CURRENT ISSUES IN STEREOPHONIC AND MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION

André Gilloire

CNET DIH/CMC, Technopole Anticipa, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
gilloire@lannion.cnet.fr

ABSTRACT

Acoustic echo cancellation is usually performed using single channel signals, i.e. 1 loudspeaker signal and 1 microphone signal. Enhanced quality teleconference systems will require multi-channel signals to achieve better speaker localization and sound spatialization. However, the problem of controlling the resulting multi-channel acoustic echo turns out to be much more difficult and complicated than in the single channel case, due to specific correlation characteristics of the loudspeaker signals. This paper describes and analyzes these difficulties, and it gives an overview of current solutions proposed for solving properly the problem in the stereophonic and multi-channel cases.

1 INTRODUCTION

It is recognized today that a significant step of improvement of teleconference systems with respect to sound aspects will rely on localization and spatialization clues, which will be provided to the parties in each teleconference room to help them to find more easily who is speaking and to get increased feeling of telepresence of the other parties. One way to provide these clues is to use in each room a stereophonic loudspeaker system (or more generally a multi-channel loudspeaker system) fed by adequate signals transmitted from the other rooms. It turns out that this way of doing deeply modifies the problem of acoustic echo handling with respect to the usual mono-channel systems. One purpose of this paper is to explain how the problem is modified and why it is much more difficult to solve. Another purpose is to describe typical algorithms and techniques that have been proposed to solve this problem and to discuss their efficiency.

Stereophonic acoustic echo cancellation has been generally viewed as a straightforward extension of the usual mono-channel scheme, as depicted in figure 1 [Son91]. Only one half of the echo path system is shown in the local room (where the echo originates), and the echo canceller dedicated to the remote room (where the speech of the distant speaker is picked-up) is not shown.

Let us assume that at some time in the remote room a unique speaker (source) is active, whose voice is filtered by the pick-up (i.e. source-to-microphone) impulse responses G_1 and G_2, and that mutually uncorrelated background noisy components n_1 and n_2 are also present in the signals x_1 and x_2 at the outputs of the microphones m_1 and m_2. The signals x_1 and x_2 are transmitted to the local (listening) room, where the dual echo canceller tries to model the acoustic echo paths W_1 and W_2 by using adaptive FIR filters H_1 and H_2 (of size L), which added outputs produce an estimate $\hat{y}$ of the true echo y. Some background noise (not shown) is also added to the microphone input in the local room.

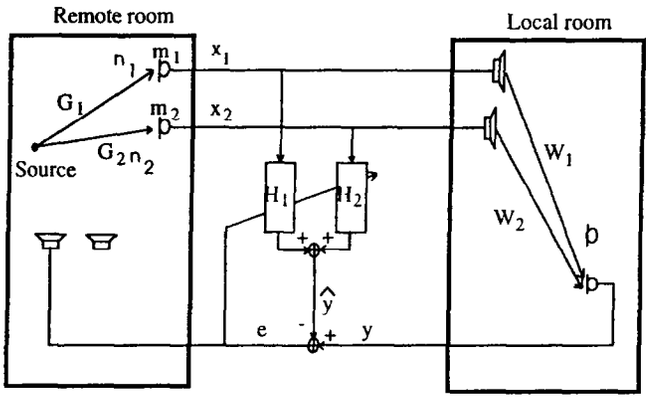

Figure 1: Basic scheme for stereophonic acoustic echo cancellation

Theoretical situations have been discussed to gain better understanding of the problem [Son95], [Ben95], [Ama96a], [Ben97], involving both finite length pick-up impulse responses G_1 and G_2 and "clean" microphone signals, i.e. without noisy components n_1 and n_2. We consider instead real situations where all the room impulse responses G_1, G_2, W_1 and W_2 are of infinite lengths, and noise components n_1 and n_2 are present.

2 CONSIDERING THE FUNDAMENTAL PROBLEM

The observed poor performance w.r.t. the usual monophonic case, of stereophonic acoustic echo cancellers based on the scheme of figure 1 is due to the correlation between the input signals x_1 and x_2. There is always a correlation because the input signals both contain the source signal filtered by the pick-up acoustic paths G_1 and G_2. A brief review of the problem is given below.

