PERFORMANCE OF ADAPTIVE DEREVERBERATION TECHNIQUES USING DIRECTIVITY CONTROLLED ARRAYS

C. Marro, Y. Mahieux*, K. U. Simmer***

*FRANCE TELECOM - CNET LAA/TSS/CMC
Technopole Anticipa, 2 avenue Pierre Marzin 22307 Lannion Cedex - FRANCE

**Houpert Digital Audio, Wiener Str 5, D-28359 Bremen, GERMANY

e-mails: marro@lannion.cnet.fr - mahieux@lannion.cnet.fr - u.simmer@proaudio.de

ABSTRACT

The use of optimal postfiltering has been previously proposed to increase the performance of microphone arrays. In this paper, an analysis of the postfilter shows that its behaviour is closely related to the one of the array. This is illustrated by considering a typical videoconferencing context. The results we provide demonstrate that the use of a directivity controlled array is a requirement to ensure a sufficient robustness of the whole system. It is also shown that the dereverberation performed by the postfilter is limited and that its main interest lies in a significant reduction of the acoustic echo even in the double talk case. This attractive property depends on the whole design of the array including its placement versus the acoustic echo sources.

1 INTRODUCTION

In hands-free voice communication systems, the first transformation of speech is due to the acoustical environment. The reverberation, the ambient noise and the acoustic echo degrade the quality of the signal produced by the talker and in adverse environments, such as in mobiles radio, can even reduce the intelligibility. Several methods for restoring the original speech signal based on a multi-sensors system, have been proposed in the literature. For example, microphone arrays take advantage of the space discrimination between the desired and undesired signal sources. The performance of microphone arrays varies as a function of the number of sensors which is inevitably limited due to practical considerations. Making microphone arrays adaptive does not really improve their performance in real acoustic environments [1].

Another family of techniques which could be named "microphone array with optimal postfilter", consists of removing the non-coherent parts of the recorded signal by filtering it. In such a system, as described in figure 1, the output signal of the array is filtered by a time varying filter, the transfer function of which is derived from the cross spectral densities of the sensor signals. This technique has been previously investigated by a number of researchers [2-4]. Most papers deal with the improvement of the transfer function expression or with the practical realisation of the filtering. In [5], a theoretical analysis of this method is proposed. It is shown that the properties of the postfilter are tightly related to the characteristics of the array. This is illustrated in the present paper by comparing the impact of the directivity control of the array over the performance of the whole system. Unlike in [5] where only the asymptotic behaviour of the postfilter and the reduction of the reverberation were considered, the results given in this paper are related to the real performance of the system in a complete videoconferencing context.

2 BASIC PROPERTIES OF THE SYSTEM

2.1 Microphone Arrays Properties

The basic properties of the rectilinear uniformly spaced arrays are described in several papers [6], [7]. Let us just recall that if the inter-sensors spacing, *d*, exceeds a threshold which depends on the steering angle φ_0 and on the wavelength λ, grating lobes appear. At each frequency, an optimal spacing exists which avoids grating lobes while maximising the diffuse noise reduction yielded by the array as measured by its directivity factor. In "directivity controlled arrays", this optimality is extended to a wide frequency band by means of (for example) harmonically nested subarrays [6].

2.2 Postfilter Properties

Among the various expressions of the postfilter transfer function, we have retained the one proposed in [3]:

$$W(f) = \frac{\sum_{i=1}^{N}\left|a_i(f)\right|^2}{\mathrm{Re}\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i(f)a_j^*(f)\right\}} \cdot \frac{\mathrm{Re}\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i(f)a_j^*(f)\hat{\Phi}_{v_iv_j}(f)\right\}}{\sum_{i=0}^{N-1}\left|a_i(f)\right|^2\hat{\Phi}_{v_iv_i}(f)} \quad (1)$$

where $\hat{\Phi}_{v_iv_j}(f)$, (resp. $\hat{\Phi}_{v_iv_i}(f)$) are the estimated cross (power) spectral densities (csd and psd) of the delayed input signals $v_i(n)$ and $v_j(n)$ (see figure 1). $\{a_i(f)\}$ is the set of microphone signals weightings at frequency *f*. Although this expression of *W(f)* overestimates the noise psd [4], it has been observed that it gives the highest dereverberation gain. Let us note that the original expression as proposed in [3] has been modified so that *W(f)* is estimated from the delayed and filtered signals $a_i(n)*v_i(n)$ [5]. Normalising factors have also been added in order to ensure a transfer function of the whole system (array + postfilter) equal to unity when only the desired signal is present.

