ELSEVIER

Signal Processing 64 (1998) 33–47

SIGNAL PROCESSING

New optimal filtering approaches for hands-free telecommunication terminals

Christophe Beaugeant*, Valérie Turbin, Pascal Scalart, André Gilloire

France Telecom, CNET DIH/CMC, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France

Received 9 October 1997

Abstract

The use of Wiener filtering to simultaneously reduce all the perturbations at the sound pick-up of hands-free terminals is investigated in this paper. This analysis of perturbation reduction is a global approach as all types of perturbations to desired signal are reduced by a unique filter. Filters derived from this approach are applied in two different contexts: hands-free radiotelephony in cars, where two distinct perturbations have to be considered, noise and acoustic echo; and the teleconference context where the global approach was performed on residual echo remaining after a classical echo cancellation. The use of psychoacoustic properties is also considered in order to improve the quality of the transmitted speech signals.

Zusammenfassung

In diesem Aufsatz wird die Verwendung eines Wiener Filters zur gemeinsamen Reduzierung aller Störungen im Mikrofonsignal eines Freisprechendgerätes untersucht. Dieser Ansatz der Geräuschreduzierung wird global genannt, da alle auftretenden Störungen mit einem einzigen Filter verringert werden. Von diesem Ansatz abgeleitete Filter werden für zwei unterschiedliche Einsatzgebiete diskutiert: Autofunktelefone mit Freisprecheinrichtung, bei denen Rauschen und akustische Echos als zwei unterschiedliche Störungen angesehen werden, und Videokonferenzanlagen, bei denen der "globale Ansatz" auf das verbleibende Restecho nach einem konventionellen Echokompensator angewendet wird. Darüber hinaus wird die Ausnutzung von psychoakustischen Eigenschaften zur Verbesserung der Qualität des zu übertragenden Sprachsignals vorgeschlagen.

Résumé

L'utilisation d'un filtre de Wiener pour réduire toutes les perturbations accompagnant une prise de son par des terminaux mains-libres est proposée dans cet article. Cette analyse de la réduction des perturbations est appelée approche globale étant donné que toutes les perturbations sont supposées être réduites par un filtre unique. Nous avons appliqué les filtres issus de cette méthode globale dans deux contextes distincts: d'une part la radio-téléphonie mains-libres dans les voitures, contexte dans lequel deux perturbations distinctes doivent être prises en compte, le bruit et l'écho acoustique d'autre part le contexte de téléconférence pour lequel l'approche globale a été utilisée afin de réduire l'écho résiduel;

* Corresponding author. E-mail: christophe.beaugeant@cnet.francetelecom.fr.

0165-1684/98/$19.00
PII S0165-1684(97)00174-6

persistant après une annulation d'écho classique. L'utilisation de propriétés psychoacoustiques est également proposée dans le but d'augmenter la qualité des signaux de parole traités.

Keywords: Wiener filter; Disturbance reduction; Echo cancellation; Noise reduction; Global approach; Psychoacoustic constraints; Masking model

1. Introduction

The development of hands-free telecommunication terminals induces an increase in technical difficulties to transmit 'natural' and 'undisturbed' signals. The terms 'natural' and 'undisturbed' are difficult to define, depending on the context and on the quality requirements of these terminals. Nevertheless, intelligibility and speaker recognition are at least required. Two main disturbances need to be reduced at the sound pick-up level, namely the acoustic echo and the background noise.

The acoustic echo is due to the feedback of the far end speaker's voice through the loudspeaker-microphone path involved in hands-free communication. This must be reduced first, to prevent howling in the hybrid/acoustic loop within two-wire connected speakerphones (the Larsen effect) and second, to prevent the far-end talker from hearing his echo. Noise reduction is of particular interest in adverse environments, for example hands-free telephones in cars where loud background noise may occur (engine, tires, etc.). Overall, increased intelligibility and decreased far-end speaker fatigue (from these perturbations) are sought.

For many years, the perturbations described above have been considered separately, and many echo cancellation methods [10,11] or noise reduction systems [6,9] have been proposed in the literature. For reducing both perturbations, classical approaches lead to cascading a noise reduction filter and an echo canceller separately. However, as shown in Section 2, these usual solutions of combined systems tend to prove that the interaction between algorithms is high. As a result, the proposed new optimal filtering can be considered as a global disturbance reduction algorithm in which separated degradation is no longer considered. This paper deals with such a global approach to perturbation reduction. The principle is based on the minimization of the mean-squared error and is explained in Section 3. To enhance such filters and to obtain sufficiently high speech quality, the use of perceptual speech properties is introduced in Section 4. Simulations were made using such an approach and the comparison between systems using the global approach and the usual combined systems is presented in Section 5.

2. State of the art: steps to the global approach

In earlier systems that combined acoustic echo cancellation and noise reduction, the problem of the arrangement of the two processing operations had to be taken into account: should noise reduction (NR) occur before acoustic echo cancellation (EC), as first proposed in [17] (structure NR/EC shown in Fig. 1(a)) or vice versa, as proposed in [14,7] (structure EC/NR shown in Fig. 1(b))? An answer to this question can be found by considering the general theoretical problem of the minimization of the mean-squared error (MMSE) between the near-end speech signal and the processed signal at the system output.

In Fig. 1(a) and Fig. 1(b), $y(t)$, $x(t)$ and $\hat{s}(t)$ represent the microphone signal, the loudspeaker signal and the output of the processing systems, respectively. In the following, $s(t)$ and $n(t)$ stand for the near-end speech signal and for the noise signal. In the frequency domain these signals will be noted $Y(f)$, $X(f)$, $\hat{S}(f)$, $S(f)$ and $\hat{N}(f)$, respectively.

Considering now the model of Fig. 2, the estimate of the near-end speech spectrum $\hat{S}(f)$ can be written as the linear combination of the far-end signal $X(f)$ and of the microphone signal $Y(f)$, as follows:

$$\hat{S}(f) = W_X(f)X(f) + W_Y(f)Y(f). \quad (1)$$

Assuming that the noise, near-end speech signal and acoustic echo are mutually uncorrelated, the minimization of $E[(S(f) - \hat{S}(f))^2]$ w.r.t. $W_X(f)$ and

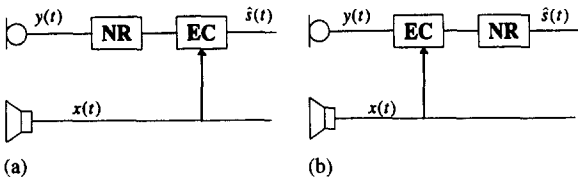

Fig. 1. (a) NR/EC combination. (b) EC/NR combination.

