USING AUDITORY PROPERTIES TO IMPROVE THE BEHAVIOUR OF STEREOPHONIC ACOUSTIC ECHO CANCELLERS

André Gilloire and Valérie Turbin

CNET DIH/CMC, Technopole Anticipa, 2 Avenue Pierre Marzin -22307 Lannion Cedex, France

andre.gilloire@cnet.francetelecom.fr valerie.turbin@cnet.francetelecom.fr

ABSTRACT

We focus on the problem of stereophonic acoustic echo cancellation for teleconference applications. To limit the well-known detrimental effect of the correlation between the loudspeaker input signals, we propose a new method which consists in adding to these signals random noises controlled by auditory properties. We describe this method in some details and we show that its complexity can be fairly low. We demonstrate experimentally that the improvement yielded by this method is higher than the one provided by a former method based on the use of a non-linearity.

1. INTRODUCTION

Stereophonic (and more generally multi-channel) teleconference systems are likely to provide enhanced sound quality and tele-presence with respect to conventional mono-channel systems. However it has been early recognized that the problem of acoustic echo cancellation is much more difficult to solve in the stereophonic case than in the mono-channel case [1]. The reason why the problem is difficult comes from the correlation between the audio signals transmitted in the different channels; this fact is presently well understood [2], [3]. Nevertheless multi-channel acoustic echo cancellation algorithms that have been proposed are not yet entirely satisfactory [4], [5]. The purpose of this paper is to describe a solution based on the use of auditory masking properties which provides enhanced behaviour of stereophonic acoustic echo cancellers and which can be used in combination with any kind of adaptive filtering algorithms.

Stereophonic acoustic echo cancellation is generally viewed as a straightforward extension of the usual mono-channel scheme, as shown in figure 1. Only one half of the echo path system is shown in the local room (where the echo originates); the echo canceller operating in the remote room (where the speech of the distant speaker is picked-up) is not shown. In a typical situation a speaker (source) is speaking in the remote room. His voice is filtered by the two pick-up (source-to-microphone) impulse responses G_1 and G_2; the signals x_1 and x_2 at the outputs of the microphones m_1 and m_2 contain the corresponding filtered speech signals that are mutually correlated accordingly, and some noise components (room background noise, electronic circuits noise) n_1 and n_2 that can be considered as mutually uncorrelated.

In the local room where the signals x_1 and x_2 are received, the echo canceller computes a model of the acoustic echo paths W_1 and W_2 by using adaptive FIR filters H_1 and H_2, which added outputs produce an estimate $\hat{y}$ of the true echo y. Some background noise (not shown) is also present in the microphone signal y.

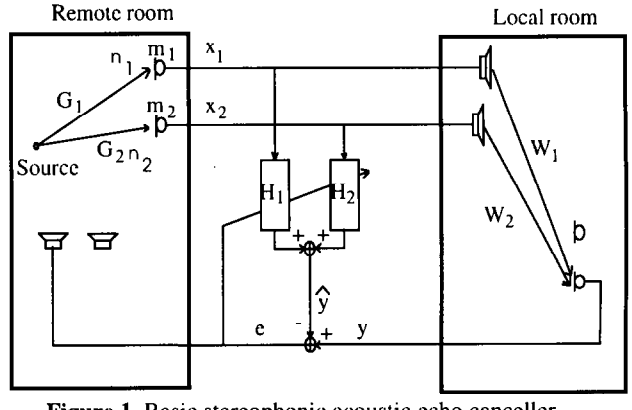

Figure 1. Basic stereophonic acoustic echo canceller

The correlation between the channel signals x_1 and x_2 makes the covariance matrix of these signals ill-conditioned; nevertheless the noise components n_1 and n_2 help to reduce the ill-conditioning. Thus a solution to improve the behaviour of the adaptive filters would be to add extra uncorrelated random noise components to each loudspeaker signal. It was noted however that the amount of extra noise components should be limited to avoid audible artefacts, thus this solution is poorly efficient if coarsely applied [6]. In this paper we show experimentally that this method can be efficient provided that the noise components are spectrally shaped according to auditory masking rules. Moreover, this method turns out to be more efficient than the method proposed in [2], which is based on the use of a non-linear function applied to each loudspeaker signal.

