USING PSYCHOACOUSTIC CRITERIA IN ACOUSTIC ECHO CANCELLATION ALGORITHMS

Valérie Turbin, André Gilloire, Pascal Scalart, Christophe Beaugeant

CNET DIH/CMC, 2, Avenue Pierre Marzin, 22307, Lannion Cedex, FRANCE

{turbin, gilloire, scalart, beaugean}@lannion.cnet.fr

ABSTRACT

We propose a general approach based on optimal filtering and use of psychoacoustic constraints to achieve acoustic echo cancellation which is applied in two contexts: teleconferencing and mobile telephones in cars. In the teleconferencing context, the acoustic echo cancellation system is composed of a conventional echo canceller combined with an optimal filter. In the mobile telephony context where not only the acoustic echo but also the ambient noise are to be cancelled, we propose to reduce globally both disturbances with only one optimal filter. We show that using psychoacoustic criteria in the optimal filter computation enables to reduce the distortion generated on the near-end speech especially when the perturbator is the acoustic echo.

1 INTRODUCTION

Recent studies have shown that acoustic echo cancellation can be performed by the combination of an echo canceller of reduced size and of an optimal post-filter [1]. An advantage of this method is to reduce the implementation cost while maintaining a high amount of echo reduction. However noticeable distortion of the near-end speech is produced by the post-filter. A way of reducing the distortion is to take into account human auditory properties in the filter computation.

In this contribution, we show the impact of including psychoacoustic criteria in our echo processing method. We first describe our echo reduction approach, which is based on optimal filtering, in two application contexts: teleconferencing and hands-free telephones in cars. In a second part, we discuss some auditory models and explain our choice of a particular one. The use of psychoacoustic criteria in the filtering method is then described. Results of simulations show that the distortion generated by our echo cancellation system can be efficiently reduced by the use of human auditory properties.

2. OPTIMAL ECHO CANCELLATION

The principle of our approach is based on a general concept of disturbance reduction and can be directly derived from noise reduction techniques. Let us assume that the observation signal $u(t)$ (e.g. microphone signal) can be written as follows:

$$u(t) = s(t) + p(t)$$

where $s(t)$ and $p(t)$ are uncorrelated signals and stand respectively for the near-end speech and the perturbation signal. Depending on the context, $p(t)$ may be composed either of the acoustic echo (teleconference) or both of the acoustic echo and of the ambient noise (mobile telephony). The linear optimal filter G, which gives an estimate of the near-end speech in the sense of the MMSE, can be expressed in the frequency domain as:

$$G(m,f) = \frac{SPR(m,f)}{1+SPR(m,f)} \qquad (1)$$

where m stands for the current block, f for the frequency. SPR stands for Signal to Perturbation Ratio and corresponds to the ratio between the power spectral density (psd) of the near-end speech and the one of the perturbation. We used the method originally designed for noise cancellation proposed in [2] to estimate the SPR, which is efficient and free of artifacts such as musical noise. It was found able to cope with the echo as well:

$$SPR(m,f) = \beta.\frac{\left|\hat{S}(m-1,f)\right|^2}{\hat{\gamma}_p(m,f)} + (1-\beta).P\left(SPR_{post}(m,f)\right)$$

$$\mathrm{SPR}_{post}(m,f) = \frac{\left|U(m,f)\right|^2}{\hat{\gamma}_p(m,f)} - 1, \quad P(x) = \frac{1}{2}\left(x+|x|\right) \qquad (2)$$

where $0<\beta<1$. $\hat{\gamma}_p$ is an estimate of the perturbation psd and U, $\hat{S}$ the Short Time Fourier Transform (STFT) of respectively $u(t)$ and the near-end speech estimate as returned by the filter G. In fact, this approach basically consists in applying a Wiener filtering to the observation signal knowing an estimation of the disturbance psd. We have developed solutions to the acoustic echo cancellation problem following these guidelines that we have applied to the teleconferencing and mobile radiotelephony contexts.

2.1 Application in a teleconferencing context

In this context where the echo path impulse response is long (because of the time reverberation much higher in teleconferencing rooms than those measured in cars), we propose the use of a "combined system" derived from [1] as shown figure 1.

