ACOUSTIC ECHO AND NOISE REDUCTION: A NOVEL APPROACH

Beghdad AYAD, Régine LE BOUQUIN-JEANNÈS

Laboratoire de Traitement du Signal et de l'Image - Université de Rennes 1
Bât. 22 - Campus de Beaulieu - 35042 RENNES CEDEX - FRANCE
Regine.Le-Bouquin-Jeannes@univ-rennes1.fr

ABSTRACT

Acoustic echo and noise cancelling is fundamental in any speech transmission system. In the solutions addressed to this problem, the acoustic echo cancellation is carried out by identification of the transfer function of the acoustic channel. In this paper, another approach is proposed where echo cancellation is realized by filtering the microphone observation. Within this approach, three systems are developed. For noise reduction, an updating of the noise characteristics in the presence of speech is studied. Measures of echo return loss enhancement, noise reduction and speech distortion are presented. It happens that the new approach performs better than the basic one.

1. INTRODUCTION

In any speech transmission system, the presence of acoustic echo and noise is undesirable and so the attenuation of these two disturbances is fundamental, particularly in hands-free telecommunications. A few solutions [1,2,3] have already been addressed to this topic. In this paper, the minimum mean square error (MMSE) criterion applied in the frequency domain is used to advance solutions to this problem. In the following, we first develop acceptable solutions for noise reduction (NR) and acoustic echo cancellation (AEC) and then we discuss how to combine the two operations. Four different combined systems are proposed and compared.

2. OPTIMAL FILTERING

The estimated output is considered as a linear combination of the inputs. In the following, $U(f)$ denotes the Fourier transform of a signal $u(t)$. Let $\underline{Y}(f)$ be the input observation vector and $\underline{H}(f)$ the filter gain vector, the estimated signal is given by:

$$\hat{S}(f)=\underline{H}^T(f)\cdot\underline{Y}(f) \tag{1}$$

where the linear optimal filter minimising the error $\mathbf{E}\left\{\left|S(f)-\hat{S}(f)\right|^2\right\}$ in the frequency domain is:

$$\underline{H}(f)=\left(\Gamma_{\underline{Y}\underline{Y}}^{-1}(f)\cdot\Gamma_{S\underline{Y}}(f)\right)^* \tag{2}$$

which corresponds to the non-causal Wiener filter; $\Gamma_{\underline{Y}\underline{Y}}(f)$ is the power spectral density (psd) matrix of the vector $\underline{Y}(f)$ and $\Gamma_{S\underline{Y}}(f)$ is the cross-psd between $S(f)$ and $\underline{Y}(f)$; the asterix denotes the conjugate.

3. NOISE REDUCTION

We consider the microphone observation x composed of the useful signal s added to the disturbing noise n. Using (1) and (2), the estimated signal in the sense of the MMSE is given by:

$$\hat{S}(f)=\frac{\gamma_{ss}(f)}{\gamma_{xx}(f)}\cdot X(f) \tag{3}$$

where $\gamma_{uu}(f)$ represents the psd of the signal u.
Since signals are non stationary, we compute the psd on each block k and we define the *a priori* Signal to Noise Ratio, $SNR_{pri}(f,k)$, by:

$$SNR_{pri}(f,k)=\frac{\gamma_{ss}(f,k)}{\gamma_{nn}(f,k)}. \tag{4}$$

The estimated signal in (3) can be expressed by:

$$\hat{S}(f,k)=\frac{SNR_{pri}(f,k)}{SNR_{pri}(f,k)+1}\cdot X(f,k) \tag{5}$$

which defines the practical implementation of the noise reduction filter on windowed signals including overlap-and-add. $U(f,k)$ is the Short Time Fourier Transform (STFT) of the signal u.
Now, the problem lies with the estimation of $SNR_{pri}(f,k)$. In [4,5], some estimation methods are compared regarding bias, variance and the purchase of rapid speech variations. All methods use a noise psd computed on sequences where noise is alone, which supposes the introduction of a voice activity detector. If we are able to estimate the noise psd on each block k, we can get rid of this voice activity detector. To this end, we exploit the orthogonality principle when minimising the squared error between $S(f,k)$ and $\hat{S}(f,k)$. It stipulates that:

$$\mathbf{E}\left\{X(f,k)\cdot\left(S(f,k)-\hat{S}(f,k)\right)^*\right\}=0. \tag{6}$$

