case when $L<M$, there is a unique solution to the normal equation, although the covariance matrix is very ill-conditioned. We can define the number

$$\zeta^2=\frac{\sum_{i=1}^{2}\mathbf{g}_{i,L}^T\mathbf{g}_{i,L}}{\sum_{i=1}^{2}\mathbf{g}_{i,M}^T\mathbf{g}_{i,M}} \tag{14}$$

to measure how close $\mathbf{x}_{1,L}$ and $\mathbf{x}_{2,L}$ are to being linearly related when we do not neglect the impulse response tails in the transmission room $(0<\zeta\leq 1)$. If $L=M$, then $\zeta=1$, hence $\mathbf{x}_{1,L}$ and $\mathbf{x}_{2,L}$ are linearly related, as shown in (1). The smaller ζ, the better the condition number of the covariance matrix of the input signals. In the next section we explain how ill-conditioning leads to a poor solution in the face of strong cross-correlation between the input signals.

III. The Misalignment Problem

The mismatch between the modeling filters $\hat{\mathbf{h}}=[\hat{\mathbf{h}}_{1,L}^T\ \hat{\mathbf{h}}_{2,L}^T]^T$ and the truncated impulse responses of the receiving room $\mathbf{h}=[\mathbf{h}_{1,L}^T\ \mathbf{h}_{2,L}^T]^T$ is quantified by the so-called "misalignment," which is defined as

$$\varepsilon=\|\mathbf{h}-\hat{\mathbf{h}}\|/\|\mathbf{h}\|. \tag{15}$$

It is possible to have good echo cancellation even when misalignment is large. However, in such a case, the cancellation will worsen if $\mathbf{g}_{1,M}$ and $\mathbf{g}_{2,M}$ change. One of the main objectives of the present work is to avoid this problem.

Let us see what happens in practice to the filter coefficients and examine the link between the cross-correlation of the input signals and the misalignment. We examine first the single-channel AEC and explain why in general the value of the misalignment is low in this case.

A. The Single-Channel Case

We have the classical normal equation

$$\mathbf{R}_{11}(n)\hat{\mathbf{h}}_{1,L}(n)=\mathbf{r}_1(n) \tag{16}$$

where

$$\mathbf{R}_{11}(n)=\sum_{p=1}^{n}\lambda^{n-p}\mathbf{x}_{1,L}(p)\mathbf{x}_{1,L}^T(p) \tag{17}$$

is an estimate of the autocorrelation matrix of the input signal x_1 and

$$\mathbf{r}_1(n)=\sum_{p=1}^{n}\lambda^{n-p}y_1(p)\mathbf{x}_{1,L}(p) \tag{18}$$

is an estimate of the cross-correlation vector between x_1 and the microphone signal

$$y_1(n)=\mathbf{h}_{1,N}^T\mathbf{x}_{1,N}(n). \tag{19}$$

This solution minimizes the single-channel version of the recursive least squares error criterion (5).

In practice, the impulse response $\mathbf{h}_{1,N}$ of the receiving room will always be very long so that any finite impulse response (FIR) modeling filter $\hat{\mathbf{h}}_{1,L}$ will only form an approximation to the actual impulse response. Let us split the impulse response into the following two parts:

$$\mathbf{h}_{1,N}=\begin{bmatrix}\mathbf{h}_{1,L}\\ \mathbf{h}_{1,\mathrm{t}}\end{bmatrix}$$

where $\mathbf{h}_{1,L}$ is a vector of size L and $\mathbf{h}_{1,\mathrm{t}}$ is the "tail" of the impulse response that is not modeled by $\hat{\mathbf{h}}_{1,L}$. Now the microphone signal is

$$\begin{aligned}y_1(n)&=\mathbf{h}_{1,N}^T\mathbf{x}_{1,N}(n)\\ &=\mathbf{h}_{1,L}^T\mathbf{x}_{1,L}(n)+\mathbf{h}_{1,\mathrm{t}}^T\mathbf{x}_{1,\mathrm{t}}(n-L)\end{aligned} \tag{20}$$

with

$$\begin{aligned}\mathbf{x}_{1,\mathrm{t}}(n-L)=[&x_1(n-L)\quad x_1(n-L-1)\\ &\cdots\quad x_1(n-N+1)]^T\end{aligned}$$

and the solution of the normal equation, assuming that $\mathbf{R}_{11}(n)$ is invertible, becomes

$$\hat{\mathbf{h}}_{1,L}(n)=\mathbf{h}_{1,L}+\mathbf{R}_{11}^{-1}(n)\mathbf{R}_{11,\mathrm{t}}(n)\mathbf{h}_{1,\mathrm{t}} \tag{21}$$

with

$$\mathbf{R}_{11,\mathrm{t}}(n)=\sum_{p=1}^{n}\lambda^{n-p}\mathbf{x}_{1,L}(p)\mathbf{x}_{1,\mathrm{t}}^T(p-L). \tag{22}$$

We note that, in addition to $\mathbf{h}_{1,L}$, the coefficients of the modeling filter also depend on the tail of the impulse response.

Actually, the misalignment that we obtain by minimizing the recursive least-squares error is

$$\begin{aligned}\varepsilon_{\min}^2(n)&=[\mathbf{h}_{1,L}-\hat{\mathbf{h}}_{1,L}(n)]^T[\mathbf{h}_{1,L}-\hat{\mathbf{h}}_{1,L}(n)]/\mathbf{h}_{1,L}^T\mathbf{h}_{1,L}\\ &=\mathbf{h}_{1,\mathrm{t}}^T\mathbf{Q}_{11,\mathrm{t}}(n)\mathbf{h}_{1,\mathrm{t}}/\mathbf{h}_{1,L}^T\mathbf{h}_{1,L}\end{aligned} \tag{23}$$

where

$$\mathbf{Q}_{11,\mathrm{t}}(n)=\mathbf{R}_{11,\mathrm{t}}^T(n)\mathbf{R}_{11}^{-2}(n)\mathbf{R}_{11,\mathrm{t}}(n).$$

There are two interesting cases, as follows.