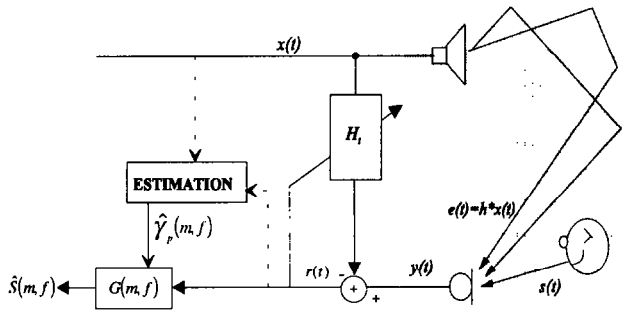

Figure 1. Combined acoustic echo cancellation system.

Acoustic echo cancellation is achieved by the combination of two filters. A partial echo attenuation is first performed by a conventional echo canceller H_t of reduced size L. The echo signal $e(t)$ which results from the convolution of the loudspeaker signal $x(t)$ and of the echo path impulse response h may be written as:

$$e(t) = \sum_{i=0}^{L-1} h(i)x(t-i) + \sum_{i=L}^{N-1} h(i)x(t-i) = e_d(t) + e_r(t) \quad (3)$$

where N is the assumed length of the impulse response. We may assume that the estimation given by the echo canceller is equal to $e_d(t)$. This is not strictly exact since the echo canceller yields a biased estimate of $e_d(t)$ dependent of the autocorrelation of $x(t)$; nevertheless the overall behaviour of the system is not modified by the bias. Additional echo reduction is obtained by the post-filter G whose task is to attenuate the residual echo $e_r(t)$. The post-filter G obeys eq. (1) and in that case, with the previous notations, we get: $u(t)=r(t)$ and $p(t)=e_r(t)$. We may rewrite $e_r(t)$ as follows:

$$e_r(t) = H_{N-L}^T x_{N-L}(t-L) = h_r * x(t-L) \quad (4)$$

where h_r is the impulse response associated to the vector $H_{N-L}^T = [h(L) \cdots h(N-1)]$ and $x_{N-L}(t-L)$ is a vector of loudspeaker observations of dimension N-L defined by: $x_{N-L}(t-L) = [x(t-L) \; x(t-L-1) \cdots x(t-N+1)]^T$.

Thus, the residual echo psd estimation problem is equivalent to the estimation of the transfer function H_r of hr. During echo only events, we get an estimate of this transfer function as follows:

$$|H_r(m,f)|^2 = \frac{|R(m,f)|^2}{|X_r(m,f)|^2} \quad (5)$$

where R is the STFT of the signal $r(t)$ and X_r the STFT of the appropriately delayed signal $x(t)$. During double talk events, the estimation of H_r is frozen and we use the last estimate of H_r to obtain an estimation of the residual echo psd.

A major advantage of this combined system is that it does not require the identification of the complete echo path impulse response. For example, with an impulse response of length N=4096, the combination of an echo canceller of size L=512 with the post-filter G yields a high amount of echo reduction, the echo being hardly audible at the system output. With the classical echo cancellation approach, an adaptive filter of several thousands coefficients is necessary to obtain similar results.

2.2 Application in a mobile radiotelephony context

In this context, the near-end speech $s(t)$ may be corrupted not only by the acoustic echo $e(t)$ but also by the ambient noise $n(t)$. Reduction of these two perturbations is classically achieved by two separate processings. Here, we propose to apply a global processing as shown figure 2.

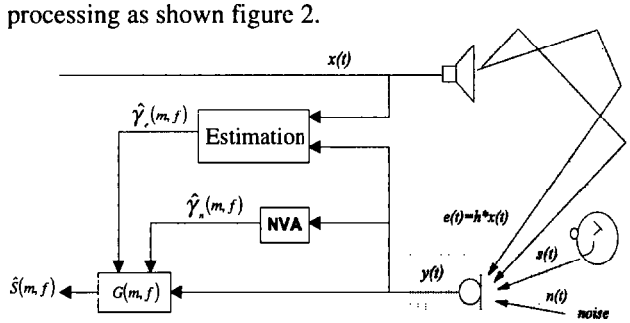

Figure 2. Echo and Noise reduction system.