Having only $\hat{S}(f,k)$ and $X(f,k)$, we can estimate a noise psd on each block k by computing:

$$\gamma_{nn}(f,k)=\mathbf{E}\left\{X(f,k)\cdot\left(X(f,k)-\hat{S}(f,k)\right)^*\right\}. \tag{7}$$

Hereafter, we detail the noise reduction algorithm:

a) first of all, we compute the noise psd on the first ten blocks of 256 samples (corresponding to silent periods):

$$\gamma_{nn}(f,k)=\rho\cdot\gamma_{nn}(f,k-1)+(1-\rho)\cdot|N(f,k)|^2 \quad (8)$$

b) we estimate the psd of the microphone signal:

$$\gamma_{xx}(f,k)=\alpha\cdot\gamma_{xx}(f,k-1)+(1-\alpha)\cdot|X(f,k)|^2 \quad (9)$$

c) the ratio $SNR_{pri}(f,k)$ is computed as follows [5]:

$$SNR_{pri}(f,k)=\beta\cdot\frac{|\hat{S}(f,k-1)|^2}{\gamma_{nn}(f,k)}+(1-\beta)\cdot\mathrm{Max}\left[\frac{\gamma_{xx}(f,k)}{\gamma_{nn}(f,k)}-1;0\right] \quad (10)$$

where $|\hat{S}(f,k-1)|$ is the amplitude of the signal estimate on the block $k-1$;

d) a signal estimate is obtained using eq. (5);

e) we compute a new estimate of the noise psd using eq. (7)

$$\gamma_{nn}(f,k)=\lambda\gamma_{nn}(f,k-1)+\mathrm{Re}\left[(1-\lambda)X(f,k)\left(X(f,k)-\hat{S}(f,k)\right)^*\right] \quad (11)$$

where Re[.] denotes the real part;

f) we return to *b)*.

ρ, α, β and λ are weighting factors between 0 and 1 ($\rho=0.9$, $\alpha=0.7$, $\beta=\lambda=0.98$).

4. ACOUSTIC ECHO CANCELLATION

In this section, the signal received by the microphone is composed of a useful signal s and a noise n constituting the disturbance d and the echo e. Two filters are derived according as we consider one or two "observations".

4.1 Optimal filter considering two observations

The input observation vector $\underline{Y}(f)$ is composed of the microphone signal $X(f)$ and the signal $Z(f)$ emitted by the loudspeaker, $\underline{Y}(f)=[X(f)\ \ Z(f)]^T$. Equations (1) and (2) lead to:

$$\hat{D}(f)=X(f)-\frac{\gamma_{xz}(f)}{\gamma_{zz}(f)}\cdot Z(f) \quad (12)$$

where $\gamma_{uv}(f)$ is the cross-psd between signals u and v. Classically, the ratio $\gamma_{xz}(f)/\gamma_{zz}(f)$, which identifies the acoustic channel transfer function, is computed in an adaptive manner [6]. Ideally, the output $\hat{d}$ is equal to $s+n$.

