A DOUBLE TALK DETECTOR BASED ON THE PARTIAL COHERENCE FUNCTION

R. LE BOUQUIN JEANNÈS - G. FAUCON

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

ABSTRACT

The growth of mobile radio and teleconference communications requires the design of efficient and robust hands-free systems. In this context, optimisation of acoustic echo and noise reduction is needed. This operation often requires double talk detection, either to choose between different structures or to stop the adaptation of the acoustic echo canceller. In this paper, two microphones and one loudspeaker are considered and a double talk detector based on the partial coherence is investigated. Results are presented on simulated and real signals.

1 INTRODUCTION

In many communication systems, the development of hands-free terminals is growing. This technology brings more comfort and flexibility and also safety when it is used in a car. In such situations, the quality of the speech signal to be transmitted must be satisfactory. Any realization of hands-free telephony has to deal with two major problems. At first, according to the distance between the speaker's mouth and the microphone, the signal-to-noise ratio may be very low. Secondly, the coupling between the loudspeaker and the microphone induces an echo on the microphone input. In consequence, echo and noise have to be reduced [1]. For a few years, a great interest has been devoted to the conception of hands-free technology including noise reduction and acoustic echo cancellation. These problems can be tackled in a combined approach to recover a near-end speech signal only slightly distorted for a sufficient attenuation of echo and noise. It has been proved that the optimal filtering (to estimate the near-end speech signal) in the sense of the minimum mean square error consists of an echo canceller followed by a noise reduction filter [2]. Other approaches (using one or two microphones) have been studied in the literature [3,4,5,6]. Adding microphones or using a microphone array may result in higher performance as an offset to an increased complexity. Moreover, incorporating multiple microphone algorithms can lead to a more efficient noise reduction in the presence of nonstationary noises and reduce reverberation and late echoes simultaneously.

2 MOTIVATION

Whatever the approach is, the Acoustic Echo Cancellation (AEC) realized by an adaptive filtering is disturbed by the other signals: the ambient noise and the speech signal to be transmitted. Regarding noise, it is omnipresent and only prefiltering techniques allow to decrease its effect [2]. Now, the influence of the near-end speech signal on echo cancellation may be reduced by stopping the adaptation or by modifying the adaptation step. Even if the Double Talk (DT) mode (*i.e.* near-end and far-end speech signals present simutaneously) occurs only 20% of time, the signal-to-echo ratio may be important and conduce to a poor acoustic echo cancellation. A Double Talk Detector (DTD) may be introduced with a view to stop the AEC adaptation in DT mode. In this way, the AEC coefficients are not disturbed by near-end speech. In [7], we compared the performance of a combined system when adaptation is stopped or continued; when the input Signal-to-Noise Ratio (SNR) and the Echo-to-Noise Ratio (ENR) are equal to 10 dB, the Echo Return Loss Enhancement (ERLE) is increased by 3 to 4 dB when the adaptation is stopped in the presence of near-end speech: the use of a DTD appears clearly.

More generally, the choice between different structures or between different processing modes may depend on the input signals. For example, in [8], a structure based on the optimal filtering is tested when the noise reduction filtering is derived either from the AEC output or from the microphone observation. In single talk (ST) mode, the first version gives the highest ERLE (from 6 to 12 dB (over the second version) when the ENR is in the range [-3 dB - 12 dB]). In DT mode, the best gain is obtained with the second one (about 2 dB overall the tested conditions). These results confirm the interest of a DTD to distinguish the DT mode from the ST mode. In the following section, we propose a double talk detector based on the ordinary and partial coherences when two microphones and one loudspeaker are available.

3 PROBLEM STATEMENT

3.1 Definitions

The original problem concerns the estimation of a near-end speech signal disturbed by acoustic echoes and ambient noise. In this context, the observations may be written:

$$x_{i,t} = s_{i,t} + n_{i,t} + e_{i,t} \qquad (i = 1,2)$$

in double talk mode, where x represents the observation, s the near-end speech signal, n the disturbing noise and e the echo, i being the channel index.