The idea consists in considering the acoustic echo and the ambient noise as a unique disturbance $p(t)$, i.e:

$$p(t) = e(t) + n(t)$$

The filter G obeys (1). The filter input signal is the microphone signal, i.e. $u(t)=y(t)$. Estimation of the echo and noise psds is necessary to provide an estimation of the disturbance psd. The noise psd is estimated during Non Vocal Activity periods (detected into the box labeled "NVA"). Besides, the echo $e(t)$ and the loudspeaker signal $x(t)$ being correlated, we get an estimation of the echo psd as follows:

$$\hat{\gamma}_e(m,f) = \frac{|\gamma_{yx}(m,f)|^2}{\gamma_x(m,f)} \quad (6)$$

where γ_x and γ_{yx} are respectively the psd of the signal $x(t)$ and the cross-psd between $y(t)$ and $x(t)$. The filter G can be implemented in different ways which lead to different results both in terms of disturbance reduction and distortion of the near-end speech [3].

This approach is interesting because it does not require the identification of the echo path impulse response to perform acoustic cancellation. Moreover, it enables to get rid of the interaction between the acoustic echo cancellation processing and the noise reduction processing by applying a global filtering.

3. USE OF PSYCHOACOUSTIC CRITERIA IN THE OPTIMAL FILTERING

The main drawback of the two systems presented in the previous section is that the optimal filter may generate audible distortion of the near-end speech. So, it is important to reduce this distortion to improve the speech transmission quality of our systems. When two sounds occur simultaneously, it can happen that one is made inaudible by the other one: this effect is commonly

known as the masking phenomenon [4]. This means that when the near-end speech masks the echo, there is no need to filter. Limiting the filtering to frequencies where the echo is not masked should enable to reduce the distortion while maintaining the same perceptual amount of echo reduction.

3.1 Choice of the masking model

The incorporation of masking properties in our system is based on a human hearing model. This model yields a spectral masking threshold. A listener tolerates the presence of the perturbator (echo, noise) as long as it is below this threshold. We considered the Johnston's masking model [5] and the ISO MPEG Psychoacoustic Model II [6]. The different steps necessary to determine the masking threshold are illustrated figure 3.

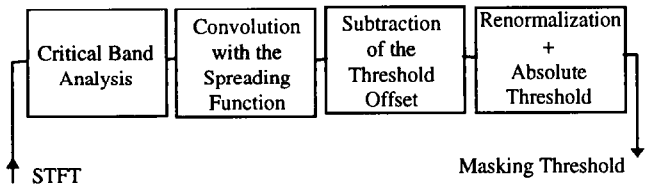

Figure 3. Calculation of the masking threshold.

The two models can be distinguished by a different threshold offset, which can be expressed as:

$$O(b) = \alpha(b)TMN(b) + (1-\alpha(b))NMT(b) \qquad (7)$$

where b stands for the Bark frequency, $TMN(b)$ is the value to apply in the case of a tone masking a noise, and $NMT(b)$ the value to apply in the case of a noise masking a tone. A constant value of 5 dB is used for $NMT(b)$ by the two models. TMN is dependent on the Bark frequency. Higher values are used in the ISO model. Moreover, in this latter, one has to evaluate a coefficient of tonality $\alpha(b)$ which varies with the critical band and which requires a significant cost of computation. Whereas, in the Johnston model, a unique coefficient of tonality is used for all critical bands and the computational cost is much lower. The model we incorporate in our systems is a "hybrid" model: we consider the Johnston's model whose TMN values have been replaced by the ISO TMN values. An advantage of this "hybrid" model is to improve masking performance of Johnston's model in low frequencies while preserving a reasonable computational cost for the global system.