1) $L\geq N$: In this case, $\varepsilon_{\min}^2(n)=0$.
2) $L<N$: For L large enough, $\mathbf{R}_{11,\mathrm{t}}(n)\approx\mathbf{0}_{L\times(N-L)}$, then $\mathbf{Q}_{11,\mathrm{t}}(n)$, and hence $\varepsilon_{\min}^2(n)$, are very small.

(Note that it is not even necessary to have $L<M$ to have a solution for the single-channel case, in contrast to the two-channel case discussed in the last section.)

The problem of bad misalignment rarely appears in the single-channel case with full-rank $\mathbf{R}_{11}(n)$ (for a reasonable length of the modeling filter with regard to the length of the impulse response). However, we next show that with stereophonic sound, the misalignment is much worse because of the strong cross-correlation between the input signals and the bad condition number of the covariance matrix.

B. The Two-Channel Case

We use the same approach as in the single-channel case. First, we split the two impulse responses into two parts each, as follows:

$$\mathbf{h}_{i,N}=\begin{bmatrix}\mathbf{h}_{i,L}\\ \mathbf{h}_{i,\mathrm{t}}\end{bmatrix},\qquad i=1,2$$

where $\mathbf{h}_{1,L}$ and $\mathbf{h}_{2,L}$ are vectors of length L, and $\mathbf{h}_{1,\mathrm{t}}$ and $\mathbf{h}_{2,\mathrm{t}}$ are the tails of the two impulse responses that are not modeled by $\hat{\mathbf{h}}_{1,L}$ and $\hat{\mathbf{h}}_{2,L}$. For the ideal noiseless case, the microphone output signal is then expressed as

$$\begin{aligned}y(n)&=\sum_{i=1}^{2}\mathbf{h}_{i,N}^{T}\mathbf{x}_{i,N}(n)\\&=\sum_{i=1}^{2}\mathbf{h}_{i,L}^{T}\mathbf{x}_{i,L}(n)+\sum_{i=1}^{2}\mathbf{h}_{i,\mathrm{t}}^{T}\mathbf{x}_{i,\mathrm{t}}(n-L)\end{aligned}\tag{24}$$

with

$$\begin{aligned}\mathbf{x}_{i,\mathrm{t}}(n-L)=[&x_i(n-L)\quad x_i(n-L-1)\\&\cdots\quad x_i(n-N+1)]^{T},\qquad i=1,2.\end{aligned}$$

Now, assuming that $\mathbf{R}(n)$ is invertible, we can rewrite the normal equation (8) as

$$\begin{bmatrix}\hat{\mathbf{h}}_{1,L}(n)\\ \hat{\mathbf{h}}_{2,L}(n)\end{bmatrix}=\begin{bmatrix}\mathbf{h}_{1,L}\\ \mathbf{h}_{2,L}\end{bmatrix}+\mathbf{R}^{-1}(n)\mathbf{R}_{\mathrm{t}}(n)\begin{bmatrix}\mathbf{h}_{1,\mathrm{t}}\\ \mathbf{h}_{2,\mathrm{t}}\end{bmatrix}\tag{25}$$

with

$$\begin{aligned}\mathbf{R}_{\mathrm{t}}(n)&=\sum_{p=1}^{n}\lambda^{n-p}\begin{bmatrix}\mathbf{x}_{1,L}(p)\\ \mathbf{x}_{2,L}(p)\end{bmatrix}[\mathbf{x}_{1,\mathrm{t}}^{T}(p-L)\ \mathbf{x}_{2,\mathrm{t}}^{T}(p-L)]\\&=\begin{bmatrix}\mathbf{R}_{11,\mathrm{t}}(n)&\mathbf{R}_{12,\mathrm{t}}(n)\\ \mathbf{R}_{21,\mathrm{t}}(n)&\mathbf{R}_{22,\mathrm{t}}(n)\end{bmatrix}.\end{aligned}\tag{26}$$

The misalignment that we obtain by minimizing the recursive least-squares error is

$$\begin{aligned}\varepsilon_{\min}^{2}(n)&=\sum_{i=1}^{2}[\mathbf{h}_{i,L}-\hat{\mathbf{h}}_{i,L}(n)]^{T}[\mathbf{h}_{i,L}-\hat{\mathbf{h}}_{i,L}(n)]/\mathbf{h}^{T}\mathbf{h}\\&=\mathbf{h}_{\mathrm{t}}^{T}\mathbf{Q}_{\mathrm{t}}(n)\mathbf{h}_{\mathrm{t}}/\mathbf{h}^{T}\mathbf{h}\end{aligned}\tag{27}$$

where

$$\begin{aligned}\mathbf{h}&=[\mathbf{h}_{1,L}^{T}\quad\mathbf{h}_{2,L}^{T}]^{T}\\ \mathbf{h}_{\mathrm{t}}&=[\mathbf{h}_{1,\mathrm{t}}^{T}\quad\mathbf{h}_{2,\mathrm{t}}^{T}]^{T}\end{aligned}$$

and

$$\mathbf{Q}_{\mathrm{t}}(n)=\mathbf{R}_{\mathrm{t}}^{T}(n)\mathbf{R}^{-2}(n)\mathbf{R}_{\mathrm{t}}(n).$$

In this formulation, we implicitly make the realistic assumption that $L<M$ so that $\mathbf{R}^{-1}$ exists.

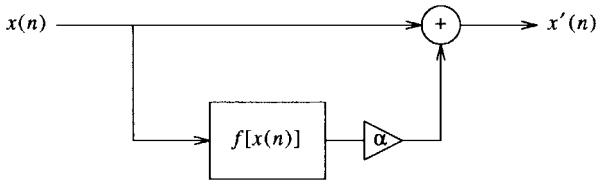


Fig. 2. Nonlinear transformation of a signal.

As in the single-channel case, if $L\geq N$ then the minimum value of the misalignment is zero. If $L<N$ and for L large enough we can make the following approximations: $\mathbf{R}_{11,\mathrm{t}}(n)\approx\mathbf{0}_{L\times(N-L)}$ and $\mathbf{R}_{22,\mathrm{t}}(n)\approx\mathbf{0}_{L\times(N-L)}$; however, we cannot say anything about the terms $\mathbf{R}_{12,\mathrm{t}}(n)$ and $\mathbf{R}_{21,\mathrm{t}}(n)$, because their value depends on the cross-correlation and delay between the two input signals x_1 and x_2. Moreover, since the covariance matrix is very ill-conditioned, then $\mathbf{Q}_{\mathrm{t}}(n)$ is not negligible. Thus, in the stereophonic case, the value of the misalignment can be high for $L<N$. Simulations confirm this analysis.