The coherence function [9] between the two observations x_1 and x_2 is defined by $\rho_{x_1x_2}(f_j)$ with:

$$\rho_{x_1x_2}(f_j)=\frac{\gamma_{x_1x_2}(f_j)}{\gamma_{x_1}^{1/2}(f_j)\gamma_{x_2}^{1/2}(f_j)}$$

where $\gamma_{x_1x_2}(f_j)$ is the cross power spectral density (cross psd) between x_1 and x_2, $\gamma_{x_i}(f_j)$ is the psd of x_i and j is the frequency bin index. This measure varies in module between 0 and 1 and it is representative of the correlated components of x_1 and x_2. The partial coherence [9] between x_1 and x_2 conditioned on the reference signal z emitted by the loudspeaker is $\rho_{x_1x_2/z}(f_j)$:

$$\rho_{x_1x_2/z}(f_j)=\frac{\rho_{x_1x_2}(f_j)-\rho_{x_1z}(f_j)\rho_{zx_2}(f_j)}{\sqrt{1-\left|\rho_{x_1z}(f_j)\right|^2}\sqrt{1-\left|\rho_{x_2z}(f_j)\right|^2}}.$$

This partial coherence corresponds to the coherence between the observations when the echoes have been removed.

3.2 Objective

In the context of mobile telephony, environmental noises are present continuously and four situations can be distinguished (see Table 1). The microphones are far apart so that the noises are considered as decorrelated. The idea is to compute the partial coherence which becomes lower than the coherence in the presence of echo components. Theoretically, the partial coherence falls to 0 when the near-end speech signal disappears. In Table 1, we sum up the expressions of the ordinary coherence and the partial coherence for each case as well as the difference between these two quantities. We note that the differences δ_1 and δ_4 are ideally equal to zero due to the absence of echo. The second and most important point is that the partial coherence is non-zero in the presence of near-end speech. So, in the presence of echo (*i.e.* δ is different from 0), we can compare the module of the partial coherence to a threshold T_1 and, if this module is greater than T_1, the double talk mode is detected.

	$x_i=n_i$ (noise)	$x_i=n_i+e_i$ (single talk mode)	$x_i=n_i+e_i+s_i$ (double talk mode)	$x_i=n_i+s_i$ (noise + near-end speech)
$\rho_{x_1x_2}$	0	$\frac{\gamma_{e_1e_2}}{\sqrt{(\gamma_{e_1}+\gamma_{n_1})(\gamma_{e_2}+\gamma_{n_2})}}$	$\frac{\gamma_{e_1e_2}+\gamma_{s_1s_2}}{\sqrt{(\gamma_{e_1}+\gamma_{s_1}+\gamma_{n_1})(\gamma_{e_2}+\gamma_{s_2}+\gamma_{n_2})}}$	$\frac{\gamma_{s_1s_2}}{\sqrt{(\gamma_{s_1}+\gamma_{n_1})(\gamma_{s_2}+\gamma_{n_2})}}$
$\rho_{x_1x_2/z}$	0	0	$\frac{\gamma_{s_1s_2}}{\sqrt{(\gamma_{s_1}+\gamma_{n_1})(\gamma_{s_2}+\gamma_{n_2})}}$	$\frac{\gamma_{s_1s_2}}{\sqrt{(\gamma_{s_1}+\gamma_{n_1})(\gamma_{s_2}+\gamma_{n_2})}}$
$\delta=\rho_{x_1x_2}-\rho_{x_1x_2/z}$	$\delta_1=0$	δ_2	δ_3	$\delta_4=0$

(Nota Bene: the frequency bin has been dropped for clarity).

Table 1

4 EXPERIMENTS

4.1 Simulations

Simulated signals have been constructed to validate the theoretical results. To create a real situation, we generate first a white gaussian noise, named z_t, to simulate the loudspeaker signal. Then, we derive two autoregressive models, $e_{1,t}$ and $e_{2,t}$, such as:

$$e_{i,t}=z_t-\alpha_i e_{i,t-1}. \qquad (i=1,2)$$

These two signals represent the echoes at the microphones inputs. Then, we generate the useful signals $s_{1,t}$ and $s_{2,t}$ received by the microphones:

$$s_{i,t}=s_t-\beta_i s_{i,t-1}-\lambda_i s_{i,t-2} \qquad (i=1,2)$$

where s_t is a white gaussian noise decorrelated from z_t. The different sequences on the first microphone are indicated on Figure 1 where n_1 represents the noise component on the first microphone. All sequences have the same length (equivalent to 10000 samples). The same sequences are repeated on the second channel.