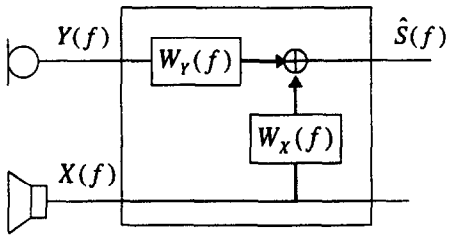

Fig. 2. Theoretical modelling of the estimation problem.

$W_Y(f)$ leads to

$$\hat{S}(f) = W_1(f)[Y(f) - W_2(f)X(f)], \tag{2}$$

with

$$W_1(f) = \frac{E[S^2(f)]}{E[S^2(f)] + E[N^2(f)]}$$

and

$$W_2(f) = \frac{E[Y(f)X(f)]}{E[X^2(f)]},$$

where $E[\cdot]$ is the expectation function.

It can be shown easily that Eq. (2) corresponds to the combination of Fig. 1(b), i.e. an echo canceller (filter $W_2(f)$) followed by a noise reducer (filter $W_1(f)$).

However, the MMSE solution can be written as well as [1]

$$\hat{S}(f) = W_1(f)Y(f) - W_3(f)W_1(f)X(f), \tag{3}$$

with

$$W_3(f) = \frac{E[(W_1(f)Y(f))(W_1(f)X(f))]}{E[(W_1(f)X(f))(W_1(f)X(f))]}.$$

In Eq. (3) the echo cancellation processing is applied to the signals (far-end and microphone signals) pre-processed by $W_1(f)$. This corresponds to the NR/EC structure of Fig. 1(a). As a result, the MMSE criterion is not sufficient to determine which structure among EC/NR and NR/EC leads to the optimal solution.

Many arguments are in favor of the EC/NR order. First of all, echo reduction can be increased in that case: if some residual echo remains, the NR device will consider it as background noise and reduce it [14]. A second argument in favor of the EC/NR structure comes from the linearity constraint of usual echo cancellers. These latter generally use adaptive finite impulse response (FIR) filters. Linearity is therefore required between the reference signal and the output of the unknown system to provide good performance and to converge correctly. However, some nonlinearity may be introduced by the NR filter in the NR/EC structure. As a result the linearity between the far-end signal and the output of NR is no longer ensured. Correct identification is then difficult to obtain by the echo canceller. The NR/EC structure has nevertheless the advantage of reducing the noise level before the echo canceller which may be necessary for noise-sensitive echo cancellation algorithms. In general terms, combined systems optimization mainly depends on the robustness of the algorithms. Our experiments have proved that an echo canceller sensitive to noisy signals should be preceded by an NR filter, conversely, a robust echo canceller in an EC/NR structure provides correct identification of the impulse response of the acoustic path, even if fairly high levels near-end noise is present.

All these remarks indicate that the connection of two independent locally optimal solutions does not lead to an optimal system in practice. In addition, the discussion so far has stressed the interaction between the two filters. Such results lead to considering systems where both filters cooperate. The crucial point is no longer the processing order but how to optimize the combination taking into account the properties of each filter. This is illustrated in Fig. 3. Such a statement can be found in [8,3] where echo cancellation is enhanced by noise reduction, or in [14] where the echo cancellation is only partially done, since noise reduction also reduces the residual echo.

The final step to simultaneously reducing all disturbances is to achieve a single global filter reducing both echo and noise, as proposed in Fig. 4.

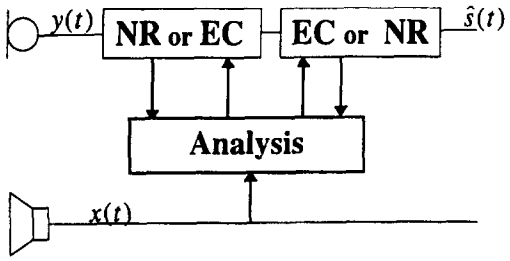

Fig. 3. Combined systems.

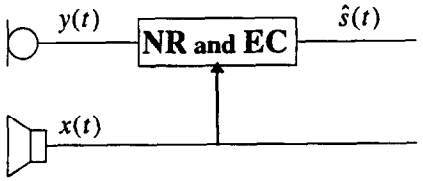

Fig. 4. Global approach.

In order to take into account the interaction between the degrees of noise and echo reduction, a single global filter is applied to the microphone signal $y(t)$, which simultaneously reduces both perturbations. The optimal filter in this approach is no longer simply a linear filtering of the microphone and the loudspeaker signals as in Eq. (1), but it corresponds to the optimal filter $G(f)$ that provides $\hat{S}(f) = G(f)Y(f)$ closest to $S(f)$ according to a given criterion which will be defined in Section 3.1. The purpose of this paper is to deal with such a global approach, to propose algorithms to compute the filters and to enhance the performance of these filters using perceptual properties.

3. Global filter approach

3.1. General principle

The principle of our global 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 $y(t)$ can be written as follows:

$$y(t) = s(t) + p(t), \tag{4}$$

where $s(t)$ and $p(t)$ are uncorrelated signals and stand for the near-end speech and for the perturbation signal, respectively. Depending on the context, $p(t)$ may be composed either of the acoustic echo (in teleconference applications for instance) or of the acoustic echo plus the background noise (in mobile telephony). Considering that $\hat{S}(f) = G(f)Y(f)$ (a filtering of the microphone signal), the optimal filter, G, which gives an estimate of the near-end speech in the sense of the MMSE, can be expressed as [16]

$$G(m,f) = \frac{\mathrm{SPR}(m,f)}{1 + \mathrm{SPR}(m,f)}, \tag{5}$$

where (m,f) indicates that we deal with block processing in the frequency domain of short-term stationary signals, m standing for the current frame. In practice, frequency representations of data are provided through short-term Fourier transform (STFT). In this equation, SPR stands for signal to perturbation ratio and corresponds to the ratio between the short-term power spectral density (psd) of the near-end speech and that of the perturbation.

As the perturbation psd is unknown in advance, the method originally designed for noise reduction proposed in [5] to estimate the SPR was used. It is efficient and has proven to be free of artifacts such as musical tones [2].