For $L<N$, we introduce an error in the filter coefficients both in the monaural and stereophonic applications. But for the stereo case, the problem is amplified because of the strong correlation between the two input signals which appears in the terms $\mathbf{R}_{12,\mathrm{t}}(n)$ and $\mathbf{R}_{21,\mathrm{t}}(n)$. So in practice we may have a bad misalignment even if there is a unique solution to the normal equation.

IV. The Impulse Response Tail Effect

We have seen that the tails of the impulse responses both in the transmission and receiving rooms play a key role. Thanks to the impulse response tails in the transmission room, we can obtain a unique solution to the normal equation. However, because of the impulse response tails in the receiving room, we have a bad misalignment. We suppose of course that $L<M$ and $L<N$, since this is the real case to be dealt with.

There are two ways to improve the misalignment. The first way is to use long adaptive filters; but when we do that, the adaptive algorithm becomes very slow in terms of convergence speed and is very expensive to implement in terms of memory, arithmetic complexity, etc. Moreover, the solution is not robust. A second way is to decorrelate partially (or in totality) the two input signals. However, up until now, there has been no completely satisfactory method to do this [2]. We next develop a new approach for reducing the cross-correlation.

V. The Coherence Function

In general, the coherence between two random signals x_1 and x_2 is defined in the frequency domain as

$$\gamma(f)=\frac{S_{x_1x_2}(f)}{\sqrt{S_{x_1x_1}(f)S_{x_2x_2}(f)}}\tag{28}$$

where

$$\begin{aligned}S_{x_kx_l}(f)&=\sum_{\tau=-\infty}^{+\infty}E\{x_k(n)x_l(n-\tau)\}e^{-i2\pi f\tau}\\&=\sum_{\tau=-\infty}^{+\infty}R_{x_kx_l}(\tau)e^{-i2\pi f\tau},\qquad k,l=1,2.\end{aligned}\tag{29}$$

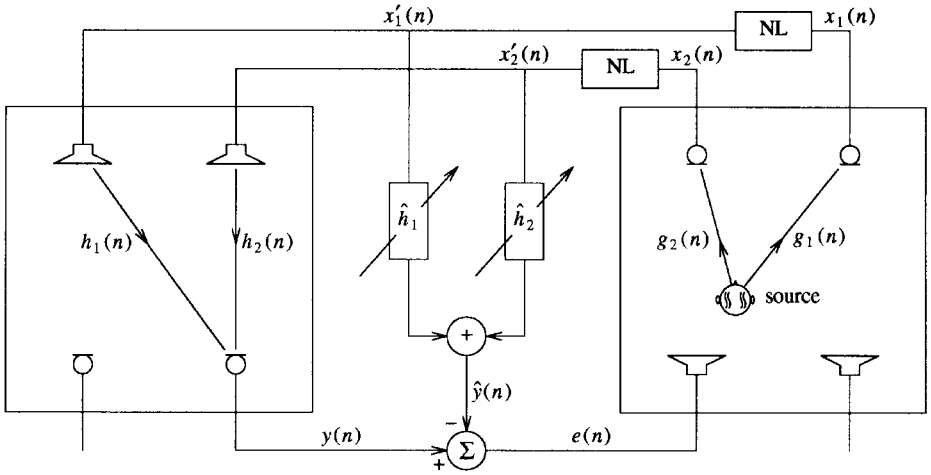


Fig. 3. Schematic diagram of stereophonic echo cancellation with nonlinear transformation of the two input signals.

The significance of the coherence function is that it can be shown to relate to the conditioning of the covariance matrix (9), and therefore determines the sensitivity of the normal equation solution to noise. It is shown in Appendix A that the eigenvalues of the covariance matrix are lower bounded by a factor $[1-|\gamma(f)|^2]$. Therefore, a magnitude-squared coherence of 0.999 at some frequency f would mean that the solution would be sensitive to noise at the -30 dB level. In the case where $|\gamma(f)| = 1$, there is of course no unique solution because the normal equation is singular.

The above definition of the coherence function takes into account an infinite number of cross-correlation coefficients $R_{x_1x_2}(\tau)$. However, in our context of minimizing the error criterion (5) (where the mathematical expectation is approximated by a weighted sum), we obtain an $L \times L$ Toeplitz cross-correlation matrix $E\{\mathbf{x}_{1,L}(n)\mathbf{x}_{2,L}^T(n)\}$ with only $2L-1$ cross-correlation coefficients (where L is the length of the modeling filters). In general, the finite length of the adaptive filters will serve to constrain the coherence and this will insure a unique solution. However, in teleconferencing systems, we use long adaptive filters (typically, $L \geq 1000$). Therefore, as a practical matter, we will use the coherence function to measure the cross-correlation between the two input signals and also as an indicator of the efficiency of the proposed decorrelation methods.

VI. The Proposed Solution: Use of Nonlinear Transformations

The very first idea to partially decorrelate the input signals (or reduce the coherence magnitude) was proposed in [2]. The idea is to simply add a low level of independent random noise to each channel in order to reduce the coherence

$$x_i'(n) = x_i(n) + \nu_i(n), \qquad i = 1,2 \tag{30}$$

where ν_1 and ν_2 are two independent white noises. Then, we can show that the noiseless coherence γ is modified to

$$\gamma'(f) = \frac{\rho}{1+\rho}\gamma(f) \tag{31}$$

where ρ is the signal-to-noise ratio (SNR) (assumed to be equal in each channel). When $\mathbf{x}_{1,M}$ and $\mathbf{x}_{2,M}$ are derived from a common source as in Fig. 1, the coherence magnitude $|\gamma(f)| = 1$ for a stationary source. An SNR $\rho = 20$ (13 dB) would therefore result in a modified coherence magnitude $|\gamma'(f)| \approx 0.95$. This reduction is enough to significantly reduce the misalignment. However, the level of white noise is quite high relative to the signal and is subjectively objectionable. It is possible that some advantage could be gained if instead of adding white noise, the noise is shaped so as to "hide" beneath the signal. This kind of noise shaping takes advantage of noise masking effects in the human auditory system [4] and has been used to advantage in perceptual audio coding [5]. However, such a procedure is quite complicated to implement and we have not determined the effectiveness of this technique for our application.

