A BETTER UNDERSTANDING AND AN IMPROVED SOLUTION TO THE PROBLEMS OF STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Jacob Benesty *Dennis R. Morgan* *M. Mohan Sondhi*

Bell Laboratories, Lucent Technologies
700 Mountain Avenue, Murray Hill, NJ 07974
Email: { jb,drrm,mms }@research.bell-labs.com

ABSTRACT

Teleconferencing systems employ acoustic echo cancelers (AECs) to reduce echos 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.

1. INTRODUCTION

Acoustic echo cancelers (AECs) 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 AECs 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 AECs, however, is deceptive. Stereophonic AECs present problems that are basically different from those of single-channel AECs [2].

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.

2. THE NON-UNIQUENESS 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 non-uniqueness 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

$$\mathbf{x}_{i,M}(n) = \begin{bmatrix} x_i(n) & x_i(n-1) & \cdots & x_i(n-M+1) \end{bmatrix}^T,$$

$$\mathbf{g}_{i,M} = \begin{bmatrix} g_{i,0} & g_{i,1} & \cdots & g_{i,M-1} \end{bmatrix}^T, i = 1, 2.$$

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

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:

$$J(n) = \sum_{p=1}^{n} \lambda^{n-p} \left[y(p) - \hat{\mathbf{h}}_{1,L}^T(n)\mathbf{x}_{1,L}(p) - \hat{\mathbf{h}}_{2,L}^T(n)\mathbf{x}_{2,L}(p)\right]^2 \tag{2}$$

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

$$y(n) = \mathbf{h}_{1,N}^T\mathbf{x}_{1,N}(n) + \mathbf{h}_{2,N}^T\mathbf{x}_{2,N}(n) \tag{3}$$

is the microphone output and

$$\mathbf{h}_{i,N} = \begin{bmatrix} h_{i,0} & h_{i,1} & \cdots & h_{i,N-1} \end{bmatrix}^T,$$

$$\mathbf{x}_{i,N}(n) = \begin{bmatrix} x_i(n) & x_i(n-1) & \cdots & x_i(n-N+1) \end{bmatrix}^T,$$

$$\hat{\mathbf{h}}_{i,L}(n) = \begin{bmatrix} \hat{h}_{i,0}(n) & \hat{h}_{i,1}(n) & \cdots & \hat{h}_{i,L-1}(n) \end{bmatrix}^T,$$

$$\mathbf{x}_{i,L}(n) = \begin{bmatrix} x_i(n) & x_i(n-1) & \cdots & x_i(n-L+1) \end{bmatrix}^T.$$

The minimization of (2) leads to the normal equation:

$$\mathbf{R}(n)\begin{bmatrix}\hat{\mathbf{h}}_{1,L}(n)\\ \hat{\mathbf{h}}_{2,L}(n)\end{bmatrix}=\mathbf{r}(n) \tag{4}$$

where

$$\mathbf{R}(n)=\sum_{p=1}^{n}\lambda^{n-p}\begin{bmatrix}\mathbf{x}_{1,L}(p)\\ \mathbf{x}_{2,L}(p)\end{bmatrix}\begin{bmatrix}\mathbf{x}_{1,L}^T(p) & \mathbf{x}_{2,L}^T(p)\end{bmatrix} \tag{5}$$

is an estimate of the input signal covariance matrix and

$$\mathbf{r}(n)=\sum_{p=1}^{n}\lambda^{n-p}y(p)\begin{bmatrix}\mathbf{x}_{1,L}(p)\\ \mathbf{x}_{2,L}(p)\end{bmatrix} \tag{6}$$

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

$$\mathbf{h}_{i,L}=\begin{bmatrix}h_{i,0} & h_{i,1} & \cdots & h_{i,L-1}\end{bmatrix}^T,\ i=1,2.$$

These nonunique "solutions" are dependent on the impulse responses in the transmission room [2]. 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].