3.2 Integration of psychoacoustic constraints in the optimal filtering

The computation of the spectral masking threshold $T(m,f)$ requires to know the near-end speech. Among several possible solutions to find an estimate of this signal, we have chosen to use a simple one, i.e. the output of the optimal filter, which has proven to work fairly well in our experiments. The threshold $T(m,f)$ enables to determine frequencies for which the perturbator is masked by the near-end speech. For these frequency components, the filter gain is forced to 1. The output of the system with psychoacoustic constraints is then given by:

$$\hat{S}(m,f) = \begin{cases} U(m,f) & \text{if } \gamma_p(m,f) \le T(m,f) \\ G(m,f)U(m,f) & \text{if } \gamma_p(m,f) > T(m,f) \end{cases} \qquad (8)$$

Note that this method was previously proposed for noise reduction techniques [7].

4 SIMULATIONS AND RESULTS

Psychoacoustic constraints were integrated in our optimal echo cancellation systems as described in the previous sections. The experimental conditions of the different simulations were carefully controlled in order to compare performance of our optimal echo cancellation systems to the one of the corresponding systems under psychoacoustic constraints. The impact on the near-end speech distortion was evaluated by informal listening tests and by an objective measure: the cepstral distance between the original near-end speech and the same signal filtered by the considered system. We noted that audible distortion occured when the average cepstral distance raised above approximately the value 1.

Impulse responses measured in a teleconferencing room and in a car interior were used with several values of the echo loss through the echo path. In the mobile telephony context, noise measured in a moving car was also added to the near-end speech with different values of the Signal to Noise Ratio. In both contexts, results of simulations showed that the distortion of the near-end speech can be efficiently reduced by the introduction of masking properties in our echo cancellation systems, provided that the Signal to Echo Ratio is greater than 0 dB. As an illustrative example, results in a teleconferencing context are presented figure 4. They correspond to a Signal to Echo Ratio at the microphone input of 3 dB and an echo canceller which provides about 12 dB of echo attenuation. With these conditions, the remaining echo components at the output of the post-filter are hardly audible: Figure 4(b) shows very high ERLE values during echo only periods and fairly high (about 20 dB) ERLE values during double talk. Moreover, the system with psychoacoustic constraints allows a significant reduction of the near-end speech distortion, while maintaining the same perceived amount of echo reduction. In fact, both systems (without and with psychoacoustic constraints) provide the same ERLE values during echo only periods and ERLE values somewhat smaller (up to 5 dB) are yielded by the system with psychoacoustic constraints during double talk. Whereas a significant distortion reduction is obtained when the perturbator is the acoustic echo, it appears that the use of psychoacoustic criteria is less efficient when the main disturbance is noise. Simulations in the mobile telephony context showed that the use of psychoacoustic

constraints in the filtering did not lead to a significant reduction of the near-end speech distortion. This can be explained by the fact that in this case the spectra of *p(t)* and *s(t)* do not overlap, contrary to the echo case.

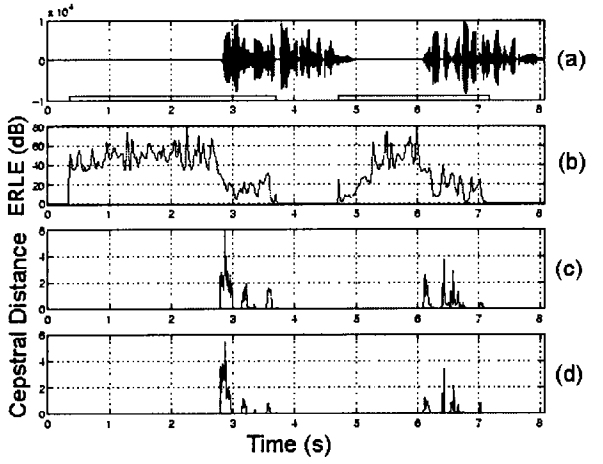

Figure 4. Results in a teleconferencing context.
(a) Near-end speech (echo events indicated by the stepped curve).
(b) ERLE provided by the combined system without constraints.
(c) Optimal filter distortion without psychoacoustic constraints.
(d) Optimal filter distortion with psychoacoustic constraints.