A second idea proposed in [6] was to modulate each input signal with independent random noise

$$x_i'(n) = [1+\epsilon_i(n)]x_i(n), \qquad i = 1,2 \tag{32}$$

with ϵ_i two independent lowpass noise processes, as follows:

$$\epsilon_i(n) = \alpha\epsilon_i(n-1) + (1-\alpha)\nu_i(n)$$

where again ν_1 and ν_2 are two independent white noises. We have determined that these methods are not satisfactory either. Indeed, many experiments show that when we add or modulate a random noise (or a "foreign" signal) to the original signal, it is clearly heard even when its level is very low. This significantly degrades the quality of the speech.

To minimize the audible degradation, it is really preferable to add something like the original signal. But how can that be done? It is well-known that the coherence magnitude between two processes is equal to one if and only if they are linearly related, and this is what happens in the stereophonic case. The new idea here is to add to the signal a nonlinear function of the signal itself (Figs. 2 and 3) [7]:

$$x_i'(n) = x_i(n) + \alpha f[x_i(n)], \qquad i = 1,2. \tag{33}$$

The function f must be nonlinear to avoid a linear relation between x_1' and x_2', thus ensuring that the coherence magnitude

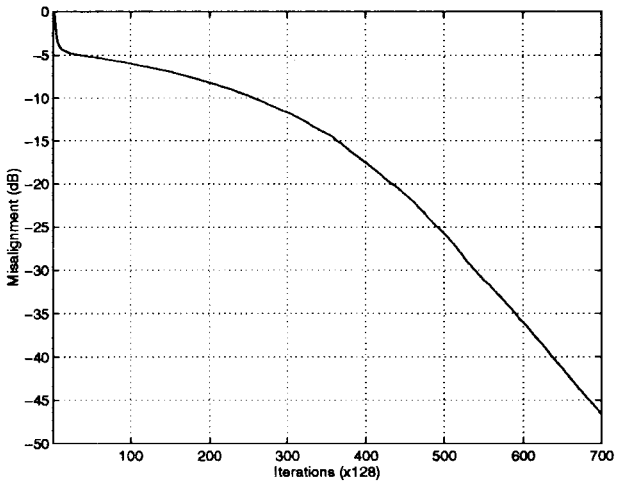


Fig. 4. Convergence of the misalignment when the length of the adaptive filters L is equal to the length N of the impulse responses in the receiving room (white noise source, measured room responses, $L = N = 1000$).

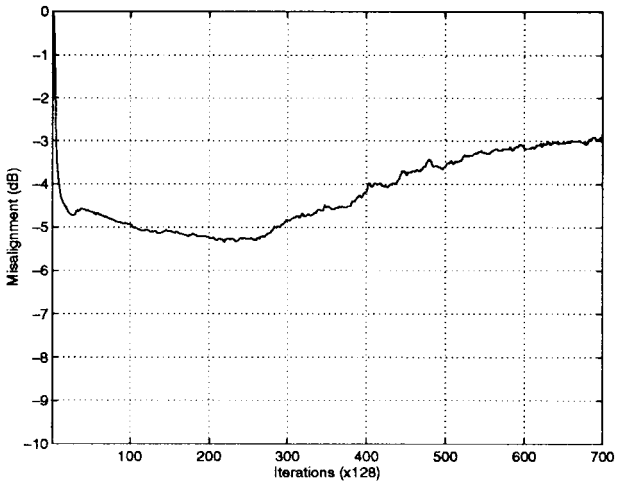


Fig. 5. Convergence of the misalignment when the length of the adaptive filters L is smaller than the length N of the impulse responses in the receiving room (white noise source, measured room responses, $L = 1000, N = 4096$).

will be smaller than one. Such a transformation reduces the coherence and hence the condition number of the covariance matrix, thereby improving the misalignment. Of course, this transformation is acceptable only if its influence is inaudible and has no effect on stereo perception.

Of the several nonlinear transformations that we have tried, a simple one that gives good performance is the half-wave rectifier

$$\tilde{x} = f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

Let us check in this case if the relation between $\mathbf{x}'_{1,M}$ and $\mathbf{x}'_{2,M}$ is linear or not. From (1) and (33), we deduce

$$\begin{aligned} &\mathbf{x}'^T_{1,M}(n)\mathbf{g}_{2,M} - \alpha\tilde{\mathbf{x}}^T_{1,M}(n)\mathbf{g}_{2,M} \\ &\quad = \mathbf{x}'^T_{2,M}(n)\mathbf{g}_{1,M} - \alpha\tilde{\mathbf{x}}^T_{2,M}(n)\mathbf{g}_{1,M} \end{aligned} \tag{35}$$

with

$$\begin{aligned} \mathbf{x}'_{i,M}(n) &= [x'_i(n) \quad x'_i(n-1) \quad \cdots \quad x'_i(n-M+1)]^T \\ \tilde{\mathbf{x}}_{i,M}(n) &= [\tilde{x}_i(n) \quad \tilde{x}_i(n-1) \quad \cdots \quad \tilde{x}_i(n-M+1)]^T, \\ & \qquad i = 1, 2. \end{aligned}$$

Therefore, there is a linear relation between $\mathbf{x}'_{1,M}$ and $\mathbf{x}'_{2,M}$ if and only if

$$\tilde{\mathbf{x}}^T_{1,M}(n)\mathbf{g}_{2,M} = \tilde{\mathbf{x}}^T_{2,M}(n)\mathbf{g}_{1,M}. \tag{36}$$

This can happen if

1) $\forall n \;\; x_1(n) \geq 0$ and $x_2(n) \geq 0$, and in this case $\tilde{\mathbf{x}}_{1,M}(n) = \mathbf{x}_{1,M}(n)$ and $\tilde{\mathbf{x}}_{2,M}(n) = \mathbf{x}_{2,M}(n)$;
2) $\exists a, \tau_1, \tau_2$ such that $a\tilde{x}_1(n-\tau_1) = \tilde{x}_2(n-\tau_2)$.