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

$$\mathrm{SPR}_{\mathrm{post}}(m,f) = \frac{|Y(m,f)|^2}{\hat{\gamma}_p(m,f)} - 1,$$

$$P(x) = \frac{1}{2}(x + |x|), \tag{6}$$

where $0 < \beta < 1$. $\hat{\gamma}_p(m,f)$ is an estimate of the perturbation psd. In fact, this approach basically consists in applying a Wiener filter to the observation signal knowing an estimation of the disturbance psd. The transposition of this solution originally designed for noise reduction to the echo cancellation problem is the first issue to be discussed. The use of such a global approach means low computation load at the expense of some additional distortions in the transmitted signal.

In this paper, the application of the global approach in two different contexts is presented. In

hands-free radiotelephony applications in cars, where low-complexity algorithms are needed, global filtering was tested directly; in teleconference systems, a high quality of speech signals is needed. In this latter context, a pre-filtering (EC) was therefore added to reduce the echo level before applying the Wiener filter on the residual echo.

3.2. *Application in a mobile hands-free radiotelephony context: direct use of the global approach*

In the mobile radiotelephony context, the near-end speech $s(t)$ may be corrupted both by the acoustic echo $e(t)$ and by the surrounding noise $n(t)$. The global approach is of interest not only because it applies a global filter to the microphone signal but also because it requires no identification of the echo path impulse response to perform acoustic echo reduction (Fig. 5).

The principle consists in considering the acoustic echo and the surrounding noise as a unique disturbance $p(t)$, i.e. $p(t) = e(t) + n(t)$. Filter G obeys Eq. (5).

Assuming that the echo signal and the noise are not correlated, mathematically equivalent expressions of the filter $G(m,f)$ may be written, such as

$$G(m,f) = \frac{1}{1 + \dfrac{1}{\mathrm{SER}(m,f)} + \dfrac{1}{\mathrm{SNR}(m,f)}}, \tag{7}$$

or

$$G(m,f) = \frac{1}{1 + \dfrac{1 + \mathrm{ENR}(m,f)}{\mathrm{SNR}(m,f)}}, \tag{8}$$

where SER, SNR and ENR express the signal to echo ratio, the signal to noise ratio and the echo to noise ratio, respectively. In practice, the choice between the filters derived from Eqs. (5), (7) and (8) is governed by the individual properties of the ratios estimation. Although these expressions are mathematically equivalent, they do not lead to the same filter properties.

In these two new expressions of G, the estimations of the ratios are determined by algorithms similar to Eq. (6). In each filter, two different ratios appear: (SER, SNR) and (ENR, SNR) in Eqs. (7) and (8), respectively. For each of these ratio pairs, Eq. (6) need two coefficients (β_1, β_2): the first one controls the noise reduction (estimation of SNR), the second one the echo reduction (estimation of SER or ENR). β_1 can be considered as the coefficient used in [6] to obtain good performance in noise reduction. Experiments showed effectively the same results as those shown in [6], i.e. $\beta_1 \approx 0.98$ for the filters (7) and (8) to obtain a good trade-off between musical tones and low distortion of the near-end signal. The choice of the value of β_2 was found to be more dependent on the expression of $G(m,f)$ than β_1. Nevertheless, it appears that good behaviors are obtained for $\beta_2 \in [0.9, 1[$ during voice activity of the far-end signal. It is also noted that the value of β_2 influences the noise reduction and the choice of β_1 influences the echo cancellation performance as well.

Separate estimation of the echo and noise psds are necessary to evaluate the different ratios SNR, SER and ENR. The noise psd is estimated during nonvocal activity periods of near-end speech and echo (detected into the box labeled 'NVA' of Fig. 4).

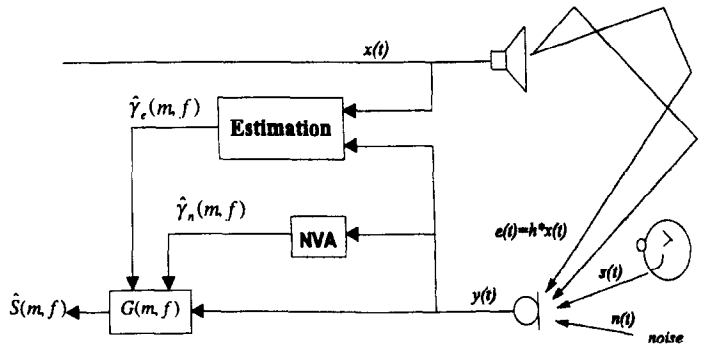

Fig. 5. Echo and noise reduction system.

Since the echo $e(t)$ and the loudspeaker signal $x(t)$ are correlated, an estimation of the echo psd can be obtained:

$$\hat{\gamma}_e(m,f) = \frac{|\gamma_{yx}(m,f)|^2}{\gamma_x(m,f)}, \tag{9}$$

where γ_x and γ_{yx} are the psd of the signal $x(t)$ and the cross-psd between $y(t)$ and $x(t)$ for the block m, respectively. All these spectral quantities are estimated using first-order IIR filters. Our experiments have shown that this estimator is fairly robust in noisy environment, so it is suited for hands-free telephony in cars. With such an estimation, no identification of the echo path is needed. Moreover, no time convergence is required contrary to the usual adaptive echo estimation.

3.3. *Application in a teleconference context: prefiltering of the signal*

In this context where the echo path impulse response is long (the reverberation time is much longer in teleconferencing rooms than in cars), we propose the use of a 'combined system' derived from [14] as shown in Fig. 6. This method can be seen as a special case of the global approach previously described, since pre-processing is added to the filtering $G(m,f)$. In addition, in this section, it is assumed that the perturbation signal reduces to the echo signal $e(t)$, as a matter of fact, much of the time the ambient noise is low and hence is not disturbing in a teleconferencing context.