For these simulations, the values of the parameters are:
$\alpha_1=-0.6,\ \alpha_2=-0.5,\ \beta_1=-0.5,\ \lambda_1=0.8,\ \beta_2=-0.6,\ \lambda_2=0.7$
and the signal to noise ratio is fixed to 6 dB (on each channel) as well as the echo to noise ratio. The first quantity we are interested in is the coherence averaged on the whole frequency bandwidth because (*i*) it is easy to interpret compared with the coherence estimated at each frequency, and (*ii*) the averaging allows to reduce the variance of the estimator. This coherence is computed on each block *k* of 256 samples (256-point FFT):

$$\rho^a_{x_1x_2}(k)=\frac{1}{129}\sum_{j=1}^{129}\left|\rho_{x_1x_2}(f_j,k)\right|.$$

In the same way, we derive the averaged partial coherence on each block *k*; it is defined as:

$$\rho^a_{x_1x_2/z}(k)=\frac{1}{129}\sum_{j=1}^{129}\left|\rho_{x_1x_2/z}(f_j,k)\right|.$$

Finally, we compute the difference between the two previous quantities:

$$\delta^a(k) = \max\left(\rho^a_{x_1x_2}(k) - \rho^a_{x_1x_2/z}(k), 0\right).$$

In this experiment, the psd are estimated using a rectangular window:

$$\gamma_{uv}(f_j,k) = \frac{1}{K}\sum_{l=k-K+1}^{k} U(f_j,l)V^*(f_j,l)$$

where k represents the current block number and K is the number of blocks on which the estimation is performed. We use a 50% overlapping factor and K is equal to 19 (equivalent to 10 adjacent blocks). $U(f_j,l)$ is the Fourier transform of the signal u_t weighted by a Hamming window and the asterisk indicates the conjugate. The estimation of the different quantities is represented on Figure 2. It is obvious that the values of $\rho^a_{x_1x_2/z}$ increase when the near-end speech signal occurs, so that we detect the arrival of this signal. When $\rho^a_{x_1x_2/z}$ and δ^a are greater than thresholds T_1 and T_2 respectively, the double talk mode is detected. However, to determine these thresholds, we must take the bias of the estimator into account. As a matter of fact, we note that the coherence computed from the noises is about 0.2 instead of 0.

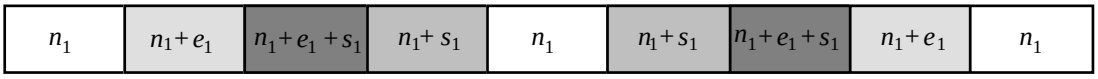

Figure 1. Sequence on the first microphone

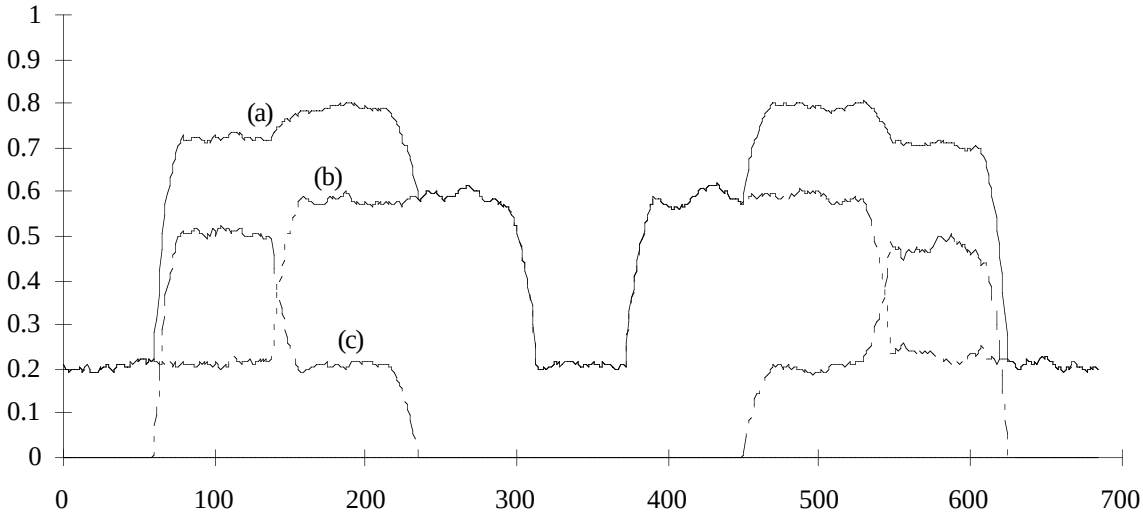

Figure 2. Coherences versus block number
(a) coherence (b) partial coherence (c) difference