For example, if we have $ax_1(n-\tau_1) = x_2(n-\tau_2)$ with $a > 0$.

However, in practice these cases never occur because we always have zero-mean signals and $\mathbf{g}_{1,M}, \mathbf{g}_{2,M}$ are never related by just a simple delay.

Experiments show that stereo perception is not affected by our method even with α as large as 0.5. Also, the distortion introduced is hardly audible because of the nature of the speech signal and psychoacoustic masking effects [4]. This kind of distortion is also acceptable for some music signals but may be objectionable for pure tones like a flute produces.

We can also use other types of nonlinearities such as the square-law, the square-sign, and so on. However, in this paper we exclusively use the half-wave rectifier, which has been found to be as effective as any other nonlinearity, is simple to implement, and is self-scaling to signal level.

VII. Simulations

First, we wish to show, by way of simulation, the effect of the impulse response tails in the receiving room. The signal source s in the transmission room is white noise. The two microphone signals were obtained by convolving s with two impulse responses g_1, g_2 of length $M = 4096$, which were measured in an actual room (HuMaNet I, room B [8]). The microphone output signal y in the receiving room is obtained by summing the two convolutions $(h_1 * x_1)$ and $(h_2 * x_2)$, where h_1 and h_2 were also measured in an actual room (HuMaNet I, room A [8]) as 4096-point responses, which are subsequently truncated to N points. The sampling frequency rate is 16 kHz. For all of our simulations, we have used the two-channel fast recursive least squares (FRLS) algorithm [3], with $\lambda = 1 - 1/(10L)$; we also tried the normalized LMS algorithm but that was ineffective because of the extremely slow convergence of the misalignment due to the ill-conditioned nature of the solution. Fig. 4 shows the convergence of the misalignment for $L = N = 1000$ and Fig. 5 shows the corresponding response with $N = 4096$. For the purpose of smoothing the curves, misalignment samples are averaged over 128 points. We point out that, as expected, there is a great difference between these two results. The first gives good misalignment whereas the second is very bad because of the impulse response tails in the receiving room, which give rise to strong cross-correlation between the two input signals.

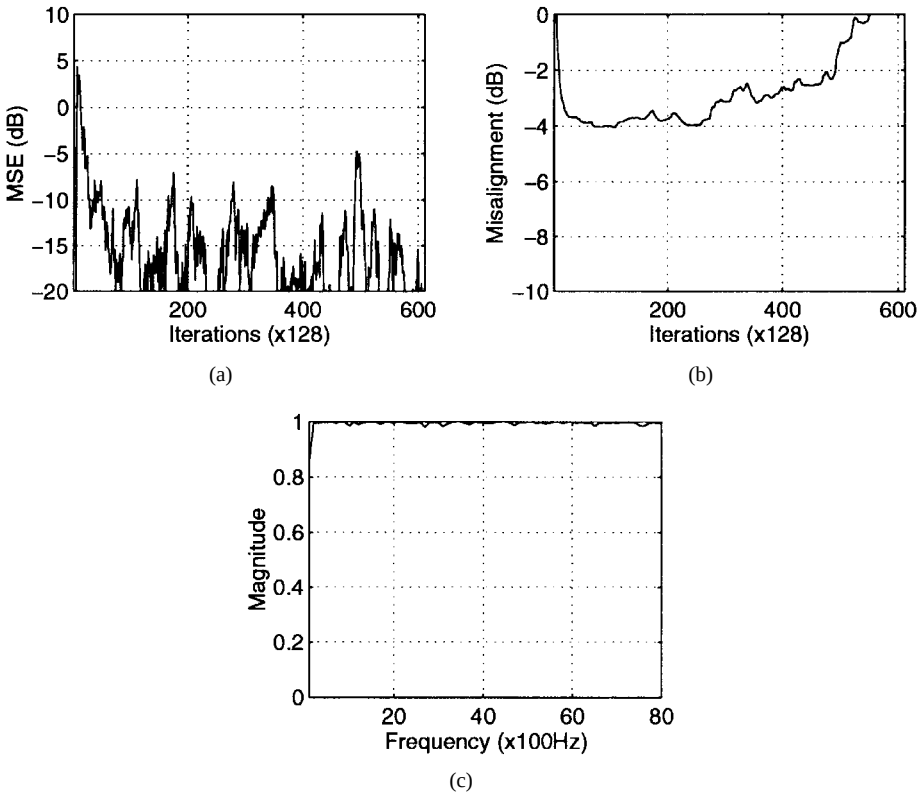


Fig. 6. Behavior of the (a) MSE, (b) misalignment, and (c) coherence magnitude with a pair of vertical microphone arrays and $\alpha = 0$ (speech source, measured room responses, $L = 1200, N = 4096$).

In the second part of these simulations, we show the effectiveness of our nonlinear transformation method using actual speech signals. Here we use the half-wave rectifier nonlinearity of Fig. 2 with $\alpha = 0.3$ as the NL block in Fig. 3. With this value, there is no audible degradation of the original signal and we can even increase α to 0.5 with only a slight degradation. Also, preliminary psychoacoustic experiments have shown that the stereophonic spatial localization is not affected. The length of the impulse responses both in the transmission and receiving room is $N = M = 4096$ in these simulations. The length of the two adaptive filters is taken as $L = 1200$; however, the misalignment was calculated only over the first 1000 samples, as there is negligible energy beyond that point. We again used the two-channel FRLS algorithm, but now the source in the transmission room is a speech signal sampled at 16 kHz, and consists of the following three sentences: "Bobby did a good deed," "Do you abide by your bid?", and "A teacher patched it up."

Figs. 6–10 show the behavior of the mean square error (MSE), the misalignment, and the coherence magnitude (for this latter, we use MATLAB's spectrum function and take the square root of the computed magnitude-squared coherence). For the purpose of smoothing the curves, error and misalignment samples are averaged over 128 points and coherence magnitude samples are averaged over 100 points.