The paper is organized as follows. Section 2 recalls the fundamental problem of stereophonic acoustic echo cancellation. In section 3 the proposed method is described. Computational complexity aspects are discussed in section 4. Simulation results obtained with the proposed method are presented in section 5 and are compared to results obtained with the method described in [2]. Then we conclude the paper.

2. THE FUNDAMENTAL PROBLEM

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 $X_i(n) = \left[x_i(n) \quad \ldots \quad x_i(n-L+1)\right]^t$, $i = 1,2$, and ^{t} denotes transposition. The Wiener solution (H_1^{opt}, H_2^{opt}) minimizing the criterion $J(n) = E\left[e^2(n)\right]$ w.r.t. the responses of the filters H_1 and H_2 satisfies the system:

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

with:

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

Consider the equations defining the input signals x_1 and x_2:

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

The additional terms $r_i(n)$, $i=1,2$ which correspond to the convolution of the source signal by the tails of the impulse responses G_1 and G_2, as well as the noise components $n_i(n)$, $i=1,2$, make the matrix $\boldsymbol{R}$ full-rank, therefore the system (1) is invertible and it has a unique solution. However, these additional terms may be relatively small depending on the acoustic characteristics of the remote room, hence the matrix $\boldsymbol{R}$ may be more or less ill-conditioned.

It is shown in [2], [3] that the under-modelization (impulse response truncation) of the echo paths impulse responses W_1 and W_2 (of infinite sizes) by the filters H_1 and H_2 of size L makes the solution of the system (1) dependent on the correlation of the input signals x_1 and x_2. 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 finite tails sizes for the mathematics):

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

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

Thus, high correlation between the input signals, which makes the norm of the matrix $\boldsymbol{R}^{-1}$ very high, creates high misalignment (this latter one is defined as the norm of the difference vector between each filter and the L first coefficients of the corresponding echo path W_1 or W_2). Consequently the echo canceller may "stick" for a long time to identified impulse responses very different from the assumed "true" solution, i.e. the L first coefficients of W_1 and W_2. Since the correlation between the input signals may change drastically within a short period, e.g. when two speakers in the remote room speak in turn, the amount of echo cancellation may be severely degraded.

As shown in [6], the misalignment may be improved by the effect of the noise components n_1 and n_2 which yield block-diagonal terms in the matrix $\boldsymbol{R}$. These terms introduce some kind of regularization which reduces the condition number of the matrix, hence improving the behavior of the adaptive filters. Indeed, in practical teleconference situations the noise components have low levels, therefore the improvement is generally modest. Our purpose is to show that adding perceptually controlled noise components does lead to significant improvements.

3. THE PROPOSED METHOD

Addition of auxiliary signals to the loudspeaker input signals, x_1 and x_2, is the basis of the proposed method with great concern to keep the subjective speech quality unchanged. The idea is here to take advantage of human auditory properties, namely the simultaneous masking i.e. masking which happens in the frequency domain. We propose to add to each channel a random noise spectrally shaped so as to be masked by the presence of the loudspeaker input. To achieve proper masking it is necessary to control carefully levels of each additive auxiliary noise. The principle of this method is depicted figure 2.

