A COMBINED FDAF/WSAF ALGORITHM FOR STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Florence Alberge, Pierre Duhamel and Yves Grenier

ENST/SIG, 46, rue Barrault, 75634 Paris Cedex 13, France

ABSTRACT

Adaptive Acoustic Echo Cancellation in stereophonic teleconferencing is a very demanding application. Characteristics are : very large number of coefficients, non-stationary input (speech), (slowly) time-varying systems to be identified, plus the specific property that both stereo signals are intrinsically very correlated. Basic versions of stochastic gradient algorithms have difficulties to meet these requirements. We show that, in a multichannel framework, only a combination of techniques can result in an algorithm which convergence is governed by a quasi-diagonal matrix. Simulations with data recorded in a conference room demonstrate the improvement in convergence of our algorithm compared to the LMS.

1. INTRODUCTION

Teleconferencing systems are expected to provide a high sound quality. In particular listeners use spatial information to locate the voice of the person they are talking with. Thus, multichannel systems are of great practical importance. A stereophonic teleconferening system is depicted in figure 1. The two loudspeakers signals x_1 and x_2 are produced by a unique signal S, filtered respectively by G_1 and G_2, the impulse responses of the far-end room. Hence, the correlation matrix of the received signals x_1 and x_2 is singular under the assumptions that the far-end impulse responses are shorter than the local adaptive filters and that the far-end room has no background noise [1], [4], [3]. In practical situations, these assumptions are not met, and the matrix is not singular, but strongly ill-conditionned. This problem is not relevant if one makes use of the Recursive Least Squares algorithm for tuning the adaptive filters, but the resulting computational cost is very high. Hence, it is necessary to find methods with low arithmetic complexity, (i.e. of a stochastic gradient type) able to provide acceptable results in such a framework.

Few algorithms have been proposed in this context (see [1]). This paper proposes an algorithm specifically tuned for fast convergence in a multi-channel situation. It is emphasized that its convergence is governed by a quasi-diagonal matrix with coefficients close to one on the diagonal. The corresponding eigenvalue spread of the matrix is small, thus resulting in an improved convergence rate. The performances of this algorithm are compared to those of the two-channel LMS algorithm, both algorithms incorporating a suitable step variation strategy in order to take into account the energy variation of the speech signal.

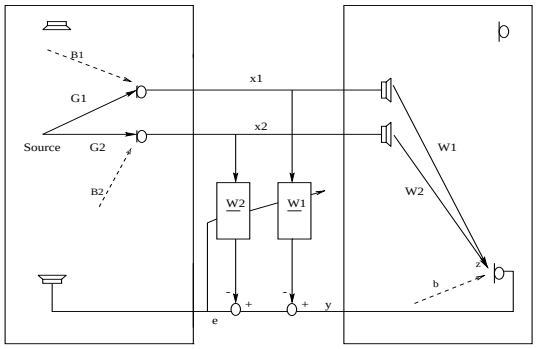

Figure 1: *Basic sheme for stereophonic acoustic echo cancellation*

2. THE TWO CHANNEL LMS ALGORITHM

The LMS adaptive algorithm minimizes the mean square error by updating the estimation of the filter as each new data sample is received. $x_1(n)$ and $x_2(n)$ are the two loudspeakers signals and y_n is the microphone one.

Let $W_1(n)$ and $W_2(n)$ be the two FIR adaptive filters, each one of length L. The symbol $(.)^T$ denotes transposition and E[.] stands for mathematical expectation. The notation σ_u^2 denotes the variance of signal u. Let,

$$W(n) = (W_1(n)^T \quad W_2(n)^T)^T$$

$$X(n) = (x_1(n), .., x_1(n-L+1), x_2(n))^T, .., x_2(n-L+1))^T$$

In the two channel case the error is $e(n) = y(n) - X(n)^T W(n)$, so the update equation of the filter writes:

$$W(n+1) = W(n) + \mu X(n) e(n)$$

where μ is a scalar stepsize. Obviously the matrix which governs the convergence of the LMS algorithm in the mean is $R_l = E[X(n)X(n)^T]$. R_l is plotted in figure 2 for L=30 and for a speech signal. Table 1 shows its condition number. Clearly, the condition number of the matrix governing the convergence is too large for the algorithm to work properly in a practical situation. More specifically, if one checks the distance between the estimated impulse response and the actual one (as shown in fig. 6), it is seen that this error decreases very slowly, even if the corresponding modeling error (as shown in fig. 5) shows an acceptable convergence. The problem comes from the expected changes in the echo paths (far-end or local) : if the echo path is not estimated