4.2 Optimal filter considering one observation

In this case, we consider that the input vector is only composed of the microphone signal. Equations (1) and (2) reduce to:

$$\hat{D}(f)=\frac{\gamma_{dd}(f)}{\gamma_{xx}(f)}\cdot X(f)=\frac{\gamma_{dd}(f)}{\gamma_{dd}(f)+\gamma_{ee}(f)}\cdot X(f). \quad (13)$$

Defining the *a priori* Disturbance to Echo Ratio by:

$$DER_{pri}(f,k)=\frac{\gamma_{dd}(f,k)}{\gamma_{ee}(f,k)}, \quad (14)$$

equation (13) may be rewritten:

$$\hat{D}(f,k)=\frac{DER_{pri}(f,k)}{DER_{pri}(f,k)+1}\cdot X(f,k). \quad (15)$$

5. COMBINED SYSTEMS

In a full duplex communication, the ambient noise is omnipresent. So, we distinguish four kinds of sequences: *i)* noise is alone ($x=n$), *ii)* the near-end speech signal is present ($x=s+n$), *iii)* an echo is present due to the far-end speaker ($x=e+n$), *iv)* near-end speech and far-end speech are simultaneously present ($x=s+e+n$); this sequence represents 20% of a communication (double talk mode). The first two sequences only require a noise reduction system. In the two others, we have to reduce both echo and noise. A strategy must be investigated to get a slightly distorted near-end speech signal. Our approach is based on the MMSE criterion to obtain the best estimate of the signal s. As previously, two input vectors are considered.

5.1 Optimal system with two observations

Considering the input observation vector $\underline{Y}(f)=[X(f)\ \ Z(f)]^T$, the linear optimal estimate $\hat{S}(f)$ is given by:

$$\hat{S}(f)=\left(X(f)-\frac{\gamma_{xz}(f)}{\gamma_{zz}(f)}\cdot Z(f)\right)\cdot\frac{\gamma_{ss}(f)}{\gamma_{ss}(f)+\gamma_{nn}(f)}. \quad (16)$$

In practice, we compute $\hat{S}(f,k)$ in the following manner:

$$\hat{S}(f,k)=\left(X(f,k)-\frac{\gamma_{xz}(f,k)}{\gamma_{zz}(f,k)}\cdot Z(f,k)\right)\frac{SNR_{pri}(f,k)}{SNR_{pri}(f,k)+1} \quad (17)$$

This equation shows that the optimal filtering consists of an echo canceller (based on the identification of the transfer function) followed by a noise reduction filter (Wiener filtering). This structure is named structure S1.

5.2 Optimal system with one observation

Now, if we consider only the microphone observation, the linear optimal estimate is given by:

$$\hat{S}(f)=\frac{\gamma_{ss}(f)}{\gamma_{xx}(f)}\cdot X(f). \quad (18)$$

The sum echo+noise is seen as a global perturbation. Let p be equal to $e+n$. We define the *a priori* Signal to Perturbation Ratio $SPR_{pri}(f,k)$ on each block k by:

$$SPR_{pri}(f,k)=\frac{\gamma_{ss}(f,k)}{\gamma_{pp}(f,k)}=\frac{\gamma_{ss}(f,k)}{\gamma_{nn}(f,k)+\gamma_{ee}(f,k)}. \quad (19)$$

1) A first realisation consists in directly implementing eq. (19) through the use of $SPR_{pri}(f,k)$:

$$\hat{S}(f,k)=\frac{SPR_{pri}(f,k)}{SPR_{pri}(f,k)+1}\cdot X(f,k). \quad (20)$$

In this global system (structure S2), we do not distinguish the two operations echo cancellation and noise reduction.

2) To separate both operations, eq. (20) can be written:

$$\hat{S}(f,k)=\underbrace{\frac{DER_{pri}(f,k)}{DER_{pri}(f,k)+1}}_{AEC\ filter}\cdot\underbrace{\frac{SNR_{pri}(f,k)}{SNR_{pri}(f,k)+1}}_{NR\ filter}\cdot X(f,k) \quad (21)$$

This filtering (structure S3, Figure 1) shows that each operation is realized by a Wiener filtering.