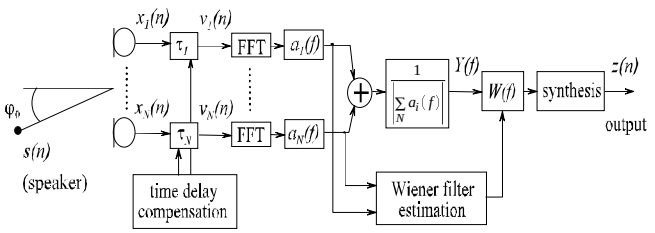

FIG. 1. *Microphone array combined with postfilter.*

Let the input signals be stationary and the delayed signals $v_i(n)$ be defined as:

$$v_i(n) = s(n) + n_i(n), \qquad i = 1,\ldots,N \quad (2)$$

where *s(n)* is the desired speech signal and $n_i(n)$ the noise signal at microphone *i*. Eq. (2) involves a perfect steering of the array. Let us furthermore assume that the noise signals $n_i(n)$ are uncorrelated with the desired signal *s(n)* and that the spectral densities are perfectly known. It can be shown [5] that the transfer function of the filter is:

$$W(f) = \frac{SNR(f)}{1+SNR(f)} + \frac{\left|\sum_{i=1}^{N} a_i(f)\right|^2 \left(NR_a(f)\right)^{-1} - \sum_{i=1}^{N}\left|a_i(f)\right|^2}{\left(1+SNR(f)\right)\left[\left|\sum_{i=1}^{N} a_i(f)\right|^2 - \sum_{i=1}^{N}\left|a_i(f)\right|^2\right]} \qquad (3)$$

where *SNR(f)* is the input signal to noise ratio and $NR_a(f)$ is the noise reduction factor of the array. $NR_a(f)$ is defined as the ratio of the psd of the noise at the input to the one at the output of the array. $NR_a(f)$ depends on the space distribution of the noise sources and on the array transfer function. We also define the noise reduction factor of the postfilter, $NR_p(f)$ as the ratio of the noise psd at the output of the array to the one at the output of the postfilter:

$$NR_p(f) = 1/\left|W(f)\right|^2 \qquad (4)$$

The variation of $NR_p(f)$ as a function $NR_a(f)$ is a somewhat complicated quadratic function which can present singularities. In spite of this, the general trend is that $NR_p(f)$ varies as a monotonously increasing function of $NR_a(f)$ [5]. Furthermore, with usual array coefficients $\{a_i(f)\}$, $NR_p(f)$ is greater than one. As a results the postfilter yields an extra reduction of the noise which is tightly related to the one of the array. Thus, to obtain a high performance of the whole system, the array should be carefully designed by taking into account the peculiarities of the acoustical environment. To maintain an approximately constant performance of the whole system over a wide frequency range, a directivity controlled array should be used. With a non controlled array, grating lobes can be steered at some frequencies towards noise sources resulting in a severe reduction of $NR_p(f)$ and in a higher spectral shaping of the output signal. It can also be shown [5] that with a directivity controlled array, the whole system is much less sensitive to the steering errors. It is worthy of note that at low frequencies, the postfilter does not improve the poor performance of the array since $NR_a(f)$ and consequently $NR_p(f)$ tend to one.

At high values of *SNR(f)*, *W(f)* tends to one and there is no noise reduction. For example, for a diffuse noise field, at *SNR(f)* values higher than 10 dB, the extra noise reduction yielded by the postfilter averaged over the whole frequency range is lower than 1 dB [5]. If the Signal to Noise Ratio is low, the behaviour of the filter is mainly influenced by the array characteristics. The fact that the postfilter is applied as well to *s(n)* as to the noise results in a spectral shaping of the desired signal. This distortion is low if the psd of *s(n)* does not overlap the one of the noise. To take advantage of such an opportunity, the frequency resolution used for the computation and the implementation of the filter should be as high as possible (see below).