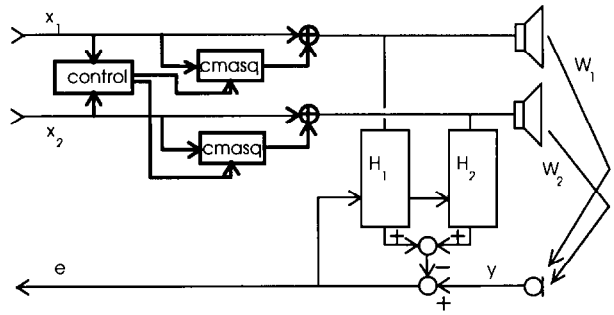

Figure 2. Principle of the proposed method

The "cmasq" boxes generate the masked noises. The "control" box enables to adjust properly the levels of the masked noises so as they are inaudible and the stereo perception is unchanged. The masking threshold under which the noise is masked can be easily computed since we have complete knowledge of the masking signals, i.e. the loudspeaker inputs $x_i(n)$, i=1, 2. This effect of noise masking by the human auditory system has been widely used in perceptual audio coding, and masking models have been developed accordingly. We have chosen to compute the masking thresholds by using a "hybrid" model described in [7] whose computation steps are recalled figure 3. The interest of the "hybrid" model is that it is easy to implement.

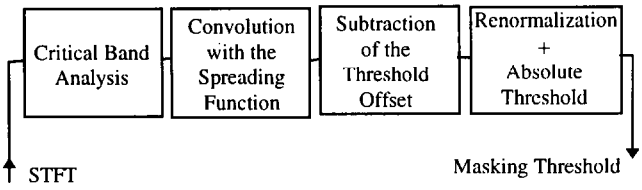

Figure 3. Computation steps of the masking threshold

Besides, we have done experiments which showed that the addition of a masked noise to each channel can affect the stereo spatialization even if it is inaudible at the output of each loudspeaker considered separately. This effect can be balanced by subtracting a correction coefficient to the masking threshold. We observed that this correction coefficient is closely dependent on the correlation of the loudspeaker signals and has to be higher for highly correlated signals. An interpretation of this effect can be given as follows: the spatial image provided by highly correlated input signals is well localized whereas the spatial image corresponding to the uncorrelated noises is spatially diffuse; therefore spatial fusion of the two images is not achieved

if the noise level is too high. We have then derived an empirical rule to adjust the correction coefficient accordingly.

The proposed method is basically independent of the adaptive algorithm. It enables to add decorrelated noises with higher levels compared to the former methods which do not take advantage of human auditory properties. This results in a better conditioning of the autocorrelation matrix and thus yields better convergence properties. Another advantage over "empirical" methods such as the use of a non-linear function [2] is that the additive signals are perceptually controlled with respect to the loudspeaker signals, which ensures that no audible degradation is generated.

4. COMPUTATIONAL COMPLEXITY

Prospects of stereophonic acoustic echo cancellation are numerous since many applications such as multimedia work-stations already use stereo sound in order to give spatial realism. It is then all the more important to improve algorithms behavior while keeping implementation cost compatible with real-time realizations. The new solution proposed in the previous section leads to an additional computational cost whatever the adaptive algorithm is. Nevertheless this additional cost is more or less critical depending on the implementation of the algorithm. Our method requires to know short term spectra of the input signals involved. Thus use of this method in a time-domain implementation would be complex and inappropriate. Frequency domain adaptive algorithms are adequate to make use of our approach with fairly low computational cost since they provide most of the spectral quantities necessary for the masking threshold calculation. In this case, the additional operations are limited to the computation steps described figure 3. Corresponding values in terms of additions and multiplications are reported in table 1, where N_b is the number of critical bands and Fast Fourier Transforms of length 2N are used.

Critical Band Analysis	Convolution with the spreading function	Subtraction of the Threshold Offset
N+1	$N_b(2N_b-1)$	$2(N+1)+4*N_b+6$

Table 1. Number of operations per block of data

For example, using N=512 and $N_b = 22$ (which corresponds to a sampling frequency of 16 kHz), the overall number of operations is 1676. Note that the renormalization (last step of figure 3) is not performed since it has no significant effect on the masking threshold for our application. Note also that the computation of the tonality index involved in the subtraction of the threshold offset needs table look-up operations not counted in the table 1.. The part of the algorithm dedicated to the control of the spatialization discussed in the previous section has a low complexity, since it relies essentially on comparisons of the short-term spectra of the input signals.