Let us examine two possible cases according to the length of the modeling filters:

(i) $L \geq M$

Consider the vector $\mathbf{u}=\begin{bmatrix}\mathbf{g}_{2,M}^T & 0 & \cdots & 0 & -\mathbf{g}_{1,M}^T & 0 & \cdots & 0\end{bmatrix}^T$, 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.

(ii) $L < M$

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

$$\mathbf{x}_{1,L}^T(n)\mathbf{g}_{2,L}+q_1(n-L)=\mathbf{x}_{2,L}^T(n)\mathbf{g}_{1,L}+q_2(n-L) \tag{7}$$

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 (7) we can see that the same is not true (in general) for $\mathbf{x}_{1,L}(n)$ and $\mathbf{x}_{2,L}(n)$; hence, in principle, the covariance matrix $\mathbf{R}(n)$ is full-rank, but it 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. 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.

3. THE MISALIGNMENT PROBLEM

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

$$\epsilon=\|\mathbf{h}-\hat{\mathbf{h}}\|/\|\mathbf{h}\|. \tag{8}$$

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.

We can easily show [4] in the mono-channel case by using the classical normal equation that if the length of the adaptive filter is smaller than the length of the impulse response in the receiving room, we introduce a bias in the coefficients of this filter. However, the problem of bad misalignment rarely appears in the monaural case for a reasonable length of the modeling filter with regard to the length of the impulse response. The same problem of course occurs in the stereo case but is much worse because of the strong cross-correlation between the input signals and the bad condition number of the covariance matrix [4].

To conclude this section, we can say that 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. So in practice we may have bad misalignment even if there is a unique solution to the normal equation.

4. 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.

5. 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:

$$x_i'(n) = x_i(n) + \nu_i(n),\ i = 1,2 \tag{9}$$

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

$$\gamma'(f) = \frac{\rho}{1+\rho}\gamma(f) \tag{10}$$

where ρ is the signal-to-noise ratio (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 $|\gamma(f)| = 1$ for a stationary source. A signal-to-noise ratio $\rho = 20$ (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 and has been used to advantage in perceptual audio coding. 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 [5] was to modulate each input signal with independent random noise:

$$x_i'(n) = [1 + \epsilon_i(n)]x_i(n),\ i = 1,2 \tag{11}$$

with ϵ_i two independent low-pass noise processes:

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

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 1 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:

$$x_i'(n) = x_i(n) + \alpha f[x_i(n)],\ i = 1,2. \tag{12}$$

The function f must be nonlinear to avoid a linear relation between x_1' and x_2', thus ensuring that the coherence magnitude will be smaller than 1. 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:

$$\tilde{x} = f(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

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 (12) we deduce:

$$\begin{aligned} &\mathbf{x}'^T_{1,M}(n)\mathbf{g}_{2,M} - \alpha\tilde{\mathbf{x}}^T_{1,M}(n)\mathbf{g}_{2,M} = \\ &\qquad \mathbf{x}'^T_{2,M}(n)\mathbf{g}_{1,M} - \alpha\tilde{\mathbf{x}}^T_{2,M}(n)\mathbf{g}_{1,M} \end{aligned} \tag{14}$$

with

$$\mathbf{x}'_{i,M}(n) = \left[\ x_i'(n) \quad x_i'(n-1) \quad \cdots \quad x_i'(n-M+1)\ \right]^T$$

$$\tilde{\mathbf{x}}_{i,M}(n) = \left[\ \tilde{x}_i(n) \quad \tilde{x}_i(n-1) \quad \cdots \quad \tilde{x}_i(n-M+1)\ \right]^T.$$

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

$$\tilde{\mathbf{x}}^T_{1,M}(n)\mathbf{g}_{2,M} = \tilde{\mathbf{x}}^T_{2,M}(n)\mathbf{g}_{1,M}. \tag{15}$$

This can happen if:

(i) $\forall n\ x_1(n) \geq 0$ and $x_2(n) \geq 0$.
In this case: $\tilde{\mathbf{x}}_{1,M}(n) = \mathbf{x}_{1,M}(n)$ and $\tilde{\mathbf{x}}_{2,M}(n) = \mathbf{x}_{2,M}(n)$.
(ii) $\exists a, \tau_1, \tau_2$ 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.