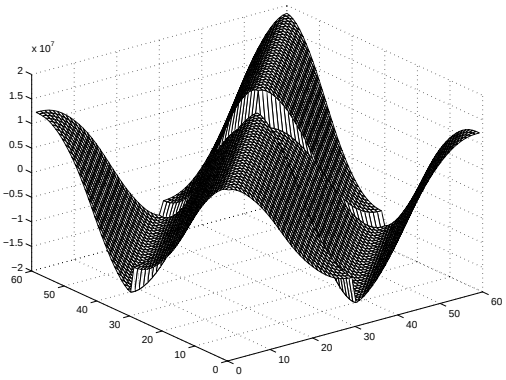

Figure 2: *Matrix governing the convergence of the LMS algorithm for L=30*

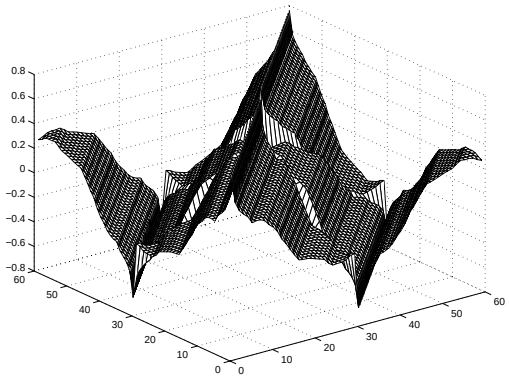

Figure 3: *Matrix governing the convergence of the WSAF for L=30 N=16 K=2*

precisely, any change in the spectrum of the input or in the far-end echo path will drastically increase the error. Classical improvements are based on transform-domain or Frequency-domain versions of the algorithm as well as subband adaptive filtering. In what follows, we illustrate that none of these methods alone is able to cope with the multichannel situation (despite a noticeable improvement), but that a mixed algorithm can make it.

3. THE TWO CHANNEL WSAF ALGORITHM

This algorithm has been chosen as a representative of the subband-based techniques. The single channel WSAF (Weighted Subbands Adaptive Filter) has been proposed in [2] in a monochannel context. This algorithm makes use of a subband decomposition of the error signal, and minimizes a sum of the appropriately weighted error components. This algorithm is of a block type, the block size being equal to the number of subbands. Let N be the number of filters in the filter bank and KN the length of the filters in the orthogonal filter bank, $H_{i,0\leq i\leq N-1}$ is one of the filter in the bank. Assume that,

$$X_l(n) = (x_l(n), ..., x_l(n-KN+1))^T \qquad l = 1, 2$$

$$\underline{X}_l(n) = (X_l(n), ..., X_l(n-L+1))_{KN\times L} \qquad l = 1, 2$$

$$\underline{X}(n) = (\underline{X}_1(n) \quad \underline{X}_2(n))$$

The error e_{kN+n}, $0 \leq n \leq N-1$, $k \geq 0$ is filtered by the N filters of the bank, thus producing N subbands error e_k^i, $0 \leq i \leq N-1$, $k \geq 0$. By definition, the criterion J^{WSAF} is the weighted sum of the N subbands errors, $J^{WSAF} = \sum_{i=0}^{N-1} \lambda_i E[|e_k^i|^2]$. The algorithm is obtained thanks to the evaluation of the instantaneous gradient estimate of the criterion [2], leading to the following update equation:

$$W((k+1)N) = W(kN) + \mu \sum_{i=0}^{N-1} \lambda_i X^i(kN) e_k^i \qquad (1)$$

where $X^i(n)^T = (x_1^i(n), .., x_1^i(n-L+1), x_2^i(n), .., x_2^i(n-L+1)) = H_i\underline{X}(n)$ is the nonsubsampled output of the i^{th} filter. The weights λ_i are chosen equal to $1/(L(\sigma^2_{x_1^i} + \sigma^2_{x_2^i}))$.

It was shown [2] that, if the filterbanks are very selective, this choice corresponds to the fastest convergence in each subband when $\mu = 1$. At this point, no specific care has been taken of the specificities of the two-channel case. Denote $\delta W(n)$ the error on the estimation of the filter $W(n)$ at time n. Under the asumptions that no noise is added to the microphone signal $y(n)$ and that x_1 and x_2 are uncorrelated from the adaptive filter taps, we obtain the equation describing the convergence in the mean,

$$E[\delta W((k+1)N)] = (I_{2L} - \mu \sum_{i=0}^{N-1} \lambda_i R_{X^i,X^i}) E[\delta W(kN)]$$

where $R_{X^i,X^i} = E[X^i(n)X^i(n)^T]$. R_{X^i,X^i} can be written in the form of a four block matrix as,

$$\begin{pmatrix} E[X_1^i(n)X_1^i(n)^T] & E[X_1^i(n)X_2^i(n)^T] \\ E[X_2^i(n)X_1^i(n)^T] & E[X_2^i(n)X_2^i(n)^T] \end{pmatrix} \qquad (2)$$