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

$$e(t) = e_d(t) + e_r(t), \tag{10}$$

with $e_d(t) = \sum_{i=0}^{L-1} h(i)x(t-i)$ and $e_r(t) = \sum_{i=L}^{N-1} h(i)x(t-i)$ and where N is the assumed length of the impulse response (infinite in practice). It may be assumed 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 $e_d(t)$, which depends on the autocorrelation of $x(t)$; nevertheless, the overall behavior of the system is not affected by this bias. Additional echo reduction is obtained using the post-filter G whose task is to attenuate the residual echo $e_r(t)$. The post-filter G obeys Eqs. (5) and (6), in that case $p(t) = e_r(t)$. The perturbation psd can be estimated as proposed in the previous section (Eq. (9)). Another interesting possibility is to take advantage of the vocal activity detection, on which the adaptation process of the filter H_t relies. $e_r(t)$ may be rewritten as follows:

$$e_r(t) = \boldsymbol{h}_{N-L}^{\mathrm{T}} \boldsymbol{x}_{N-L}(t-L) = h_r * x(t-L), \tag{11}$$

where h_r is the impulse response corresponding to vector $\boldsymbol{h}_{N-L}^{\mathrm{T}} = [h(L) \ \ldots \ h(N-1)]$ and $\boldsymbol{x}_{N-L}(t-L)$ is a vector of loudspeaker observations of dimension $N-L$ defined by $\boldsymbol{x}_{N-L}(t-L) = [x(t-L)\, x(t-L-1) \ldots x(t-N+1)]^{\mathrm{T}}$. The residual echo psd estimation can thus be obtained from

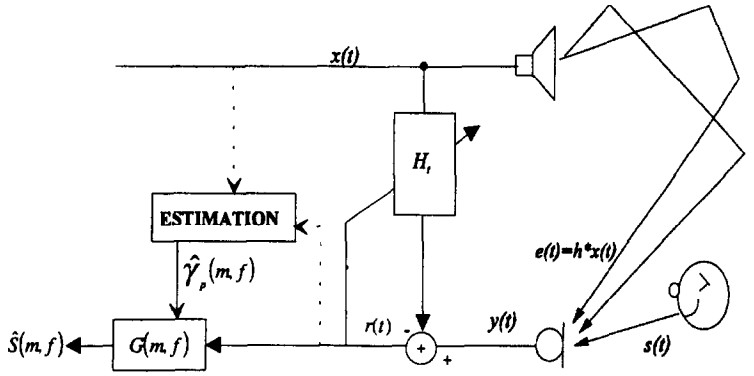

Fig. 6. Combined acoustic echo cancellation system.

the estimation of the transfer function H_r of h_r. During echo only events, this transfer function is estimated as follows:

$$|\hat{H}_r(m,f)|^2 = \frac{|\gamma_r(m,f)|^2}{|\gamma_{x_r}(m,f)|^2}, \quad (12)$$

where γ_r is the estimation of the psd of the signal $r(t)$ and γ_{x_r} the estimation of an appropriately delayed version of the signal $x(t)$. During double-talk events or when no echo is detected, the estimation of H_r is frozen and the last estimate of H_r is used to obtain an estimation of the residual echo psd. In teleconference context, this way of estimating the residual echo is more reliable than Eq. (9). This is particularly true during double-talk events where the decorrelation hypothesis between near-end speech and echo signals often fails on short-term periods.

The choice of the value of parameter β (post-filter G) is guided by a trade-off between the degree of smoothing of SPR and the acceptable level of distortion brought to the near-end speech signal [2]. A high value of β (very close to 1, e.g. 0.98) is needed in the presence of a dominant disturbance signal, whereas a smaller value at the beginning of double-talk periods must be used to avoid clipping effects. In teleconference context, experiments showed that using a parameter β which varies according to SPR is adequate to obtain the desired behavior of the post-filter, especially when talking conditions change. Hence, we suggest the following rule for varying β:

$$\beta = 0.98 - 0.3\frac{\mathrm{SPR}}{1+\mathrm{SPR}}. \quad (13)$$

Relation (13) leads to values of β ranging between [0.68, 0.98]. In case of high SPR values, which indicate the presence of the near-end speech signal, β becomes close to 0.68 allowing the reduction of clipping effects. With low SPR values which correspond to echo only events, β raises to about 0.98 which ensures a good attenuation of the perturbation signal.

As explained in [14], a major advantage of combining an echo canceller with a post-filter is that identification of the complete echo path impulse response is not necessary. For example, simulations have shown that with an impulse response of assumed length $N = 4096$, combining an echo canceller of size $L = 512$ with the post-filter G yields a high amount of echo reduction. The echo is hardly audible at the system output. With the usual echo cancellation approach, an adaptive filter of several thousands coefficients would be necessary to obtain similar results.

3.4. Complexity and delay

Real-time implementation requires a system with reduced complexity and acceptable delay. It is interesting to compare the two different approaches described in the two previous sections to the classical ones in terms of complexity and delay.

In the mobile radiotelephony context, the global approach was compared to a system which cascades a conventional echo canceller in the time domain and a noise reduction Wiener filter in the frequency domain (i.e. NR/EC or EC/NR). The use of a unique Wiener filter to reduce all perturbations with the global approach allows the proposition of a processing really close to the noise reduction Wiener filtering in terms of complexity. The difference is, on one hand, the echo psd estimation and, on the other hand, the computational load due to the adaptive filtering of the echo in the cascaded system. As an example, on one hand, a cascaded structure with a frequency-domain Wiener filtering based on fast Fourier transform (FFT) of size 256 and an NLMS adapted echo canceller of 512 filter taps was considered and, on the other hand, a global filtering implemented with FFT of size 256. The complexity of the cascaded system was found to be 80 times more important than the one of the global filter.

In the teleconference context, the 'combined' system was compared to a classical echo canceller which would provide an equivalent echo reduction at the system output. Implementation of the adaptive echo canceller was assumed to be following the NLMS algorithm in both cases. As already mentioned, a frequency-domain implementation was used for the post-filter. Experiments showed that the 'combined' system composed of a 512 filter taps echo canceller and a frequency-domain post-filtering based on FFT of 1024 taps yielded an echo reduction similar to the one obtained with an NLMS

adaptive filter of size $L = 2048$. The complexity of the 'combined' system was found to be 70 times smaller.

Frequency-domain implementation introduces a delay which depends on the FFT size. As a result, compared to cascaded systems that use a frequency analysis for the noise reduction, no additional delay is introduced by the algorithm proposed in Section 3.2 for radiotelephony applications. In the teleconference context, the method proposed in Section 3.3 introduces a delay compared to a single time-domain adaptive echo canceller, which is due to the frequency-domain post-filtering.