Let us define the estimation error or residual echo $e(n)$ as:

$$e(n) = y(n) - X_1^t(n).H_1(n) - X_2^t(n).H_2(n)$$

where $X_i(n) = [x_i(n) \quad \dots \quad x_i(n-L+1)]^t$, $i = 1,2$,

with ^{t} denoting transposition. Let us consider the least mean squares (Wiener) solution (H_1^{opt}, H_2^{opt}) which minimizes the criterion $J(n) = E\left[e^2(n)\right]$ w.r.t. the responses of the filters H_1 and H_2. The Wiener solution, which can be viewed as the asymptotic average of the solutions found by e.g. the LMS algorithm, satisfies the linear system:

$$\boldsymbol{R}\begin{bmatrix} H_1^{opt} \\ H_2^{opt} \end{bmatrix} = \boldsymbol{r} \qquad (1)$$

with :

$$\boldsymbol{R} = E\begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix}\begin{bmatrix} X_1^t(n) & X_2^t(n) \end{bmatrix} \ , \ \boldsymbol{r} = E\begin{bmatrix} y(n).X_1(n) \\ y(n).X_2(n) \end{bmatrix}$$

Similar equations can be obtained with other criteria, e.g. the weighted least squares one [Ben97]. Note that these equations can be straightforwardly extended to any number of channels by stacking the input signal vectors accordingly. The question of the invertibility of the system (1) led to many discussions linked to the sizes of the adaptive filters w.r.t. the sizes of the impulse responses G_1 and G_2. In real situations, it turns out that the matrix $\boldsymbol{R}$ is of full-rank, therefore the system (1) is invertible and it has a unique solution ; however, $\boldsymbol{R}$ may be more or less severely ill-conditioned. The considerations above can be illustrated by looking at the equations defining the input signals x_1 and x_2:

$$x_i(n) = \sum_{j=0}^{\infty} g_{i,j}\, s(n-j) + n_i(n) \ , \ i = 1,2$$

which can be written as:

$$x_i(n) = \sum_{j=0}^{L-1} g_{i,j}\, s(n-j) + r_i(n) + n_i(n) \ , \ i = 1,2$$

the additional terms $r_i(n)$, $i=1,2$ correspond to the convolution of the source signal by the remaining parts (tails of infinite lengths) of the impulse responses G_1 and G_2. These additional terms make the matrix $\boldsymbol{R}$ full-rank, except if the remaining parts of G_1 and G_2 have a huge number of common zeros, which is extremely unlikely [Ben97]. However, these additional terms may be very small depending on the characteristics of the impulse responses, hence at first view the matrix $\boldsymbol{R}$ may be extremely ill-conditioned. For example, with closely spaced pick-up microphones both facing the source and real speech, the condition number was found of the order of 10^9 for L=200 (other data can be found in [Ama95a). High ill-conditioning leads to severe degradation of the adaptive behaviour of the algorithm, which may « stick » for a long time to identified impulse responses very different from the « true » solution, i.e. the L first coefficients of each echo path W_1 and W_2. In real situations, the impulse responses G_1 and G_2 may change drastically within a short period, e.g. when two speakers in the remote room speak in turn, which may degrade severely the amount of echo cancellation since the echo canceller has to find a new solution which still depends on the current responses G_1 and G_2 as discussed below.

The convergence speed in terms of system distance (i.e. the norm of the difference vector between each filter and the L first coefficients of the corresponding echo path) also named *misalignment*, may be significantly improved by the effect of the noise components n_1 and n_2 which yield block-diagonal terms in the matrix $\boldsymbol{R}$:

$$R = \begin{bmatrix} R_{11} + R_{n1} & R_{12} \\ R_{21} & R_{22} + R_{n2} \end{bmatrix}$$

with :

$$\boldsymbol{R}_{ij} = E\left[\tilde{X}_i(n)\tilde{X}_j^t(n)\right] \ , \ i = 1,2$$

and, with obvious notations :

$$\boldsymbol{R}_{ni} = E\left[N_i(n)N_i^t(n)\right] \ , \ i = 1,2$$

with $\tilde{X}_i(n) = X_i(n) - N_i(n) \ , \ i = 1,2$.