Figure 3 plots the matrix $R = \sum_{i=0}^{N-1} \lambda_i R_{X^i,X^i}$ for a speech signal and for L=30, N=16 and K=2. It is seen that the correlation matrix governing the convergence is better shaped (and conditioned, see table 1) that the initial LMS algorithm one, but the correlations between signals show up as strong sub-diagonals.

4. THE TRANSFORM-DOMAIN ALGORITHMS

When looking at the covariance matrices of the LMS algorithm and WSAF, it is clearly seen that the very strong correlation between the channels is still the problem. Hence, one could wonder what result would be provided by a very simple transform-domain algorithm, the transform being made from the simple sum and difference between the channels. Such a transform is the matrix $F = 1/\sqrt{2}\begin{pmatrix} I_L & I_L \\ I_L & -I_L \end{pmatrix}$. It is illustrated on table 1 that this approach is not sufficient to provide a noticeable improvement (in case of speech signal the condition number is only divided by 3).

5. THE TWO CHANNEL TD-WSAF ALGORITHM

When observing both correlation matrices of the WSAF and LMS, one can guess that the transform F will be more efficient on the WSAF than on the LMS algorithm. This can be explained as follows : The WSAF matrix is almost tridiagonal. When working with stationary signals, the coefficients of the diagonal within each block are equal. Moreover by constructions the values on the diagonal of R tend to be similar. Finally, R can be written : $\begin{pmatrix} a_1 I_L & a_2 I_L \\ a_2 I_L & a_1 I_L \end{pmatrix}$ with a_1 and a_2 two parameters depending on the input signals and on the filter bank. To improve the convergence rate of the WSAF we have to reduce the eigenvalue spread of R. This is done by diagonalizing R, using matrix F as defined above.

Now, if we come back to the WSAF and if we replace in (1) the adaptive filter $W(kN)$ with $W'(kN)=FW(kN)$ and $X^i(k)$ with $\tilde{X}^i(k) = FX^i(k)$ we obtain the WSAF in the transform domain. Since F is an orthogonal transform, the resulting algorithm is strictly equivalent to the WSAF.

Finally, as classically done in a TDAF, introduce a weight matrix to adjust the stepsize in an appropriate way for each tap. The corresponding update equation is:

$$W'((k+1)N) = W'(kN) + \mu,\ \sum_{i=0}^{N-1} \lambda_i \tilde{X}^i(k) e_k^i$$

with , $= diag(\gamma_1, ..., \gamma_{2L})$ and γ_j is the inverse of the power spectrum of the j^{th} entry of the vector $\sum_{i=0}^{N-1} \lambda_i \tilde{X}^i$. Notice that $\tilde{X}^i(k) = 1/\sqrt{2}(X_1^{iT}(k) + X_2^{iT}(k) \quad X_1^{iT}(k) - X_2^{iT}(k))^T$, so the transform consists only in replacing the inputs x_1 and x_2 respectivly with the sum $(x_1+x_2)/\sqrt{2}$ and the difference $(x_1-x_2)/\sqrt{2}$. Now we just have to compute the coefficients γ_j. With stationary signals $\gamma_1 = ... = \gamma_L$ and $\gamma_{L+1} = \gamma_{2L}$. So γ_1 and γ_{L+1} determine , . Since the analysis bank is composed of Losless Perfect Reconstruction filters ,the components $\tilde{X}^i(k)$ and $\tilde{X}^j(k)$ are uncorrelated for $i \neq j$. Then, $\gamma_1 = 2/\sum_{i=0}^{N-1} \lambda_i \sigma^2_{x_1^i + x_2^i}$ and $\gamma_{L+1} = 2/\sum_{i=0}^{N-1} \lambda_i \sigma^2_{x_1^i - x_2^i}$. So the matrix $R' = , FRF$ which governs the convergence of the TD-WSAF is quasi-diagonal with coefficients close to one when using stationary inputs. Then, its eigenvalue spread had been disminished compared to that of R. The matrix R' is plotted in (4) for speech signal with L=30, N=16 subbands and K=2.

In the next table we compare the eigenvalue spread of the matrices governing the convergence of the four algorithms we considered for a colored noise and a speech signal. We keep the values L=30 N=16, K=2 of the previous examples to run the algorithms.