4.2 Experiments on real data

We consider a real recording composed of the following sequence: "noise (0 - 0.8 s); echo + noise (0.8 s - 2.6 s); noise (2.6 s - 3 s); echo + noise (3 s - 3.14 s); echo + near-end speech + noise (3.14 s - 5 s); near-end speech + noise (5 s - 5.6 s); noise (5.6 s - 6.5 s); near-end speech + noise (6.5 s - 7.2 s)".
Noise is stationary and proceeds from a car moving at 130 km/h and the echo comes from the coupling between the loudspeaker and the microphones. The signals are recorded independently to choose different SNR and ENR. For the present experiment, the input SNR and ENR are equal to 10 dB. Results on the averaged coherences are shown on Figure 3. The coherence between noises only is not zero, that is due to the fact that the noises are slightly correlated and to the bias on the low coherences. As we expected it, the difference δ^a is different from zero in presence of the echoes and the partial coherence remains high when the near-end speech signal is present. The threshold on the partial coherence is not easy to fix: in ST mode, this coherence is not strictly equal to the coherence between noises. This is explained by the fact that the coherence between each echo and the loudspeaker is less than 1 (about 0.8). Nevertheless, on the Figure 3, a threshold around 0.6 allows to detect the useful signal. In fact, if we replace e_1 and e_2 by z, the partial coherence in ST mode becomes equal to the coherence between noises (Figure 4). These results indicate that a high coherence between the echoes and the loudspeaker is necessary to detect the presence of near-end speech.

5 CONCLUSION

To conclude, the DTD we developed is of great importance to decide if we continue or stop the adaptation of the AEC and more generally to switch from one structure to another according to the presence or the absence of the near-end speech signal. The DTD we propose is able to distinguish the different input conditions and to control the noise and echo cancellation.

References

[1] E. HÄNSLER, "*The Hands-Free Telephone Problem: an Annotated Bibliography Update*", Signal Processing, vol. 27, pp. 259-271, 1992.

[2] G. FAUCON, R. LE BOUQUIN JEANNÈS, "*Joint System for Acoustic Echo Cancellation and Noise Reduction"*, EUROSPEECH, Madrid, pp. 1525-1528, Sept. 1995.

[3] 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, Sept. 1996.

[4] G. FAUCON, R. LE BOUQUIN JEANNÈS, "*Echo and Noise Reduction for Hands-Free Terminals -State of the Art-*", EUROSPEECH, Rhodes, pp. 2423-2426, Sept. 1997.

[5] R. MARTIN, J. ALTENHÖNER, "*Coupled Adaptive Filtered for Acoustic Echo Control and Noise Reduction*", Proc. ICASSP, Detroit, pp. 3043-3046, May 1995.

[6] C. BEAUGEANT, V. TURBIN, P. SCALART, A. GILLOIRE, "*New Optimal Filtering Approaches for Hands-Free Telecommunication Terminals*", Signal Processing, vol. 64, 1, pp. 33-47, 1998.

[7] R. LE BOUQUIN JEANNÈS, G. FAUCON, B. AYAD, "*A Two-Microphone Approach for Speech Enhancement in Hands-Free Communications*", International Conference on Communication Technology, Beijing, pp. 424-427, May 5-7 1996.

[8] R. LE BOUQUIN JEANNÈS, G. FAUCON, B. AYAD, "*How to Improve Acoustic Echo and Noise Cancelling using a Single Talk Detector"*, Speech Communication, vol. 20, pp. 191-202, 1996.

[9] J.S. BENDAT, A.G. PIERSOL, *Random Data: Analysis and Measurement Procedures*, Wiley-Interscience, 1971.