For Figs. 6–8, the "talker" (actually a loudspeaker) is in the right rear of an actual room (HuMaNet I, room B [8]) and the pair of microphones in the transmission room (HuMaNet I, room A [8]) are vertical arrays. In Fig. 6, there is no nonlinear transformation of the input signals $(\alpha = 0)$; in Fig. 7, we use the half-wave rectifier with $\alpha = 0.3$; and in Fig. 8, a white noise with 30 dB SNR is added to the microphone signal $y(n)$. Again, the nonlinearity is seen to greatly reduce the misalignment. The addition of white noise deteriorates the misalignment; however, this could be compensated for somewhat by increasing α since the nonlinear distortion will be less noticeable in the presence of noise. We also note that the MSE in Fig. 7 is slightly increased in comparison with the MSE in Fig. 6. However, this is not a real problem since the MSE is bounded by the level of noise in the receiving room, which will usually dominate.

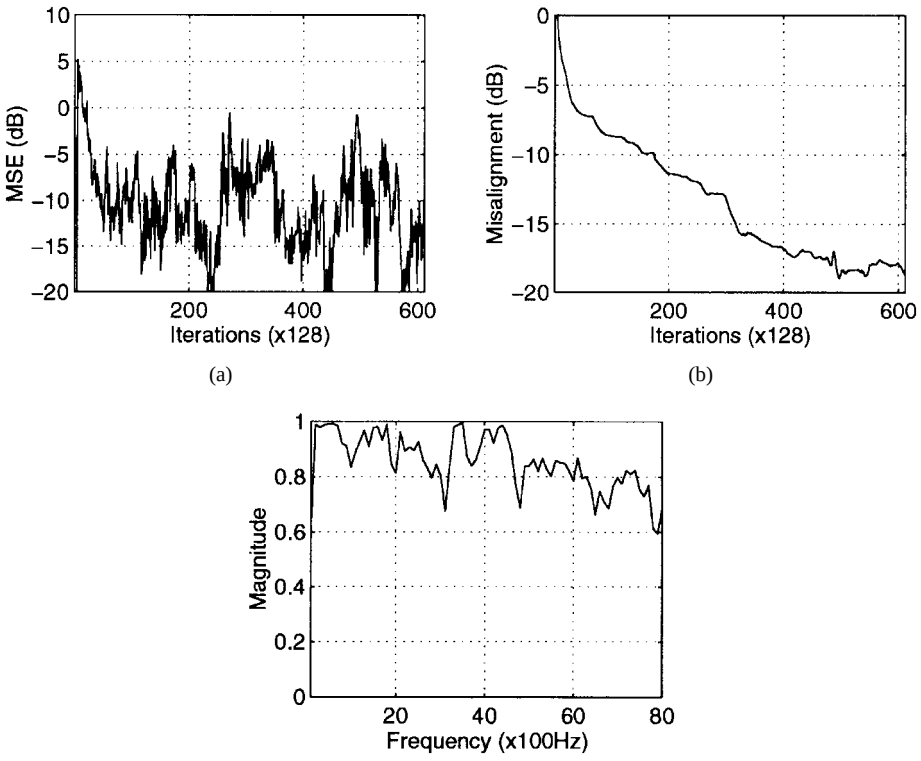
In order to further validate the method, we also simulated an image-derived [9] transmission room with cardioid microphones and source located in the rear of the room at the center. Figs. 9 and 10 give the results with, respectively, $\alpha = 0$ and $\alpha = 0.3$.

We note that in all of our simulations, the misalignment is greatly reduced by the nonlinear transformation.

VIII. Conclusion

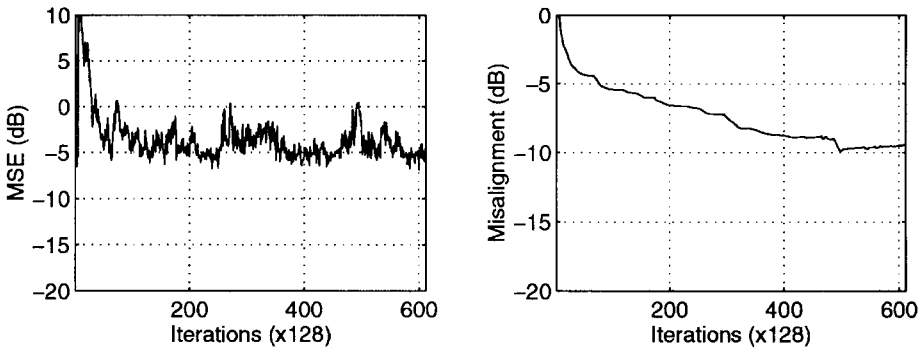
In this paper, we have given an original interpretation of the fundamental problems that occur in stereophonic acoustic echo cancellation, which are explained as the effect of the impulse response tails of the transmission and receiving rooms, respectively, on the condition number of the input signal covariance matrix and on the misalignment.

Thanks to this better understanding, we have proposed a new solution based on nonlinear transformations of the input signals to improve both the condition number of the



(a) (b) (c)

Fig. 7. Behavior of the (a) MSE, (b) misalignment, and (c) coherence magnitude with a pair of vertical microphone arrays and $\alpha = 0.3$ (speech source, measured room responses, $L = 1200, N = 4096$).



(a) (b)

Fig. 8. Behavior of the (a) MSE and (b) misalignment with a pair of vertical microphone arrays, $\alpha = 0.3$, and output SNR = 30 dB (speech source, measured room responses, $L = 1200, N = 4096$).

covariance matrix and the misalignment. Several simulations and experiments confirm our analysis and validate our method.