Each part of the algorithm reduces the disparity

	LMS	TD	WSAF	TD-WSAF
colored noise	768.4	147.9	95.8	18
speech	7.10^6	$2.38.10^6$	2.10^3	723.3

Table 1: *Condition number for the matrix governing the convergence of LMS, TD, WSAF and TD-WSAF*

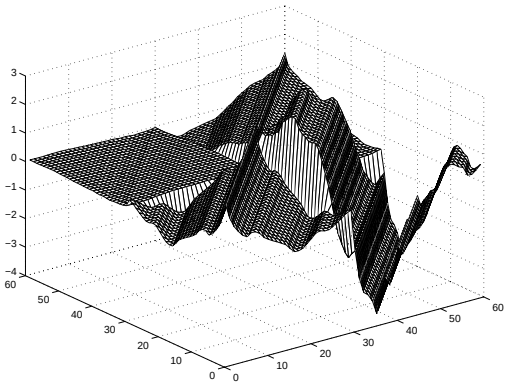

Figure 4: *Matrix governing the convergence of the TD-WSAF for L=30 N=16 K=2*

beetween the eingenvalues. When checking how this is obtained, it can be seen that such a result comes from a pre- and a post- multiplication of the LMS correlation matrix by appropriate quantities. With a speech signal, the eigenvalue spread of the matrix , FRF (TD-WSAF) is about 10^4 times smaller than the eigenvalue spread of the matrix governing the convergence of a LMS. The corresponding improvement in the convergence rate compared to the LMS is illustrated in the simulation section.

6. WEIGHTS IN NOISY SITUATIONS

When working with speech signals, one usually discovers that most adaptive algorithms are very sensitive to the variations of the input power. In some cases, one can observe a disadaptation in parts of the signal where the energy is not sufficient, and the global behavior is much worse than expected from simulations on stationary signals.

The origin of the problem is due to poor signal to noise ratios in some parts of the reference signal. The remedy is a time-varying strategy for the adaptation steps of the TDAF and of the WSAF. Let $b(n)$ be the white noise that is added to the microphone signal $y(n)$. All signals $x_1(n)$, $x_2(n)$, $y(n)$, $e(n)$, $b(n)$ are assumed ergodic and wide-sense stationnary. Noise and signals are assumed independent.

6.1. Transform Domain algorithm

Denote $S(n) = (s_1(n)...s_{2L}(n))^T$ the transformed input ($s_i(n)$ is the output of the i_{th} filter of the transform), $\epsilon(n) = \delta W^T(n) S(n)$ the noiseless error and $t_n = E[\delta W^T(n) \delta W(n)]$ the expectation of the norm of $\delta W(n)$, we have:

$$\delta W(n+1) = (I_{2L} - \mu,\ S(n)S(n)^T)\delta W(n) + \mu,\ S(n)b(n)$$

$$\begin{aligned} t_{n+1} &= t_n - \mu E[\delta W^T(n),\ S(n)\epsilon(n)] \\ &\quad -\mu E[\epsilon(n)S^T(n),\ \delta W(n)]) + \mu^2 (E[\epsilon(n)S^T(n),\ S(n)\epsilon(n)] \\ &\quad +E[b(n)^2 S^T(n),\ ^2 S(n)]) \end{aligned} \quad (3)$$

At convergence $t_{n+1} = t_n$ and $\sigma^2_{\epsilon_i} = \sigma^2_{\epsilon_j}$, $0 \leq i, j \leq 2L-1$. The output of two different filters are considered uncorrelated. Finally, define a_i as $E[|\epsilon_i^2(n)||s_i^2(n)|] = a_i \sigma^2_{\epsilon_i} \sigma^2_{s_i}$, then (3) becomes,

$$\sum_{i=0}^{2L-1} \mu\gamma_i(\mu\gamma_i\sigma_{s_i}^2((a_i+2L-1)\sigma_{\epsilon_i}^2+\sigma_{b_i}^2)-2\sigma_{\epsilon_i}^2)=0 \quad (4)$$

Assuming that the filters decorrelate sufficiently the signals, each term of the sum is zero (the error in the i^{th} subband is independant of all signals in the others subbands). Ideally, the error $\epsilon(n)$ should be an attenuated version of the microphone signal y(n). Then, we require the power $\sigma_{\epsilon_i}^2$ to be less than $\rho_i\sigma_{y_i}^2$, $0 \leq \rho_i \leq 1$ leading to:,

$$\gamma_i \leq \frac{2\rho_i}{\sigma_{s_i}^2((a_i+2L-1)\rho_i+\frac{\sigma_{b_i}^2}{\sigma_{y_i}^2})} \quad 0 \leq i \leq 2L-1 \quad (5)$$

In the case of the transform F and of stationary inputs, $\gamma_1 = ... = \gamma_L$ and $\gamma_{L+1} = ... = \gamma_{2L}$ which reduces the complexity. The variances $\sigma_{s_i}^2$, $\sigma_{y_i}^2$ and $\sigma_{b_i}^2$ are estimated with exponential windows.

