IDENTIFYING THE TRUE ECHO PATH IMPULSE RESPONSES IN STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Fabrice Amand*, André Gilloire **, Jacob Benesty ***

* CEFRIEL, Politecnico di Milano, Via Emanuelli, 15, 20126 - Milano, Italy
e-mail: amand@mailer.cefriel.it

** CNET LAA/TSS/CMC Technopole Anticipa, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
e-mail: gilloire@lannion.cnet.fr

*** Lucent Technologies, Bell Labs Innovations, 600 Mountain Avenue, Murray Hill, NJ 07974, USA

ABSTRACT

A fundamental problem in stereophonic acoustic echo cancellation for teleconferencing is the possibility to identify the true impulse responses of the acoustic echo paths. This problem arises from the correlation between the two signals picked up in the remote room. We demonstrate by simple theoretical considerations and experiments that in real situations, due to the characteristics of the acoustic environment in the remote room, the identified impulse responses converge to the true echo path impulse responses.

1 INTRODUCTION

Stereophonic acoustic echo cancellation has been generally considered as a straightforward extension of the usual mono-channel scheme to the two-channel situation, as depicted in figure 1 (for sake of simplicity, only one half of the echo path system is shown in the local room, and the echo canceller dedicated to the remote room is not shown):

![Figure 1: Standard structure for stereophonic acoustic echo cancellation](image)

In this scheme, the acoustic echo paths $W_1$ and $W_2$ in the local room are modeled by adaptive FIR filters $H_1$ and $H_2$, which added outputs produce an estimate $\hat{y}$ of the true echo $y$. Indeed, the physical impulse responses $W_1$ and $W_2$ are of infinite length; nevertheless it is assumed that the filters $H_1$ and $H_2$ are "sufficiently long", in the sense that the tails of $W_1$ and $W_2$ not modeled by $H_1$ and $H_2$ have low energy and thus can be neglected. Speaking in the sequel of "true" impulse responses means that we only consider the first parts of $W_1$ and $W_2$ which contain most of the energy, and which are assumed of the same size $L$ as the model filters $H_1$ and $H_2$.

This paper addresses a fundamental problem in the stereophonic case, namely the possibility to identify the true echo path impulse responses, depending on the characteristics of the input signals $x_1$ and $x_2$. This problem was addressed in a recent paper [1], the conclusion of which was that these true impulse responses cannot be identified, unless ad-hoc processing is applied to the input signals $x_1$ and $x_2$ to decorrelate them. This conclusion is drawn from a theoretical analysis involving both finite length pick-up (i.e. source-to-microphone) impulse responses $G_1$ and $G_2$ in the remote room, and "clean" microphone signals, i.e. with zero uncorrelated noisy components $n_1$ and $n_2$. The purpose of our paper is to show both theoretically and experimentally that in real situations, i.e. when $G_1$ and $G_2$ are of infinite length and when there are non-zero, uncorrelated noisy components $n_1$ and $n_2$ in the input signals $x_1$ and $x_2$, the true responses $W_1$ and $W_2$ can be identified and that the adaptive filters do converge towards this true solution.

2 CONSIDERING AN IDEALIZED SITUATION

The impact of the correlation between the input signals $x_1$ and $x_2$ was already recognized as prominent to explain the performance of stereophonic acoustic echo cancellers, based on the scheme of figure 1 [2], [3], [4]. There is always a correlation because these input signals both contain the source signal filtered by the pick-up acoustic paths $G_1$ and $G_2$.

Let us consider the least mean squares solution $(H_1^{opt}, H_2^{opt})$ which minimizes the criterion $J(n) = E[\hat{e}^2(n)]$ w.r.t. the responses of the filters $H_1$ and $H_2$, where: the residual echo $e(n)$ is defined as: $e(n) = y(n) - \hat{Y}_1(n), H_1(n) - \hat{X}_2(n), H_2(n)$.