Acknowledgment. *The authors wish to thank Matra Communication (Paris) for the database.*

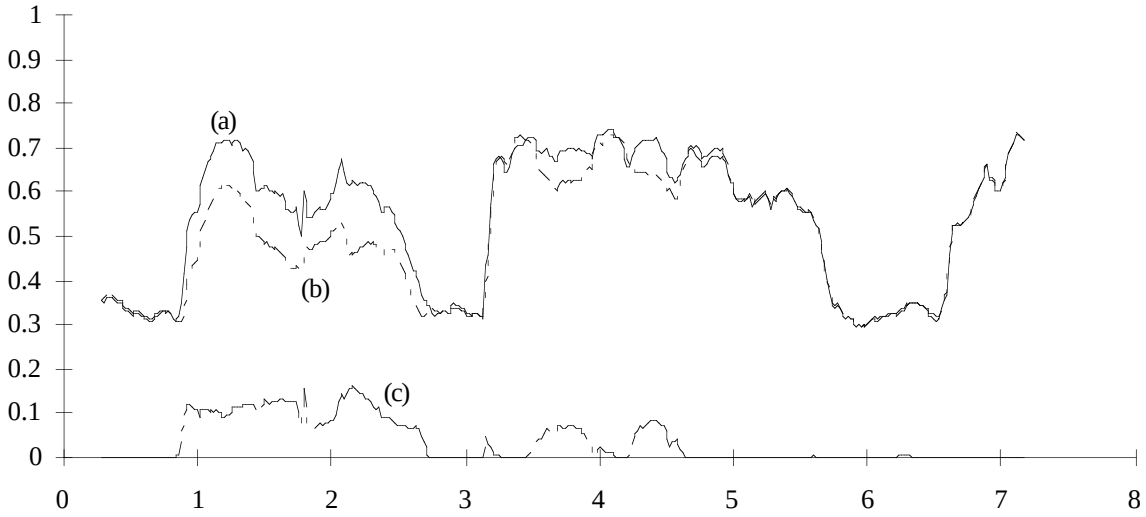

Figure 3. Coherences versus time (in seconds)
(a) coherence (b) partial coherence (c) difference

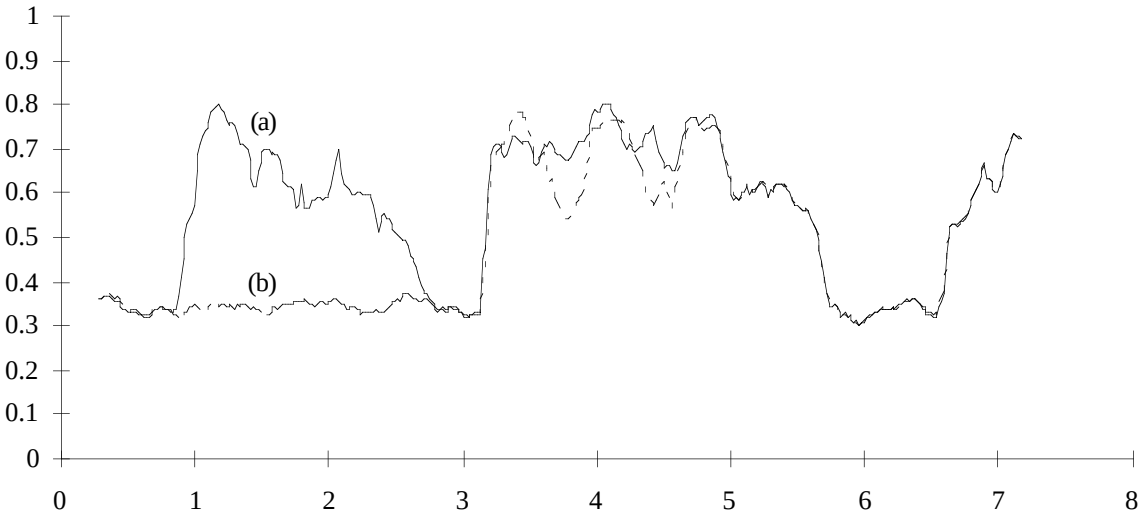

Figure 4. Coherence and partial coherence versus time (in seconds)
(a) coherence (b) partial coherence