6. SIMULATIONS

In these simulations, we show the effectiveness of our nonlinear transformation method using actual speech signals. The signal source s in the transmission room is then 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?"
"A teacher patched it up."

The two microphone signals were obtained by convolving s with two impulse responses, g_1 and g_2 of length $M = 4096$ which were measured in an actual room (HuMaNet I, room B [6]). 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 [6]) as 4096-point responses, which are subsequently truncated to N points; a white noise with 40 dB SNR is added to the microphone signal y. The length of the adaptive filters is taken as $L = 1200$. For all of our simulations, we have used the two-channel FRLS algorithm [3], with $\lambda = 1 - 1/(12L)$; 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. Figure 2 shows the behavior of the misalignment when there is no nonlinear transformation of the input signals ($\alpha = 0$) and when we use the half-wave rectifier with $\alpha = 0.5$. (For the purpose of smoothing the curves, misalignment samples are averaged over 128 points.) With $\alpha = 0.5$ there is only a slight audible degradation of the original signal. Also, psychoacoustic experiments have shown that the stereophonic spatial localization is not affected.

We point out that, as expected, there is a great difference between these two results. The second ($\alpha = 0.5$) gives good misalignment whereas the first ($\alpha = 0$) is very bad because of the impulse response tails and the additive noise in the receiving room, both of which perturb the ideal solution [4].

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

ACKNOWLEDGMENTS

The authors would like to thank B. S. Atal, G. W. Elko, J. L. Hall, and S. L. Gay from Bell Labs for their support, encouragement, and fruitful 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, No. 8, August 1995, pp. 148-151.

[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] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," submitted to *IEEE Trans. Speech Audio Processing.*

[5] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE ICASSP*, 1995, pp. 3059-3062.

[6] 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.

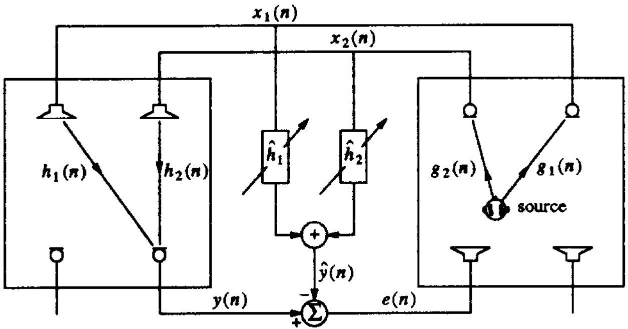

Figure 1. Schematic diagram of stereophonic echo cancellation.

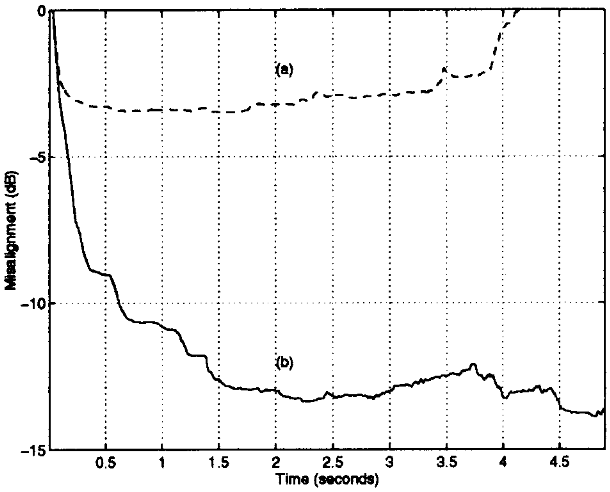

Figure 2. Behavior of the misalignment with (a) $\alpha = 0$ and (b) $\alpha = 0.5$. Output SNR = 40 dB (speech source, measured room responses, $L = 1200$, $M = N = 4096$).