This solution satisfies the following system of linear equations (Wiener solution):
that the true impulse responses (displayed figure 2) are generally not found, as shown in figure 3. The differences are clearly obvious between the "left" impulse responses (W₂ figure 2b and H₂opt figure 3b). Note that the above results were obtained after convergence, using a standard recursive least squares (RLS) algorithm to adapt the filters H₁ and H₂. The source signal was a USASI noise (average speech spectrum). The asymptotic mse obtained was zero within the finite precision effects in the computations. Similar results were obtained with a standard LMS.

Consider that (i) the impulse responses G₁ and G₂ are of finite length M, and (ii) that the uncorrelated noise components n₁ and n₂ are zero. In that idealized situation, no unique solution to the system of linear equations above can be found, since the system matrix is not invertible [5].

\[
E\left[ X(n), X_1(n) \right] H_{1opt} + E\left[ X(n), X_2(n) \right] H_{2opt} = E\left[ x(n), X_1(n) \right] \
E\left[ X_2(n), X_1(n) \right] H_{1opt} + E\left[ X_2(n), X_2(n) \right] H_{2opt} = E\left[ x(n), X_2(n) \right]
\]

Moreover, let us assume that the filters H₁, H₂ and the echo paths W₁, W₂ (shown figure 2 for a particular case) have the same length L (here L=50). Then, the linear system above degenerates into a single vector equation:

\[
G'_1 H'_{1opt} + G'_2 H'_{2opt} = G'_1 W_1 + G'_2 W_2
\]

where the matrices G₁ and G₂ (of size Lx(L+M-1)) are Hankel convolution matrices corresponding to the acoustic pick-up channels G₁ and G₂.

Simulations performed with artificially created input signals according to this idealized model indeed show

Figure 2: true echo path impulse responses W₁ and W₂ (to be identified) - L=50

Figure 3: impulse responses identified in the ideal case M=50, no noise

The sizes of the pick-up acoustic paths in the remote room were taken equal to M=50. Other small sizes were used for these acoustic paths and led to similar results. L=50 was chosen to get asymptotic convergence when using the LMS.

3 A MORE REALISTIC SITUATION WHERE THE TRUE ECHO PATH RESPONSES ARE FOUND

Let us now turn to a "real" case, i.e. the input signals x₁ and x₂ were taken from recordings made in a real room.
Therefore, the impulse responses $G_1$ and $G_2$ have infinite length, and there are non-zero uncorrelated components $n_1$ and $n_2$ in the signals $x_1$ and $x_2$ (background noise in the room). The covariance matrices involved in the Wiener system above can be written:

$$E[X_i(n).X'_i(n)] = E[X_i(n).X'_i(n)] + \delta_{ij}E[N_i(n).N'_j(n)]$$

with $i,j=1,2$, $\delta_{ij}$ is the Kronecker symbol and $X_{si}$ denotes the component in $X_i$ correlated with the other input signal. Using obvious notations, the Wiener system reads now:

$$\begin{bmatrix}
R_{r_1x_1} + R_{n_1n_1} & R_{r_1x_2} \\
R_{r_2x_1} & R_{r_2x_2} + R_{n_2n_2}
\end{bmatrix}
\begin{bmatrix}
H_{1}^{opt} \\
H_{2}^{opt}
\end{bmatrix} =
\begin{bmatrix}
R_{r_1x_1} + R_{n_1n_1} & R_{r_1x_2} \\
R_{r_2x_1} & R_{r_2x_2} + R_{n_2n_2}
\end{bmatrix}
\begin{bmatrix}
W_1 \\
W_2
\end{bmatrix}$$

The block covariance matrix appearing in both sides of the above equation is invertible as soon as the covariance matrices of the noises $R_{n_1n_1}$ and $R_{n_2n_2}$ are both non-zero. This fact can be demonstrated easily, even in the case of impulse responses $G_1$ and $G_2$ of finite length: indeed, the block covariance matrix above contains a block diagonal component which is invertible. Another situation where the matrix is invertible occurs when there are several mutually uncorrelated active sources in the remote room, e.g. when two or more speakers are speaking simultaneously. In the real case considered, the unique solution of the system is:

$$\begin{bmatrix}
H_{1}^{opt} \\
H_{2}^{opt}
\end{bmatrix} =
\begin{bmatrix}
W_1 \\
W_2
\end{bmatrix}$$

Thus, using any adaptive filtering algorithm (LMS, RLS...) intended to estimate the Wiener solution, the adaptive filters $H_1$ and $H_2$ ultimately converge to the true impulse responses, as the comparison of figure 4 with figure 2 shows.