The distortion generated on the near-end speech by the echo cancellation system is strongly dependent on the Signal to Perturbator Ratio. The filter *G* attenuates frequencies where the perturbator has more power than the near-end speech. In the case of a high *SPR*, the attenuation needed is moderate and thus the system yields a moderate distortion. Simulations were then carried out to determine the limits of the contribution of psychoacoustic constraints in our echo cancellation systems. For that, we assumed that the near-end speech was known in order to evaluate properly the spectral masking threshold. Different *SPR*s were tested in the range from -15 dB to 20 dB. Simulations results can be seen figure 5.

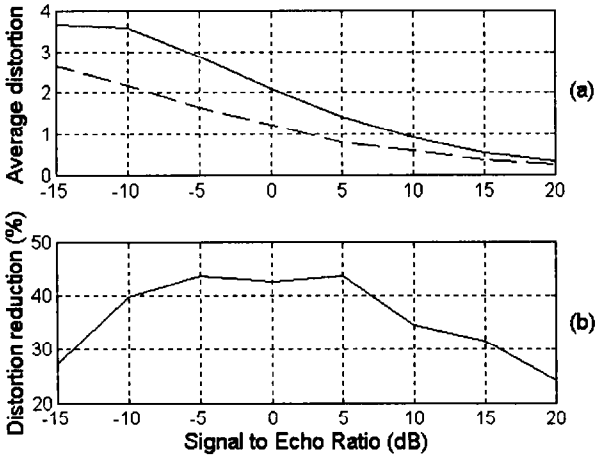

Figure 5. Reduction of the distortion for different *SPR*s.
(a) Average distortion provided by the system without (-) and with (--) psychoacoustic constraints.
(b) Reduction in percentage yielded by the system with constraints.

It appears clearly that the lower the *SPR* is, the higher the distortion is. The use of masking properties is particularly efficient when 0 dB < *SPR* < 10 dB, yielding in that case a reduction of the near-end speech distortion comprised between 30% and 40%, which leads to an average distortion below the audible distortion threshold. In the case of *SPR*>15 dB, the distortion generated is very small, so the improvement obtained with the masking properties is not significant. In the case of very low *SPR*, the improvement obtained with the masking properties is not sufficient enough to make the distortion inaudible at the output of the system, but it is still significant.

5 CONCLUSION

Acoustic echo cancellation can be achieved at least in part by the use of optimal filters which do not require the identification of the acoustic echo path impulse response. The global filtering that we have proposed in the mobile telephony context is a very efficient and interesting technique, since it enables to get rid of the interaction of two different processings classically used in this context. We have shown that the distortion generated on the near-end speech by the optimal filtering can be efficiently reduced by the integration of psychoacoustic criteria in the processing when the perturbator is the acoustic echo. Moreover, the hybrid masking model that we use may lead to a rather low additional cost of computation.

6 REFERENCES

[1] R. MARTIN, J. ALTENHONER, "Coupled adaptive filters for acoustic echo control and noise reduction", ICASSP'95, Detroit, USA, pp. 3043-3046.

[2] Y. EPHRAIM, D. MALAH, "*Speech enhancement using optimal non-linear spectral amplitude estimation*", ICASSP'83, Boston, pp 1118-1121.

[3] C. SINQUIN, "*Global optimisation of noise reduction and acoustic echo cancellation for mobile hands-free radiotelephones*", CNET internal report in French, July 1996.

[4] E. ZWICKER, R. FELDTKELLER, "*Das Ohr als Narichtenempfänger*", Stuttgart, West Germany : Hirzel Verlag, 1967.

[5] J. D. JOHNSTON, "*Transform coding of audio signals using perceptual noise criteria*", IEEE Journal on selected areas in communications, vol. 6, n°2, pp. 314-323, February 1988.

[6] Draft standard ISO 11172-3 MPEG Audio, London, November 1992.

[7] D. TSOUKALAS, M. PARASKEVAS, J. MOURJOPOULOS, "*Speech enhancement using psychoacoustic criteria*", ICASSP'93, Minneapolis, pp. II.359-II.362.