4. Use of psychoacoustic criteria to improve performance

The global approach described in the previous section induces modifications of the microphone signal spectrum, which are dependent on the spectrum of the perturbation involved. Frequency components of the near-end speech may thus be affected by the processing. This may occur especially when the perturbation is an echo (i.e. a speech signal) whose spectrum and the near-end speech spectrum often overlap. In this case, the global filter can generate an audible distortion of the near-end speech. It is therefore necessary to reduce that distortion to improve the speech transmission quality. A possible solution is to take advantage of some properties of the human auditory system. It is well known that when two sounds occur simultaneously, it may happen that one of them is made inaudible by the other: this effect is commonly known as the masking phenomenon [18]. This means that when near-end speech masks the perturbation, there is no need to filter it. Limiting the filtering to frequencies at which the perturbation is not masked reduces the distortion while maintaining the same perceptual amount of disturbance reduction. In this section the main properties of the masking model we used are first recalled. Integration of this model in the optimal filter is then explained.

4.1. Choice of the masking model

The masking model provides means to compute a masking threshold, which is then used to control the filter attenuation. This model only takes simultaneous masking into account, i.e. masking which occurs in the frequency domain. Temporal masking (backward and forward masking) is not considered, as frequency masking is likely to provide the highest distortion reduction; moreover it is important not to add a great computational load to the enhancement process in the applications considered here. The incorporation of masking properties into our systems is based on a human hearing model which yields appropriate spectral masking thresholds. A listener tolerates the presence of disturbances (echo, noise) as long as they do not exceed the masking threshold. The Johnston's masking model [13] and the ISO MPEG Psychoacoustic Model II [12] were considered. The different steps necessary to determine the masking threshold are illustrated in Fig. 7.

1. *Critical band analysis*: The energies of frequency components within each critical band are added up. This addition is based on the energy perception properties of the human ear.
2. *Convolution with the spreading function*: The spreading function is used to estimate the effect of masking across critical bands. The convolution is performed on a Bark scale.
3. *Subtraction of the threshold offset*: The offset is a function of the noiselike or tonelike nature of the masking signal.

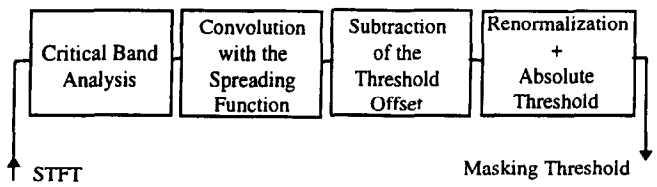

Fig. 7. Calculation of the masking threshold.

4. Renormalization and comparison with the absolute hearing threshold are then used.

The method used to compute the threshold offset mainly distinguishes the Johnston's model from the ISO one. A general expression of the threshold offset is

$$O(b) = \alpha(b)\mathrm{TMN}(b) + (1 - \alpha(b))\mathrm{NMT}(b), \quad (14)$$

where b stands for the Bark frequency, TMN(b) is the value to be used in the case of a tone masking a noise, and NMT(b) the value to be used in the case of a noise masking a tone. The coefficient of tonality $\alpha(b)$, which expresses the noise-like or tone-like nature of the signal, is used to weight the two corrective quantities TMN and NMT. In Johnston's model the computation method leads to an identical coefficient of tonality over all critical bands. This is not the case with the ISO model which requires the computation of a coefficient of tonality for each critical band. The cost of computation is therefore much higher with this latter model. A constant value of 5 dB is used for NMT(b) in both models. TMN depends on the Bark frequency. Higher values are used in the ISO model, especially in the low frequencies. This means that the threshold offset is higher in the ISO model as soon as the signal is not entirely noise-like (i.e. $\alpha(b) \neq 0$). This results in a higher masking threshold for Johnston's model than for the ISO one. A direct consequence is that a perturbation can be considered either masked or unmasked by the near-end speech, depending on the masking model.

The model incorporated in our systems is a 'hybrid' model. The coefficient of tonality is evaluated following the method proposed in the Johnston's model, not only because it is important not to add a heavy computational load to our systems but also because only a rough estimate of the masking signal spectra is available. To compute the threshold offset, the TMN values indicated in the ISO model were considered instead of Johnston's ones. The resulting model is called 'hybrid' since a 'hybrid' method is used to compute the threshold offset which can be written as

$$O_{\mathrm{hybrid}}(b) = \alpha_{\mathrm{JOHNSTON}}\mathrm{TMN}_{\mathrm{ISO}}(b) + (1 - \alpha_{\mathrm{JOHNSTON}})\mathrm{NMT}_{\mathrm{ISO}}(b). \quad (15)$$

The 'hybrid' model provides a masking threshold lower than that obtained with the Johnston's model, as it can be seen in Fig. 8. This decreases the risk of

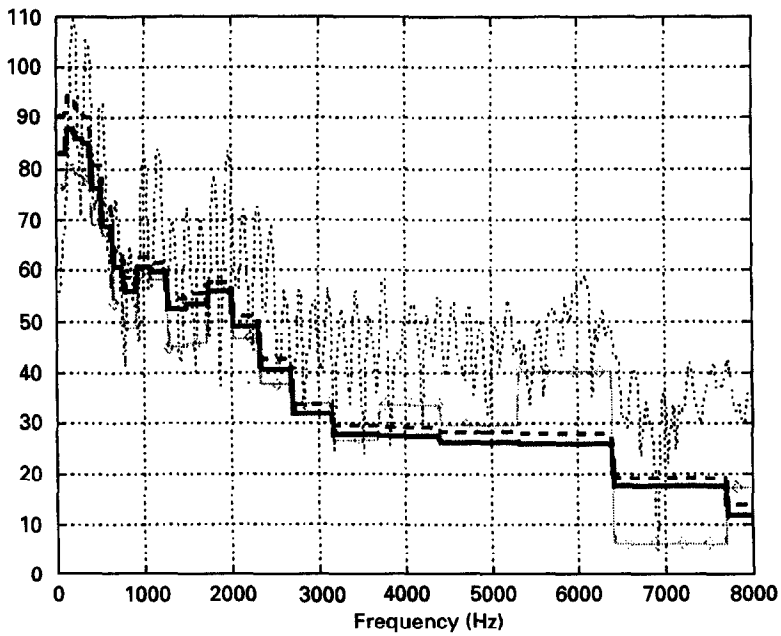

Fig. 8. Example of masking thresholds. ···: Masking signal psd; - - -: Johnston's model; ——— : 'Hybrid' model; – ◇ –: ISO model.

making wrong decisions, which would indicate that the near-end speech masks the perturbation when it actually does not. In comparison with Johnston's model, the 'hybrid' model guarantees a better perceptual disturbance reduction while preserving a reasonable computational cost for the global system.