The block-diagonal terms $\boldsymbol{R}_{n1}$ and $\boldsymbol{R}_{n2}$ introduce some kind of regularization which reduces the condition number of the matrix $\boldsymbol{R}$, hence improving the convergence of the adaptive filters [Ama96b]. Indeed, in practical teleconference situations the noise components have low levels, therefore the convergence improvement is generally modest.

Coming back to the « true » solution, it appears that the solution of the system (1) (or of an equivalent system corresponding to another criterion like the weighted LS) depends on the correlation of the input signals x_1 and x_2. This fact can be explained by the under-modelization (impulse response truncation) of the echo paths impulse responses W_1 and W_2 (of infinite sizes) modeled by the finite size filters H_1 and H_2. It is well known that in the mono-channel case under-modelization introduces a bias in the solution w.r.t. the « true » impulse response which depends on the correlation of the input signal. In the stereophonic case, it can be shown easily (full version of [Ben97]) that the solution of the system (1) is biased by the « tails » of the echo paths impulse responses according to :

$$\begin{bmatrix} H_1^{opt} \\ H_2^{opt} \end{bmatrix} = \begin{bmatrix} W_{1,L} \\ W_{2,L} \end{bmatrix} + \boldsymbol{R}^{-1}\boldsymbol{R}_t \begin{bmatrix} W_{1,t} \\ W_{2,t} \end{bmatrix}$$

with (assuming a finite tail size for the mathematics):

$$W_{i,L} = \begin{bmatrix} w_{i,0} & \dots & w_{i,L-1} \end{bmatrix}^t \ , \ W_{i,t} = \begin{bmatrix} w_{i,L} & w_{i,L+1} & \dots \end{bmatrix}^t \ , \ \text{for } i = 1,2 \ , \text{ and :}$$

$$\boldsymbol{R}_t = E\begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix}\begin{bmatrix} X_{1,t}^t(n) & X_{2,t}^t(n) \end{bmatrix}$$

with $X_{i,t}(n) = \begin{bmatrix} x_i(n-L) & x_i(n-L-1) & \dots \end{bmatrix}^t \ , \ i = 1,2$

Therefore, the bias depends not only on the auto-correlation of each input signal (as in the mono-channel case), but also on the inter-correlation between these signals, which may be relatively high as observed in [Ben97] ; fairly high misalignments may result, which depend on the acoustic conditions in the remote room. Note that in practice it is often 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 modelized by H_1 and H_2 have low energy and thus can be neglected. In any case the misalignment will have some detrimental effects that must be controlled accordingly.

3 INVESTIGATING CURRENT SOLUTIONS

The detrimental effect of the correlation between the input signals x_1 and x_2 on the performance of stereophonic acoustic echo cancellers has been early recognized [Son91], [Mah93] and many solutions to cope with this problems have been proposed. We can discriminate between solutions based on appropriate adaptive algorithms designed to be less sensitive to this effect, and solutions trying to decorrelate the signals before applying them to the reference inputs of the adaptive filters.

3.1 *Decorrelation techniques*

A fairly complete list of these techniques can be found in [Son95]. Let us consider and discuss some typical ones.

Decorrelation filters [Mah93]: the principle is to filter the input signals by adaptive decorrelating filters the outputs of which are mutually uncorrelated. This solution suffers from basic limitations because the uncorrelated components in the input signals may have very low energy ; moreover, the decorrelation filters may be very long, which would increase substantially the complexity.

Interleaving comb filters [Son95] : the principle is to filter the input signal by complementary comb filters which would null in each signal frequency components left unchanged in the other signal, hence drastically reducing the coherence between these signals. However, it was noted that this method cannot be applied to frequencies below 1000 Hz in order to avoid strong disturbances of the stereophonic effect ; therefore the method is ineffective since a large part of the correlation still remains in the signals.

Addition of random noise : the principle is to add to each input signal uncorrelated noise components which would improve the conditioning of the matrix $\boldsymbol{R}$, as the uncorrelated background noise components in the pick-up room do. However it was noted [Son95] that significant improvements of the conditioning would be achieved only with relatively high noise levels making the noise clearly audible and even disturbing ; nevertheless it was observed elsewhere [Ama96b] that this method is able to reduce significantly the misalignment with a SNR of 30 dB.