The results shown in figure 4 a-b were also obtained with a standard RLS adaptation algorithm. They correspond to the asymptotic convergence, the asymptotic mse being zero within the finite precision effects. It is worth noting that the convergence time is the same as in the case depicted in figure 3. The source signal in the remote room was a USASI noise, as in the previous experiment. The SNR at the microphone inputs was estimated at about 35 dB.

4 SIMULATION OF A "REAL" SITUATION

Another experiment was performed to get further insight in the effect of the additive noises $n_1$ and $n_2$. The input signals $x_1$ and $x_2$ were obtained as convolutions of a common source signal (speech sentence) with long (but finite) impulse responses $G_1$ and $G_2$ ($M=4096$), drawn from measurements performed in a real room. To simulate the room background noise ($n_1$ and $n_2$), two mutually uncorrelated white noise sequences were added, with various levels to get different "pickup SNRs" for the signals $x_1$ and $x_2$.

The results shown in figure 4 a-b were also obtained with a standard RLS adaptation algorithm. They correspond to the asymptotic convergence, the asymptotic mse being zero within the finite precision effects. It is worth noting that the convergence time is the same as in the case depicted in figure 3. The source signal in the remote room was a USASI noise, as in the previous experiment. The SNR at the microphone inputs was estimated at about 35 dB.

A version of the normalized LMS with independent normalizations of the adaptation step sizes for each channel was used to learn the filters $H_1$ and $H_2$.

The figure 5 shows the evolution along the time of the misalignment of the right adaptive filter $H_1$, for 4 different pick-up SNRs. The misalignment is defined as:

$$\text{Misalignment} = \frac{< (H_1(n) - W_1^*)^2 >}{W_1^* W_1}$$

where $< >$ denotes time averaging over 256 iterations. The lengths of $H_1$, $H_2$, $W_1$ and $W_2$ were fixed at $L=256$.

It appears clearly that when the noise level is increased (i.e. the pick-up SNR is decreased), the decay of the misalignment is faster, i.e. the adaptive filter converges faster towards the true echo channel impulse response. A similar behaviour was observed for the left filter $H_2$.

Note that with the lowest pick-up SNR (30 dB) yielding the fastest filter convergence, the noise was clearly audible when listening the signals $x_1$ and $x_2$, but not annoying. Note also that the usual MSE curves (decay of the short term power of the residual echo) did not show
significant differences among the various pickup SNRs used.

![Figure 5: misalignment of the right adaptive filter $H_1$ for different pickup SNRs in the remote room](image)

It is also worth noting that with a pickup SNR of 36 dB, corresponding approximately to the situation considered in section 3 (recorded signals), the misalignment still decreases slowly (only a slight improvement is observed w.r.t. the case without noise). This fact calls for adaptation algorithms with faster convergence than the LMS [5],[6].

The figures 6 a-b show the identified impulse responses after 38400 iterations of the adaptation algorithm (which correspond to the right end of the misalignment curves of figure 5). The agreement between the true (continuous line) and the identified (dots) impulse responses is fairly good. One can observe a peak at the beginning of the left identified response, coming from the influence of the other (right) channel due to the strong mutually correlated parts in the input signals $x_1$ and $x_2$.

5 CONCLUSION

The simple theoretical considerations and the experiments presented above demonstrate that in real situations, the stereophonic acoustic echo cancellation problem can be solved unambiguously, i.e. there is a unique solution to the problem since the identified impulse responses converge to the true echo path impulse responses. This fact is of particular importance, because this solution is no longer sensitive to unknown acoustic changes in the remote room. Nevertheless, appropriate adaptation algorithms must be used to achieve reasonably short convergence times and good tracking capability [5], [6].

REFERENCES


![Figure 6: identified impulse responses vs. true impulse responses after 38400 iterations](image)