4.2. *Integration of psychoacoustic constraints into the optimal filtering*

The computation of the spectral masking threshold $T(m,f)$ requires a rough estimate of the near-end speech spectrum. Among several possible solutions to finding an estimate of this signal, a simple one consists in using the output of the optimal filter. It has proven to work fairly well in our experiments. The threshold $T(m,f)$ allows to determine frequencies at 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

$$\hat{S}(m,f) = \begin{cases} U(m,f) & \text{if } \hat{\gamma}_p(m,f) \leqslant T(m,f), \\ G(m,f)U(m,f) & \text{if } \hat{\gamma}_p(m,f) > T(m,f), \end{cases} \tag{16}$$

where $U(m,f)$ stands for the STFT of the signal to be filtered, i.e. $U(m,f) = Y(m,f)$ in the mobile radiotelephony context and $U(m,f) = R(m,f)$ in the teleconference context. This method was previously proposed for noise reduction techniques [15].

5. Simulations and results

Simulations were made both in a teleconferencing application and in hands-free radiotelephony in cars. The experimental conditions of the different simulations were carefully controlled in order to compare the performance of the filters issued from the global approach to those of combined systems. The first observations made during informal listening tests show clearly that the reduction of echo by Wiener filtering induces more distortions in the near-end speech than usual echo cancellers. Therefore, it is important to determine how annoying this distortion is and how the use of perceptual properties may reduce it. Depending on the context and on the quality requirements of the application, the distortion introduced can be less critical than the residual perturbations or other drawbacks due to transmission. This section tries to deal with the difficult problem of finding a compromise between the distortion introduced and the decrease of perturbations.

5.1. *Hands-free telephony in cars: the importance of perturbations*

For the experiments conducted in hands-free telephony, the microphone signal is composed of echo, noise and near-end speech. Each of these were recorded separately in a real environment. As an example, Fig. 9 represents a significant microphone signal at a sampling frequency of 8 kHz. The stepped curve under the time-domain representation indicates the vocal activity at the loudspeaker input, hence the echo voice activity. This example deals with signals in a very noisy environment. The segmental signal-to-noise ratio value is 2.5 dB during the near-end speech activity period. The average value of the segmental signal-to-echo ratio calculated during the double-talk periods (near-end speech and echo at the same time) is 4 dB. These values correspond to a typical application of hands-free phone in a car at 100 km/h.

In such conditions, the algorithms provided by the global approach were compared with combined solutions proposed in [8]. More precisely, the curves of the Fig. 10 were obtained using an EC/NR structure with an echo cancellation based on the APA (Affine Projection Algorithm); the noise reduction was obtained using a Wiener filter limited to 10 dB noise reduction (to limit musical tones). Figs. 11 and 12 represent the same measurements obtained using the global filter given by Eq. (5) and with the global filter enhanced by the perceptual properties explained in Section 4.2, respectively. It should be noted that the maximum value of the disturbance reduction is limited to 10 dB in order to obtain a noise reduction equal to that given by the combined system.

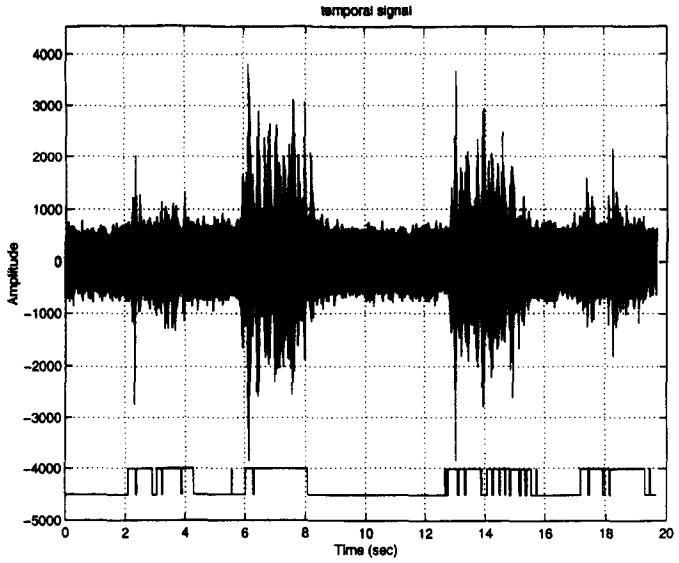

Fig. 9. A signal typical of a mobile application.

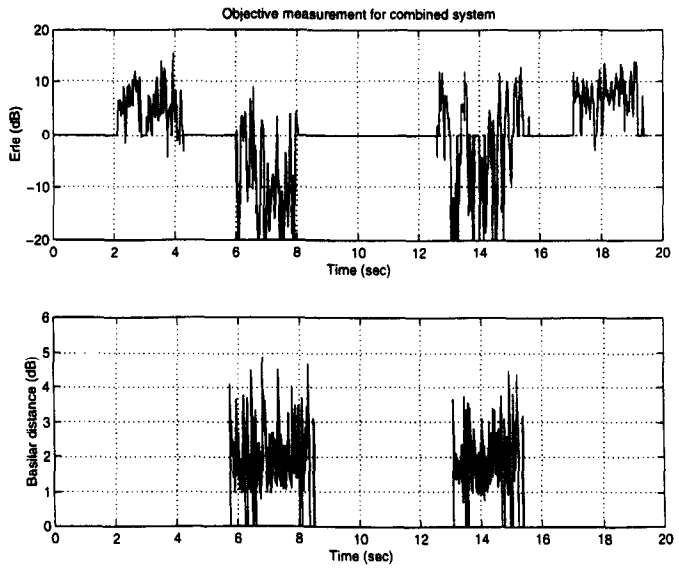

Fig. 10. Results with a combined system.

In order to compare the performance of the algorithms, echo attenuation is quantified using the echo return loss enhancement parameter (ERLE) which is computed on blocks of 256 samples with no overlapping. The speech distortion measurement is provided by the basilar distance from the perceptual objective measure system [4].