APPENDIX A
LINK BETWEEN THE COHERENCE FUNCTION AND THE COVARIANCE MATRIX

The covariance matrix of two concatenated stationary processes x_1 and x_2 is defined as

$$\mathbf{R} = E\left\{\begin{bmatrix}\mathbf{x}_1(n)\\ \mathbf{x}_2(n)\end{bmatrix}[\mathbf{x}_1^T(n) \quad \mathbf{x}_2^T(n)]\right\} = \begin{bmatrix}\mathbf{R}_{x_1x_1} & \mathbf{R}_{x_1x_2}\\ \mathbf{R}_{x_2x_1} & \mathbf{R}_{x_2x_2}\end{bmatrix} \tag{A1}$$

where $E\{\cdot\}$ denotes mathematical expectation

$$\mathbf{x}_i(n) = [x_i(n) \quad x_i(n-1) \quad \cdots \quad x_i(n-L+1)]^T, \quad i = 1,2$$

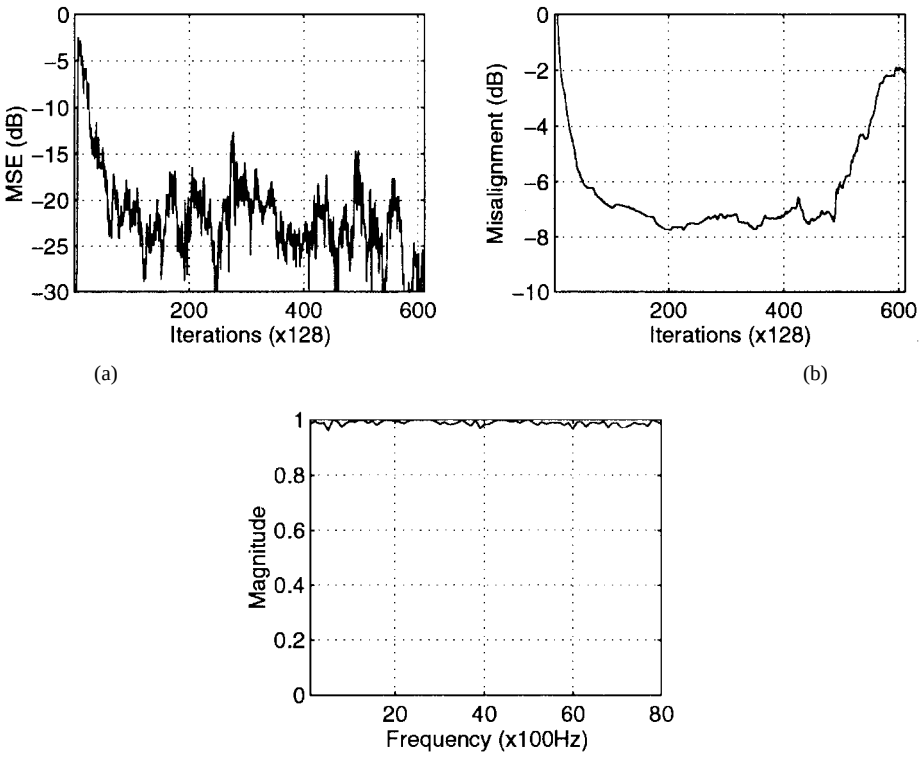
and $\mathbf{R}_{x_ix_j}$ are $L \times L$ Toeplitz matrices.

Now suppose $L \to \infty$. In this case, a Toeplitz matrix is asymptotically equivalent to a circulant matrix if its elements are absolutely summable, which is the case for the intended application. Hence, we can decompose $\mathbf{R}_{x_ix_j}$ as

$$\mathbf{R}_{x_ix_j} = \mathbf{F}^{-1}\mathbf{S}_{x_ix_j}\mathbf{F}, \qquad i,j = 1,2 \tag{A2}$$

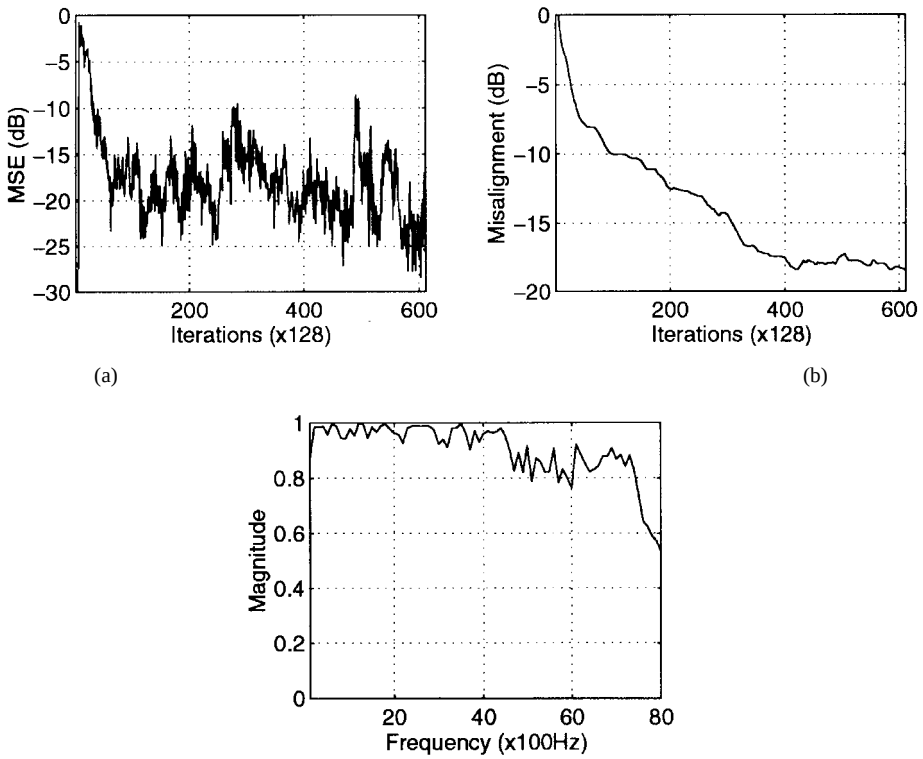
where $\mathbf{F}$ is the discrete Fourier transform matrix and

$$\mathbf{S}_{x_ix_j} = \text{diag}\{S_{x_ix_j}(0), S_{x_ix_j}(1), \cdots, S_{x_ix_j}(L-1)\} \tag{A3}$$



(a) (b) (c)

Fig. 9. Behavior of the (a) MSE (b) misalignment and (c) coherence magnitude with a pair of cardioid microphones and $\alpha = 0$ (speech source, image-derived transmission room responses, measured receiving room responses, $L = 1200, N = 4096$).