Use of nonlinear transformations [Ben97] : the principle is to add to each input signal non-linearly related components which would reduce the coherence between these signals. Simple transformations like half-wave rectification prove efficient, while perceptual self-masking properties make the resulting distortion only slightly audible for distortion amounts up to 50%, and the stereophonic effect is not affected. This method looks attractive since it is efficient though it has a very low computational cost.

3.2 *Appropriate multi-channel adaptive algorithms*

Many proposals of design of appropriate multi-channel adaptive filtering algorithms have been published (they consider mainly the two-channel case). The common goal of these algorithms is to limit as far as possible the degradation of the misalignment due to the correlation between the input signals. One may assume that the worst algorithm should be the Normalized Least Mean Squares (NLMS) since it is very sensitive to the conditioning of the covariance matrix of the input signals. This fact was indeed observed by many authors, e.g. [Ben95]. One can note that frequency-domain implementations of the multi-channel NLMS were proposed in [Mah93] and [Ama95a] both to reduce the computational complexity and to improve the convergence speed, taking advantage of the normalization of the adaptation step size in the frequency domain. Let us consider on the other side the two-channel recursive least squares algorithm (RLS), which minimizes the weighted squared error criterion $J(n)=\sum_{j=1}^{n}\lambda^{n-j}e^2(j)$, $0<\lambda<1$ being the forgetting factor. A numerically stabilized fast version of this algorithm is given in [Ben95]. As in the mono-channel case, the recursive least squares solution can be viewed as a reference of performance, since the convergence of the MSE is independent of the covariance matrix conditioning. Nevertheless even the fast versions have a prohibitive complexity which amounts to 28L multiplications per sample ; therefore lots of efforts were spent to find algorithms of lower complexity though yielding acceptable performance. We consider hereafter the two-channel case.

Algorithms with a single filter per channel [Hir92] : this proposal assumes a causality relation between the two input signals, namely one signal can be deduced from the other one by a causal filter. The single adaptive filter is intended to identify a mixture of the two echo channels W_1 and W_2 weighted in some way by the transfer functions in the remote room. This device may work properly in particular situations e.g. if one input signal is a delayed and weighted version of the other one; moreover it is very sensitive to changes in the transfer functions in the remote room. Therefore it is likely to fail in many practical situations.

Extended versions of the NLMS algorithm [Ama95b], [Ben95]: these so-called extended versions rely on rough approximations of the two-channel covariance matrix of the inputs, and the resulting algorithms can be viewed as degenerate forms of the two-channel RLS algorithm. A partial decorrelation of the inputs is thus obtained intrinsically in the algorithms, instead of being obtained by the means of external decorrelation filters as in [Mah93]. Note that these extended NLMS algorithms can also be viewed as variants of low-order projection algorithms discussed hereafter. These extended versions were found less sensitive to the correlation of the inputs than the two-channel NLMS in terms of convergence of the MSE.

Projection algorithms: these algorithms have been considered fairly extensively since they are able to decorrelate partially the input signals. In [Shi95] the authors explain how the two-channel projection algorithm can use the variations in time of the correlation between the input signals more efficiently than the two-channel NLMS. The same authors [Mak97] propose a subband implementation of the stereo projection algorithm both to gain better use of the variations of the correlation and to save computations. A generalized form of projection algorithms to the multichannel case is derived in [Ben96], where the authors introduce a constraint of orthogonality of each filter increment w.r.t. the signals of the other channels, in adition to the minimum norm of the increment classically used for the derivation of the affine projection algorithms. The resulting algorithm yields improved convergence w.r.t. the « standard » projection algorithm. Since the complexity of the algorithms described in the works cited above is still high, fast versions have a practical interest. In [Ama96c] a fast version of the two-channel projection algorithm with exponential weighting is derived, which has a complexity of 12L+O(P), P being the projection order (to be compared with the complexity $O(P^2L)$ of the « standard » versions). A more complete derivation of fast versions is given in [Ama96a], where a minimal complexity of 6L+O(P) is obtained.

Other adaptive filtering algorithms could be used as well, like the *multi-channel Fast Newton algorithm* [The95], which rely on AR models of the input signals of reduced order P (e.g. 10) suitable to speech ; hence its performance would be close to the one of the fast RLS whereas its complexity would be of the order of 4L+20P multiplications instead of 28L multiplications for the fast RLS algorithm.