5. EXPERIMENTAL RESULTS

We have carried out simulations using real speech signals and impulse responses measured in a telecconference room. The pick-up impulse responses G_1 and G_2 of size 8192 samples correspond to closely spaced unidirectional microphones directed 90° apart; the source was placed on the symmetry axis of the microphone pair at a distance of 1.75m. This arrangement led to highly correlated input signals x_1 and x_2. The microphone signal y in the local room was obtained by summation of the convolutions of the input signals x_1 and x_2 with the impulse responses W_1 and W_2 respectively, of size 3840 samples. A white noise (snr ~ 40 dB) was added at the microphone input to simulate local room noise. The length of the adaptive filters of the stereophonic echo canceller was L=1024 samples.

The experimental results presented below compare three pre-processing methods: no modification of the input signals (M0), application of the non-linear distortion proposed in [2] with α=0.5 (MNL), addition of masked noise (MMN).

Two adaptive filtering algorithms were evaluated. The first one (2-FRLS) is a numerically stabilized version of the two-channel fast recursive least squares algorithm described in [8]; the value of the forgetting factor was 1-1/(12L) as in [2]. The second one (2-NLMS) is a straightforward generalization of the normalized least mean squares algorithm to the two-channel case.

5.1 Effect on the Coherence of the input signals

The figure 4 shows the magnitude squared coherence (MSC) of the loudspeaker signals obtained with the three considered methods. The MSC was averaged over a sentence of about 3 seconds. It can be seen that the MSC of the unmodified signals (M0) is high, especially in the high frequencies, which corresponds to an ill-conditioned covariance matrix $\boldsymbol{R}$ [3]. The non-linear processing (MNL) reduces significantly the MSC in the high frequencies, but our proposed method (MMN) appears more efficient, since the MSC is very much reduced even in the medium frequencies. Thus better behavior of the adaptive algorithms can be expected with this latter method.

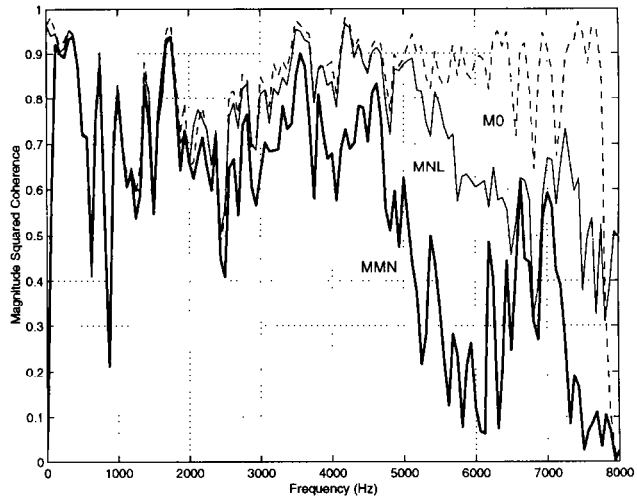

Figure 4. MSC of the two input signals for the three methods

5.2 Effect on the misalignment

The figure 5 shows the misalignment obtained with the 2-FRLS algorithm on the filter H_2 w.r.t. the left impulse response W_2. It

appears that both methods MNL and MMN lead to an acceptable behavior of the algorithm, i.e. an overall decreasing misalignment. Note that our method provides an improvement of 2-3 dB over the method MNL. The overall higher misalignment obtained, in comparison with [2], may be due to more critical sound pick-up conditions used in our experiments.