6.2. WSAF algorithm

The expression of λ_i in a noisy environment is established in [2], for a monophonic acoustic echo canceller. The generalisation to the stereophonic case is straightforward:

$$4\lambda_i \leq \frac{2r_i}{L(\sigma_{x_1^i+x_2^i}^2+\sigma_{x_1^i-x_2^i}^2)(\alpha_i r_i+\frac{\sigma_{b_i}^2}{\sigma_{y_i}^2})} \quad 0 \leq i \leq N-1 \quad (6)$$

where α_i and r_i are similar to a_i and ρ_i in the previous section. In this equation the weights λ_i depend on the signal to noise ratio. The algorithm is able to slow down the adaptation when the reference data are excessively corrupted by the noise.

7. SIMULATIONS, CONCLUSION

In this section we compare the two-channel LMS algorithm against the TD-WSAF. The impulse response W_1 and W_2 to be identified are truncated to L=80 points. They were measured in an actual teleconference room. The length of the adaptive filters is also L=80. The input is a speech signal. White noise is added to the microphone signal $y(n)$, the output SNR is 30dB. The TD-WSAF has 64 subbands and K=2 (MLT). Both algorithms include a stepsize variation chosen to enable the fastest convergence rate. There is absolutely no vocal activity detection. Our time-varying strategy for the steps takes care of this problem. We plot in fig.5 the microphone signal power to the error signal power ratio in dB (ERLE) and in fig.6 the square norm of the estimation error on the filters ($||\delta W(n)||^2$). It is seen (fig.6) that the graph $\delta W(n)$ versus n decreases much faster for the TD-WSAF than for the LMS algorithm. The gap beetween the two graphs is increasing with time. Hence the TD-WSAF performs a better estimation of the filters than the LMS algorithm. Acoustic echo cancellation specifications are usually given in terms of rejection of the echo. Hence we should concentrate on the lower part of fig.5, when the rejection is smaller. The TD-WSAF clearly outperforms the LMS algorithm by several dB in this region (low energy part of speech). The TD-WSAF seems a good candidate for stereophonic acoustic echo cancellation.

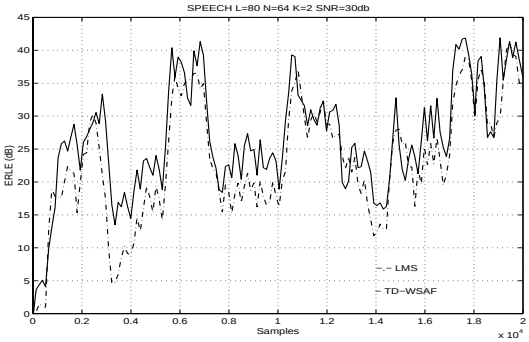

Figure 5: *ERLE for the LMS algorithm and TD-WSAF*

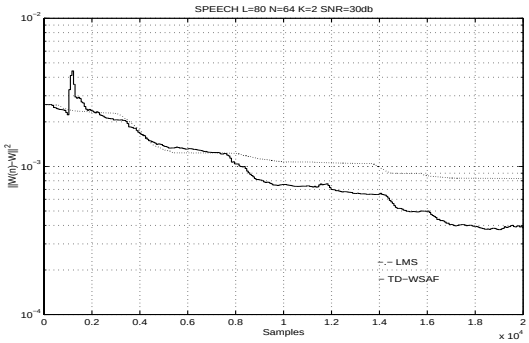

Figure 6: *Square norm of the estimation error of the filters*

8. REFERENCES

[1] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier. Adaptive filtering algorithms for stereophonic acoustic echo cancellation. In ICASSP Proc., 1995.

[2] M. de Courville and P. Duhamel. Adaptive filtering in subbands using a wheighted criterion. In ICASSP Proc, volume 2, pages 985–988, May 1995.

[3] S. Shimaushi and S. Makino. Stereo projection echo canceller with true echo path estimation.

[4] M.M. Sondhi, D.R. Morgan, and J.L. Hall. Stereophonic acoustic echo cancellation : An overview of the fundamental problem. IEEE SP Letters, 1995.