$DER_{pri}(f,k)$ can also be expressed by using the magnitude squared coherence between x and z:

$$DER_{pri}(f,k)=\frac{1}{MSC_{xz}(f,k)}-1 \qquad (22)$$

where $$MSC_{xz}(f,k)=\frac{|\gamma_{xz}(f,k)|^2}{\gamma_{xx}(f,k)\cdot\gamma_{zz}(f,k)}. \qquad (23)$$

So, the echo cancellation can be carried out by using $MSC_{xz}(f,k)$, which gives another realization (structure S4, Figure 1):

$$\hat{S}(f,k)=(1-MSC_{xz}(f,k))\cdot\frac{SNR_{pri}(f,k)}{SNR_{pri}(f,k)+1}\cdot X(f,k). \qquad (24)$$

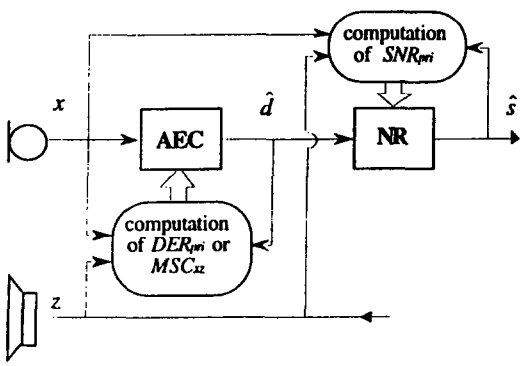


Figure 1. Structures S3 and S4

5.3 Implementation

The adaptive acoustic echo canceller used in eq. (17) is the GMDF algorithm [7]. The estimation of $SNR_{pri}(f,k)$ in structures S1, S3 and S4 can be deduced from eq. (10):

$$SNR_{pri}(f,k)=\beta_s\cdot\frac{|\hat{S}(f,k-1)|^2}{\gamma_{nn}(f,k)}+(1-\beta_s)\cdot \text{Max}\left[\frac{\gamma_{xx}(f,k)-\gamma_{ee}(f,k)}{\gamma_{nn}(f,k)}-1;0\right] \qquad (25)$$

where $|\hat{S}(f,k)|^2$ is the periodogram of the final system output, $\gamma_{nn}(f,k)$ is computed as in section 3 and $\gamma_{ee}(f,k)$ is given by:

$$\gamma_{ee}(f,k)=\frac{|\gamma_{xz}(f,k)|^2}{\gamma_{zz}(f,k)}. \qquad (26)$$

The second term of $SNR_{pri}(f,k)$ is computed by carrying out a spectral subtraction to remove echo from the observation instead of using the AEC output in order to get a less distorted signal [8].
In structure S2, $SPR_{pri}(f,k)$ is estimated in the same way as $SNR_{pri}(f,k)$ in eq. (10):

$$SPR_{pri}(f,k)=\beta_p\cdot\frac{|\hat{S}(f,k-1)|^2}{\gamma_{pp}(f,k)}+(1-\beta_p)\cdot\text{Max}\left[\frac{\gamma_{xx}(f,k)}{\gamma_{pp}(f,k)}-1;0\right] \qquad (27)$$

where $|\hat{S}(f,k)|^2$ is the periodogram of the global system output and $\gamma_{pp}(f,k)$ is the sum of $\gamma_{nn}(f,k)$ and $\gamma_{ee}(f,k)$. For $DER_{pri}(f,k)$ (structure S3), we use a similar form:

$$DER_{pri}(f,k)=\beta_e\cdot\frac{|\hat{D}(f,k-1)|^2}{\gamma_{ee}(f,k)}+(1-\beta_e)\cdot\text{Max}\left[\frac{\gamma_{xx}(f,k)}{\gamma_{ee}(f,k)}-1;0\right] \qquad (28)$$

where $|\hat{D}(f,k)|^2$ is the periodogram of the estimated AEC output using a Wiener filter.
All psd (except $\gamma_{nn}(f,k)$) are computed using a recursive formula:

$$\gamma_{uv}(f,k)=\alpha\cdot\gamma_{uv}(f,k-1)+(1-\alpha)\cdot U(f,k)\cdot V^*(f,k) \qquad (29)$$

where α is a forgetting factor between 0 and 1.