Finally, a multi-channel variable loss scheme without any adaptive filter was proposed in [Hei95] ; this solution may be competitive in low cost applications.

4 CONCLUSION AND ACKNOWLEDGEMENTS

It is hoped that the given extensive - though not exhaustive - review of the basic problem of multi-channel acoustic echo cancellation and of algorithms and techniques for solving it will be helpful for guiding future work on the subject.

The author wishes to acknowledge Jacob Benesty for useful advices and informations from the full version of [Ben97].

REFERENCES

[Ama95a] F. Amand, J. Benesty, A. Gilloire, Y. Grenier : "Multi-channel acoustic echo cancellation", *in Proc. Int. Workshop on Acoustic Echo and Noise Control,* Roros, 1995, pp. 57-60.

[Ama95b] F. Amand, J. Benesty, A. Gilloire, Y. Grenier : "Un algorithme d'annulation d'écho stéréo de type LMS prenant en compte l'inter-corrélation des entrées", *in Proc. Quinzième Colloque GRETSI*, Juan les Pins, France, Sept. 1995, pp. 407-410 (in French).

[Ama96a] F. Amand, "Etude de l'Annulation d'écho multi-voie et application à la téléconférence de haute qualité", thesis. University of Rennes 1, March 1996 (in French).

[Ama96b] F. Amand, A. Gilloire, J. Benesty, "Identifying the True Echo Path Impulse Responses in Stereophonic Acoustic Echo Cancellation", *in Signal Processing VIII : Theories and Applications,* LINT Ed., Trieste, 1996, pp. 1119-1122.

[Ama96c] F. Amand, J. Benesty, A. Gilloire and Y. Grenier, "A fast two-channel projection algorithm for stereophonic acoustic echo cancellation", *in Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, 1996, pp. II-949-II-952.

[Ben95] J. Benesty, F. Amand, A. Gilloire and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation", *in Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, 1995, vol. 5, pp. 3099-3102.

[Ben96] J. Benesty, P. Duhamel and Y. Grenier, "A Multichannel Affine Projection Algorithm with Applications to Multichannel Acoustic Echo Cancellation", IEEE Signal Processing Letters, vol. 3, no.2, Feb. 1996, pp. 35-37.

[Ben97] J. Benesty, D.R. Morgan, M.M. Sondhi : "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation", *in Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1997, Munich, pp. 303-306. A full paper will appear in the IEEE Trans. On Speech and Audio Processing.

[Hei95] P. Heitkämper, M. Walker, "Stereophonic and Multichannel Hands-Free Speaking", *in Proc. Int. Workshop on Acoustic Echo and Noise Control,* Roros, 1995, pp. 53-57.

[Hir92] A. Hirano and A. Sugiyama, "A new multi-channel echo canceller with a single adaptive filter per channel", in Proc. IEEE Int. Symp. Circuits and Systems, 1992, vol. 4, pp. 1922-1925.

[Mah93] Y. Mahieux, A. Gilloire and F. Khalil, "Annulation d'écho en téléconférence stéréophonique", *in Proc. Quatorzième Colloque GRETSI*, Juan les Pins, France, Sept. 1993, pp. 515-518 (in French).

[Mak97] S. Makino, K. Strauss, S. Shimauchi, Y. Haneda, A. Nakagawa, "Subband Stereo Echo Canceller using the Projection Algorithm wit Fast Convergence to the True Echo Path", *in Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1997, Munich, pp. 299-302.

[Shi95] S. Shimauchi, S. Makino : "Stereo Projection Echo Canceller with True Echo Path Estimation ", *in Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, 1995, vol. 5, pp. 3059-3062.

[Son91] M.M. Sondhi and D.R. Morgan, "Acoustic Echo Cancellation for Stereophonic Teleconferencing", *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, 1991.

[Son95] M.M. Sondhi and D.R. Morgan, "Stereophonic Acoustic Echo Cancellation - An Overview of the Fundamental Problem", IEEE Signal Processing Letters, vol. 2, no.8, Aug. 1995, pp. 148-151.

[The95] S. Theodoridis, G.V. Moustakides, K. Berberidis, "A Fast Newton Multichannel Algorithm for Decision Feedback Equalization ", IEEE Trans. Signal Processing, vol. 43 no.1, Jan. 1995, pp. 327-331.