Considering Figs. 10 and 11, we can see that the global filter provides a fairly constant echo attenuation of 10 dB (the limitation value of the filter),

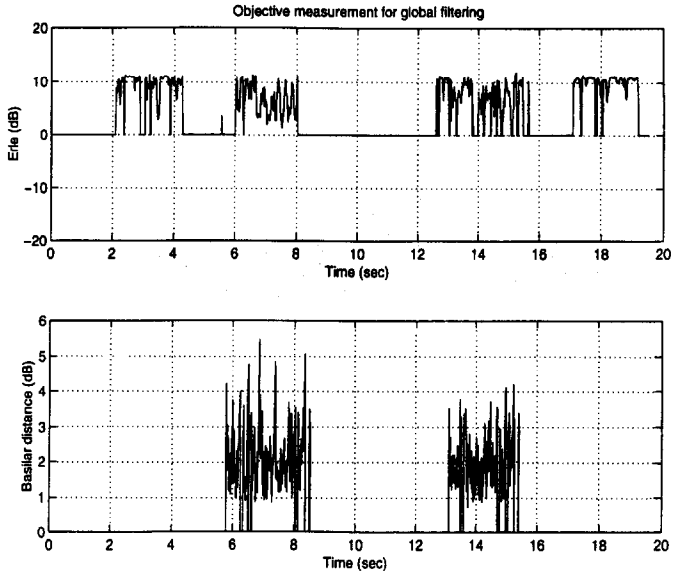

Fig. 11. Results with a global filter.

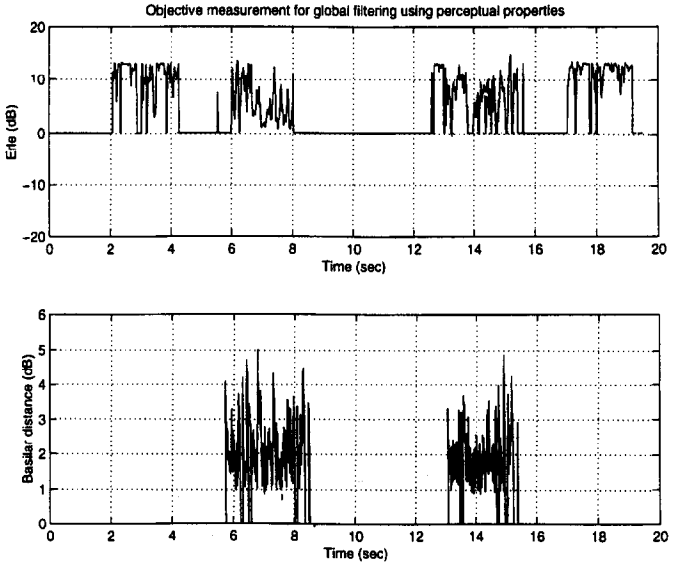

Fig. 12. Results with a global filter using perceptual properties.

whereas the combined system provides variable echo attenuation, mainly during double talk periods when adaptation is frozen. The distortion in the near-end speech signal is more audible for the global filter. Nevertheless, the distortion difference is not so significant and the informal listening tests confirm this observation.

As a result and as shown in Fig. 12, although taking the perceptual properties of speech into account decreases distortion of the near-end speech, the enhancement proposed in Section 4.2 is not so efficient in the loud noise context. It should be noted however that the lower the noise, the more noticeable the distortion provided by the global

filters becomes. Exploiting perceptual properties is therefore more productive in a low-noise environment, which is the case with low car speeds. Nevertheless, in such conditions, experiments and listening tests lead to the conclusion that the combined system is more efficient than the global filtering.

The filters given by Eqs. (7) and (8) were also compared to the combined system in both listening tests and objective measurements (ERLE and basilar distance). The objective measurement leads to the same conclusions as those made from the inspection of Figs. 10–12. Listening tests revealed that the overall quality of the processed signal is equivalent whatever the filter used. The only difference that can be noticed is in the coloration of the near-end speech that arises from different algorithms used for the computation of the filter.

5.2. Teleconference application

Simulations were carried out with impulse responses measured in a teleconferencing room, which were used with several values of the echo loss through the echo path. As an illustrative example, results in a teleconferencing context are presented in Fig. 13. They correspond to a signal-to-echo ratio (evaluated during double-talk periods) at the microphone input of 3 dB. An echo canceller of 512 coefficients was used, which corresponds to 32 ms at the sampling frequency of 16 kHz (the measured impulse response was 256 ms long). In this case, the echo canceller provides about 12 dB echo attenuation. With these conditions, the remaining echo components at the output of the post-filter were hardly audible. Fig. 13(b) shows very high ERLE values during echo only periods and fairly high

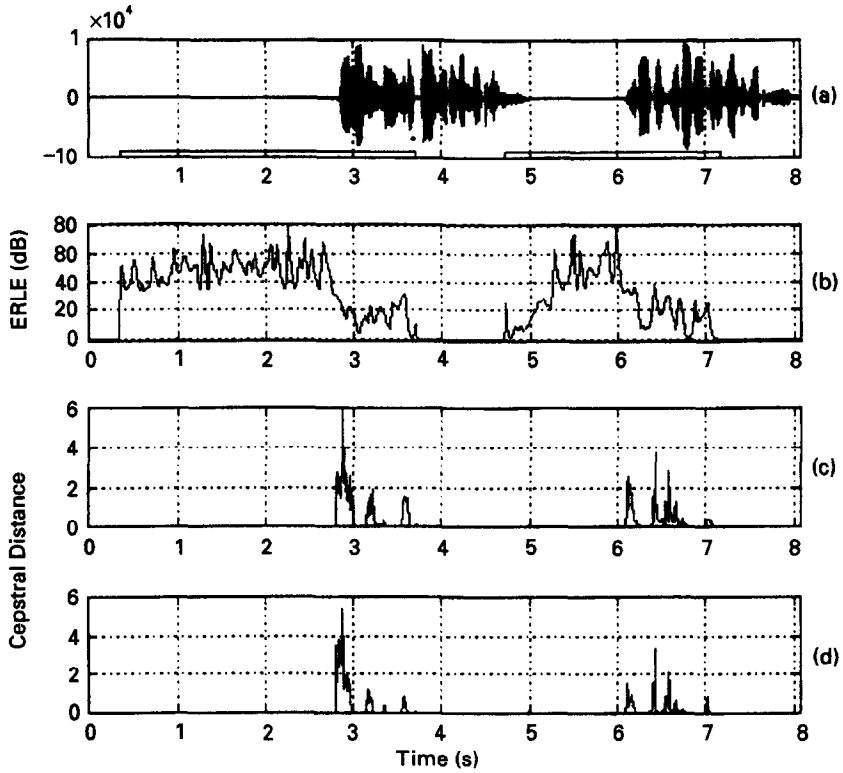

Fig. 13. Results in a teleconference context. (a) Near-end speech with echo events indicated by the stepped curve. (b) ERLE provided by the system (EC + post-filter). (c) Near-end speech distortion yielded by the system without psychoacoustic constraints. (d) Near-end speech distortion yielded by the system with psychoacoustic constraints.