In practice, each structure is performed on blocks of 256 samples, with a 75% overlapping. We choose the following parameters: for the GMDF algorithm, the length of the impulse response is 256, it is divided into 2 segments and the successive input blocks are shifted by 32 samples, the adaptation step is 0.33. The forgetting factor in (29) is equal to 0.7 and the weighting factors are $\beta_s=\beta_p=\beta_e=0.98$.

5.4 Evaluation and results

An evaluation of performance to compare the four structures is realized. From signals separately recorded in a car, we consider the simulated recording given Figure 2: the first part includes a noisy echo (Single Talk (ST) Mode) and the second part corresponds to speech added to a noisy echo (Double Talk (DT) Mode).

Three objective measures are tested:

- the Echo Return Loss Enhancement, $ERLE$, which compares the power of the echo at the input and the residual echo at the output:

$$ERLE(k)=10\log\left[\frac{\sum_f\left(|\gamma_{xz}(f,k)|^2/\gamma_{zz}(f,k)\right)}{\sum_f\left(|\gamma_{\hat{s}z}(f,k)|^2/\gamma_{zz}(f,k)\right)}\right]. \qquad (30)$$

This measure does not require the original echo but only the observations x and z, so this measure can be applied on real signals;

- a noise reduction factor which compares the noise power at the input and at the output:

$$R(k)=10\log\left[\frac{\sum_f\gamma_{nn}(f,k)}{\sum_f\left(|\gamma_{\hat{s}n}(f,k)|^2/\gamma_{nn}(f,k)\right)}\right] \qquad (31)$$

- a distortion factor :

$$D(k)=10\log\left[\frac{\sum_f\left(|\gamma_{s(x-\hat{s})}(f,k)|^2/\gamma_{ss}(f,k)\right)}{\sum_f\gamma_{ss}(f,k)}\right]. \qquad (32)$$

These measures are averaged on all blocks corresponding to the presence of the echo or/and the near-end speech, and averaged on ten simulated files. We can compute these measures for different signal-to-noise ratios *SNR* and echo-to-noise ratios *ENR*, computed on the length of speech and echo respectively.

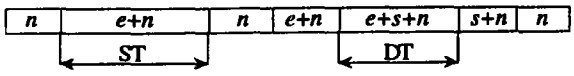
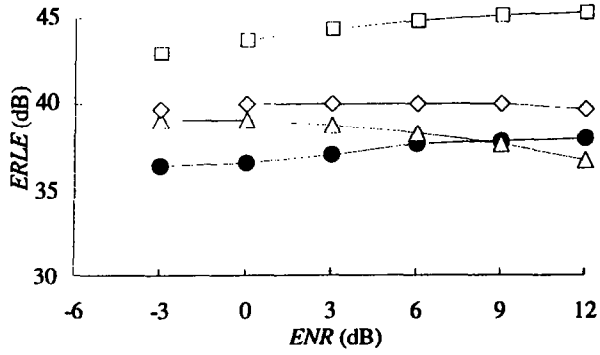
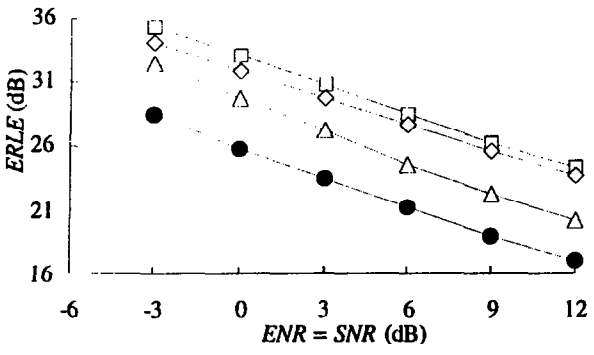