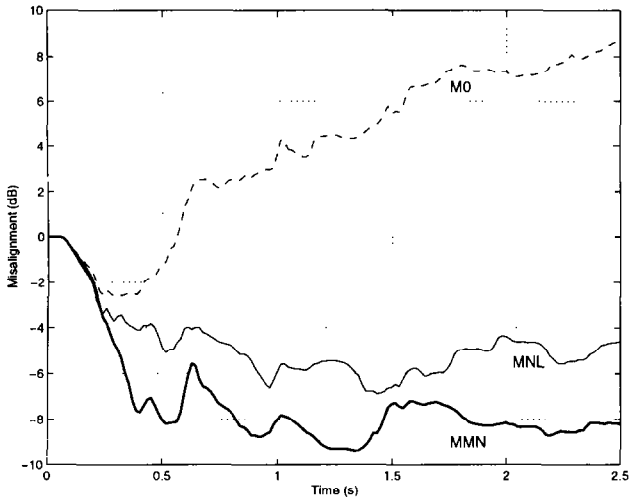

Figure 5. Misalignment obtained with the 2-FRLS

The figure 6 shows the misalignment obtained with the 2-NLMS algorithm on the filter H_2. The simulation was carried out on several consecutive sentences to take into account the low convergence speed of the algorithm. Similar conclusions can be drawn: both methods MNL and MMN improve the behavior of the algorithm, and the method MMN yields an improvement of a few dB over the method MNL.

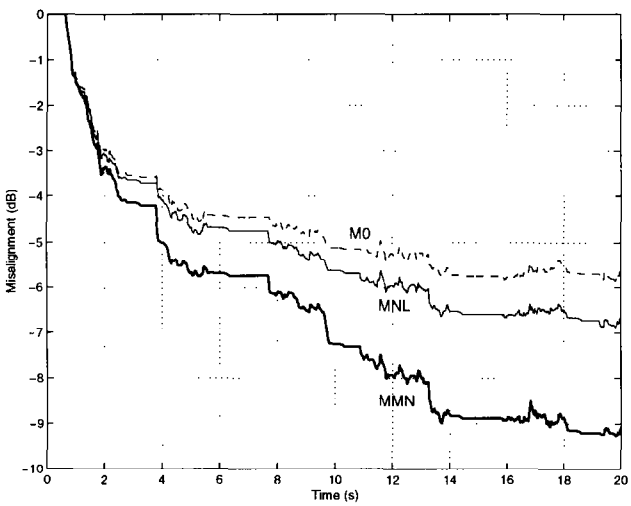

Figure 6. Misalignment obtained with the 2-NLMS

We observed in our experiments that the behavior of the Mean Square Error (MSE) obtained at the output of the echo canceller was fairly independent of the method used. This result confirms that the MSE is insufficient to characterize the behavior of multi-channel echo cancellers. Finally it appears that neither MNL nor MMN seem able to provide very low misalignments; however it is thought that achieving about -10 dB within less than 1 second, would be acceptable to avoid severe increase of the residual echo when the pick-up conditions in the remote room are drastically and quickly changed.

6. CONCLUSION

We have developed and tested a method based on the use of human auditory properties, which improves the behavior of stereophonic acoustic echo cancellers. This method was found more efficient than a previous proposal based on the use of a non-linear function. Moreover this method does not yield an important additional computational cost if frequency domain implementations of adaptive filters are used.

7. REFERENCES

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

[2] Benesty J., Morgan D. R. and Sondhi: M. M. "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation". *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Munich, April 1997, pages 303-306.

[3] Benesty J., Morgan D. R. and Sondhi M. M. "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation". *To appear in the IEEE Trans. On Speech and Audio Processing*, beginning 1998.

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

[5] Gilloire A. "Current Issues in Stereophonic and multi-Channel Acoustic Echo Cancellation". *Proceedings of IWAENC*, London, September 1997, pages K5-K8.

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

[7] Turbin V., Gilloire A., Scalart P. and Beaugeant C. "Using Psychoacoustic Criteria in Acoustic Echo Cancellation Algorithms". *Proceedings of IWAENC*, London, September 1997, pages 53-56.

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