(a) (b) (c)

Fig. 10. Behavior of the (a) MSE (b) misalignment, and (c) coherence magnitude with a pair of cardioid microphones and $\alpha = 0.3$ (speech source, image-derived transmission room responses, measured receiving room responses, $L = 1200, N = 4096$).

is a diagonal matrix formed by the first column of $\mathbf{F}\mathbf{R}_{x_ix_j}$. With this representation, the covariance matrix $\mathbf{R}$ can be expressed in the frequency domain as

$$\mathbf{S} = \begin{bmatrix} \mathbf{F} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{F} \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{F}^{-1} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{F}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{x_1x_1} & \mathbf{S}_{x_1x_2} \\ \mathbf{S}_{x_2x_1} & \mathbf{S}_{x_2x_2} \end{bmatrix}. \quad \text{(A4)}$$

However, since $\mathbf{S}_{x_ix_j}$ $(i,j=1,2)$ are diagonal matrices, the eigenvalue equation of $\mathbf{S}$ (which is the same as that of $\mathbf{R}$) has the very simple form

$$\prod_{f=0}^{L-1} \{[S_{x_1x_1}(f)-\lambda][S_{x_2x_2}(f)-\lambda] - S_{x_1x_2}(f)S_{x_2x_1}(f)\} = 0 \quad \text{(A5)}$$

or, assuming that $\forall f, S_{x_ix_i}(f) \neq 0$ $(i=1,2)$

$$\prod_{f=0}^{L-1} \{S_{x_1x_1}^{-1}(f)S_{x_2x_2}^{-1}(f)\lambda^2 - [S_{x_1x_1}^{-1}(f) + S_{x_2x_2}^{-1}(f)]\lambda + 1 - |\gamma(f)|^2\} = 0 \quad \text{(A6)}$$

where $\gamma(f)$ is the coherence function (28) between the two signals x_1 and x_2 at frequency f. Thus, the eigenvalues are obtained from L quadratics. Expression (A6) shows that the minimum eigenvalue is lower bounded by the factor $[1-|\gamma(f)|^2]$ and that if any $|\gamma(f)|=1$, then the covariance matrix is singular. The eigenvalues, and hence condition number of the covariance matrix can be obtained by trivially finding the roots of the quadratic factors.

ACKNOWLEDGMENT

The authors would like to thank J. L. Hall for the experimental psychoacoustic work on nonlinear self-masking and stereo perception and S. L. Gay and G. W. Elko for helpful discussions.

REFERENCES

[1] M. M. Sondhi and D. R. Morgan, "Acoustic echo cancellation for stereophonic teleconferencing," in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1991.
[2] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation—An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.
[3] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099–3102.
[4] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. New York: Academic, 1989, ch. 3.
[5] J. D. Johnston and K. Brandenburg, "Wideband coding—Perceptual considerations for speech and music," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, ch. 4.
[6] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE ICASSP*, 1995, pp. 3059–3062.
[7] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1997, pp. 303–306.
[8] D. A. Berkley and J. L. Flanagan, "HuMaNet: an experimental human-machine communications network based on ISDN wideband audio," *AT&T Tech. J.*, vol. 69, pp. 87–99. Sept./Oct. 1990.
[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

Jacob Benesty was born in Marrakech, Morocco, on April 8, 1963. He received the Master's degree in microwaves from Pierre and Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991.

During his doctoral work (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecommunications (CNET), Paris, France. From January 1994 to July 1995, he was with Telecom Paris working on multichannel adaptive filters and acoustic echo cancellation. He joined Bell Laboratories, Lucent Technologies (formerly AT&T), in October 1995, first as a Consultant and then as a Member of Technical Staff. Since then, he has been working on stereophonic acoustic echo cancellation, adaptive filters, and blind deconvolution.

Dennis R. Morgan (S'63–S'68–M'69–SM'92) was born in Cincinnati, OH, on February 19, 1942. He received the B.S. degree in 1965 from the University of Cincinnati, and the M.S. and Ph.D. degrees from Syracuse University, Syracuse, NY, in 1968 and 1970, respectively, all in electrical engineering.

From 1965 to 1984, he was with the Electronics Laboratory, General Electric Company, Syracuse, NY, specializing in the analysis and design of signal processing systems used in radar, sonar, and communications. He is now a Distinguished Member of Technical Staff, Bell Laboratories, Lucent Technologies (formerly AT&T), where he has been employed since 1984: from 1984 to 1990, he was with the Special Systems Analysis Department, Whippany, NJ, where he was involved in the analysis and development of advanced signal processing techniques associated with communications, array processing, detection and estimation, and adaptive systems; since 1990, he has been with the Acoustics Research Department, Murray Hill, NJ, where he is engaged in research on adaptive signal processing techniques applied to electroacoustic systems.

Dr. Morgan has served as Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING since 1995.

Man Mohan Sondhi received the B.Sc. (honors) in physics in 1950, from Delhi University, Delhi, India, the D.I.I.Sc. in communications engineering in 1953 from Indian Institute of Science, Bangalore, India, and the M.S. degree in electrical engineering and Ph.D. degree in 1955 and 1957, respectively, both from the University of Wisconsin, Madison.

He joined Bell Laboratories, Lucent Technologies, Murray Hill, NJ, in 1962. Before joining Bell Laboratories, he was with the Avionics Division, John Oster Mfg. Co., Racine, WI, and spent a year at the Central Electronics Research Institute, Pilani, India. He also taught for one year at Toronto University, Toronto, Ont., Canada. He was a guest scientist for one year (1971–1972) at the Royal Institute of Technology, Stockholm, Sweden. He was also guest scientist for short periods at CNET, Lannion, France (1982), and NTT Human Interface Lab, Musashino, Japan (1990). His research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory, speech and visual processing by human beings, acoustical inverse problems, speech recognition, and analysis and synthesis of speech using articulatory models. He holds nine patents and has authored or coauthored over 95 published articles. He is co-editor, with S. Furui, of *Advances in Speech Signal Processing*.

Dr. Sondhi was Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING for several years, and was also a distinguished lecturer for two years. He is currently on the editorial board of the *International Journal of Imaging Systems and Technology*. He is a Bell Labs Fellow.