(about 20 dB) ERLE values during double talk. The distortion generated on the near-end speech was evaluated by the cepstral distance between the original near-end speech and the same signal filtered by the system. Results for the system without and with psychoacoustic criteria are shown in Fig. 13(c) and Fig. 13(d), respectively. 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.

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 system. For that, we assumed that the near-end speech was known in order to evaluate properly the spectral masking threshold. Different SPRs were tested in the range from -15 to 20 dB. Simulations results are shown in Fig. 14. It is clear that the lower the SPR is, the higher the distortion is. The use of masking properties is particularly efficient when $0\,\text{dB} < \text{SPR} < 10\,\text{dB}$, 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 system output, but it is still significant.

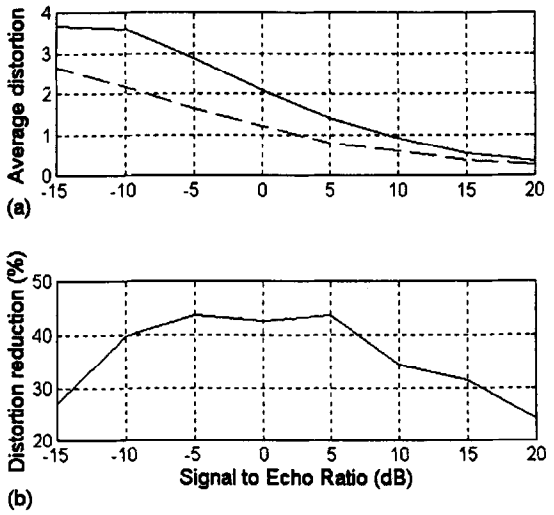

Fig. 14. Reduction of the distortion for different SPRs. (a) Average distortion provided by the system without (—) and with (— —) psychoacoustic constraints. (b) Reduction in percentage yielded by the system with constraints.

6. Conclusions

The use of a single global filter to reduce all the perturbations leads to efficient systems in hands-free radiotelephony in car and in teleconference applications as well. To reduce both noise and acoustic echo in car mobile environment, the comparison with classical combined systems shows that the global approach leads to acceptable results. In teleconference context, a Wiener filter was introduced as an additional echo attenuation, since a FIR acoustic echo canceller alone does not deliver sufficient echo attenuation in all circumstances. The main drawback of these kinds of Wiener global filters is the distortion introduced on the near-end speech. The use of psychoacoustic constraints was added to decrease this distortion. The experimental results lead to the conclusion that in noisy environment like sound pick-up in moving car, the introduction of these properties does not enhance the performance in a significant way. However, when the noise power is sufficiently low, as in teleconference contexts, the perceptual point of view is well suited to ameliorate the distortion and to provide high-quality near-end speech signals.

References

[1] A. Benamar, Etude et implantation de la fonction de contrôle de l'écho acoustique pour la radiotéléphonie

mains-libres, Thèse de l'Université Paris XI d'Orsay, No. d'ordre 4504, October 1996, pp. 117–119.
[2] O. Cappé, Elimination of the musical noise phenomenon with the Ephraim and Malah suppressor, IEEE Trans. Speech Audio Process. 2 (2) (April 1994) 345–349.
[3] F. Capman, J. Boudy, P. Lockwood, Acoustic echo cancellation and noise reduction in the frequency domain: a global optimisation, Proc. of EUSIPCO'96, 1996, pp. 29–32.
[4] C. Colomes, M. Lever, Y.F. Dehery, A perceptual objective measurement system (POM) for the quality Assessment of perceptual codecs, AES, 96th Convention, February 1994.
[5] Y. Ephraïm, D. Malah, Speech enhancement using optimal non-linear spectral amplitude estimation, ICASSP'83, Boston, 1983, pp. 1118–1121.
[6] Y. Ephraïm, Statistical-model-based speech enhancement systems, Proc. IEEE 80 (10) (October 1992) 1526–1555.
[7] G. Faucon, R. Le Bouquin, Joint system for acoustic echo cancellation and noise reduction, Proc. Eurospeech'95, Madrid, September 1995, pp. 1525–1528.
[8] Y. Guélou, A. Benamar, P. Scalart, Analysis of two structures for combined acoustic echo cancellation and noise reduction, Proc. of ICASSP'96, 1996, pp. 637–640.
[9] J.H.L. Hansen, Speech enhancement and quality assessment with applications to robust recognition and coding, Tutorial Internat. Conf. Acoust. Speech Signal Process., Detroit, MI, USA, 1995.
[10] E. Hänsler, The hands-free telephone problem – An annotated bibliography, Signal Processing 27 (1992) 259–271.
[11] E. Hänsler, The hands-free telephone problem: an annoted bibliography update, Ann. des Télécommun. 49 (7–8) (1994) 360–367.
[12] ISO, Draft standard ISO 11172-3 MPEG Audio, London, November 1992.
[13] J.D. Johnston, Transform coding of audio signals using perceptual noise criteria, IEEE J. Selected Areas Commun. 6 (2) (February 1988) 314–323.
[14] R. Martin, J. Altenhoner, Coupled adaptive filters for acoustic echo control and noise reduction, ICASSP'95, Detroit, USA, 1995, pp. 3043–3046.
[15] D. Tsoukalas, M. Paraskevas, J. Mourjopoulos, Speech enhancement using psychoacoustic criteria, ICASSP'93, Minneapolis, pp. II.359–II.362.
[16] S.V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction, Chapter 9, Wiley Teubner Communications, Queen's University of Belfast, UK, 1996.
[17] H. Yakasuma, Acoustic echo canceller with sub-band noise cancelling, Electr. Lett. 28 (15) (July 1992) 1403–1404.
[18] E. Zwicker, R. Feldkeller, Das Ohr als Narichtenempfänger, Hirzel Verlag, Stuttgart, West Germany, 1967.