Figure 2. Simulated recording



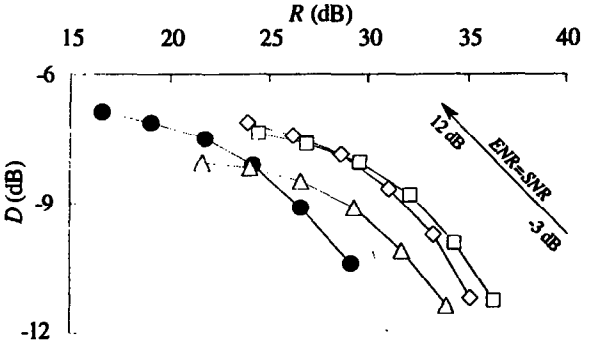
-●- struct. S1, -◇- struct. S2, -□- struct. S3, -△- struct. S4

Figure 3. ERLE in ST mode



-●- struct. S1, -◇- struct. S2, -□- struct. S3, -△- struct. S4

Figure 4. ERLE in DT mode



-●- struct. S1, -◇- struct. S2, -□- struct. S3, -△- struct. S4

Figure 5. D=f(R) in DT mode

Represented in Figures 3 and 4 are the values of the *ERLE* in ST and DT modes. Whatever the mode, the structure S1 realizing the optimal filtering considering two observations shows the lowest ERLE. Among the other structures, the highest *ERLE* in ST and DT modes is obtained by S3 with a more important difference in ST mode. Figure 5 represents *D* versus *R* for different values of *ENR=SNR* varying from -3 dB to 12 dB. We can see that the structure S1 remains less performant. The other structures are practically equivalent, except for low values of *ENR=SNR* for which the structure S4 leads to the best noise reduction for an identical level of distortion. Moreover, informal listening tests show (i) the interest of updating the noise psd in the noise reduction step, mainly for non stationary noises, and (ii) the effective echo cancellation in the new approach.

6. CONCLUSION

We first presented a new noise reduction algorithm where the noise power spectral density is updated during the processing. A new approach to reduce acoustic echo by filtering the microphone observation is investigated. The structures we develop give promising results and they must be further studied to optimize the updating of the noise characteristics as well as the estimators used in the echo cancellation.

Acknowledgements

The authors wish to thank Telecom Paris for the GMDF algorithm and Matra Communication (Paris) for the database.

REFERENCES

[1] R. Martin, P. Vary, "*Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony - State of the Art and Perspectives*", EUSIPCO, Trieste, pp. 1107-1110, Sep. 1996.

[2] F. Capman, J. Boudy, P. Lockwood, "*Acoustic Echo Control and Noise Reduction in the Frequency Domain: A Global Optimisation*", EUSIPCO, Trieste, pp. 29-32, Sep. 1996.

[3] Y. Guélou, A. Benamar, P. Scalart, "*Analysis of Two Structures for Combined Acoustic Echo Cancellation and Noise Reduction*", ICASSP, Atlanta, pp. 637-640, May 1996.

[4] O. Cappé, "*Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor*", IEEE Trans. on Speech and Audio Processing, vol. 2, n°2, pp. 345-349, Apr. 1994.

[5] A. Akbari Azirani, R. Le Bouquin Jeannès, G. Faucon, "*Speech Enhancement Using a Wiener Filtering Under Signal Presence Uncertainty*", EUSIPCO, Trieste, pp. 971-974, Sep. 1996.

[6] S. Haykin, *Adaptive Filter Theory*, 2nd edition, Prentice-Hall, Englewood Cliffs, New Jersey, 1991.

[7] J. Prado, E. Moulines, "*Frequency-Domain Adaptive Filters with Applications to Acoustic Echo Cancellation*", Third International Workshop on Acoustic Echo Cancellation, Lannion, pp. 249-258, Sep. 1993.

[8] B. Ayad, "*Systèmes Combinés d'Annulation d'Écho Acoustique et de Réduction de Bruit pour les Terminaux Mains-Libres*", Thèse de l'Université de Rennes 1, July 1997.