3 PRACTICAL REALIZATION

3.1 Array and Postfilter Realisation

The array we consider in this paper is composed of eleven unidirectional microphones grouped into 4 subarrays as follows:

- subarray 1 : sensors	1, 2, 6, 10, 11	- spacing $d(1)$=32 cm
- subarray 2 :	2, 3, 6, 9, 10	$d(2)$ = 16 cm
- subarray 3 :	3, 4, 6, 8, 9	$d(3)$ = 8 cm
- subarray 4 :	4, 5, 6, 7, 8	$d(4)$ = 4 cm

The four subarrays filters are linear phase FIR band-pass filters (61 taps) which are computed in such a way that the transfer function of the array equals unity for the steering direction [7]. The high cut-off frequencies, $f_c(i)$ of these filters are chosen to optimise the directivity factor of the array and are therefore dependent on the steering direction φ_0:

$$f_c(i) = 0.8 * \frac{c}{d(i).\left(1+\sin(\varphi_0)\right)} \qquad i = 1, ..., 4 \qquad (5)$$

In the subsequent sections, this array is compared to a "non directivity controlled" one, the output signal of which is simply the one delivered by subarray 1. The choice of this subarray is justified by the fact that it provides the highest noise reduction at low frequencies. Although their total number of sensors are different (11 vs. 5), the two arrays have the same number of sensors at each frequency and almost the same directivity factor. The talker localisation procedure is not addressed in this paper. Unless it is explicitly mentioned, an exact localisation of the desired source is assumed.

The postfilter implementation is performed with the help of the WOLA technique. The window length is 512 samples (32 ms at a sampling rate of 16 kHz) with an overlap of 256 samples. The postfilter implementation is made in order to obtain a linear convolution. The analysis and synthesis windows verify the perfect reconstruction property. Time aliasing is avoided by zero padding the block of samples before FFT (FFT size equal to 1024 samples) and by constraining the convolution to be linear. This is performed by computing the inverse DFT of *W(f)* and truncating the resulting impulse response, *w(n)*, to 511 samples. *w(n)* is then transformed back to the frequency domain for fast convolution with the array output signal. The constraint of linear convolution requires 2 additional FFT's but if it is not respected, the resulting circular convolution produces perceptible time artefacts.

The spectral densities are estimated by means of an exponentially weighted averaging procedure of the Welch periodogram. The time constant is equal to 64 ms which has been found to be a compromise between a low variance of the estimator and a fast updating of *W(f)* to follow the non-stationarities of the input signals.

It can be shown [5] from Eq. (3), that for linear phase array coefficients, $\{a_i(f)\}$, $\left|W(f)\right| \le 1$. But due to estimation errors of the spectral densities, this property is not always respected. To avoid artificial amplification of the signal, $\left|W(f)\right|$ is constrained to the range [-1, 1].

3.2 Experimental Set-up

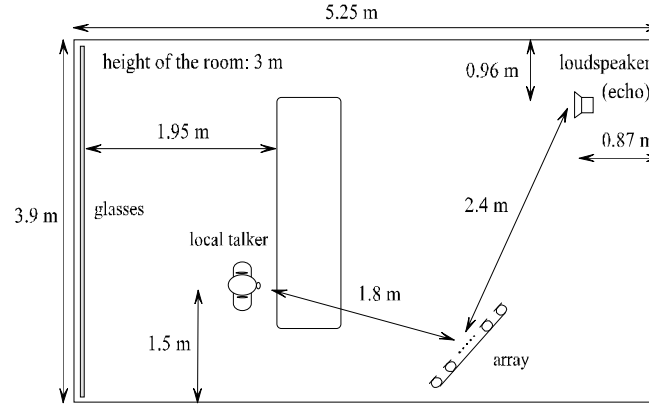

FIG. 2. *Experimental set-up.*

To evaluate the performance of the method in a real videoconferencing context, impulse response measurements have been performed in the room described in figure 2. Its dimensions are 5.25 x 3.9 x 3m and its reverberation time is approximately 500 ms.

The microphone signals $x_i(n)$ are obtained by filtering anechoic speech signals by the measured impulse responses between the two loudspeakers (one simulating the talker plus one for the acoustic echo) and the eleven sensors of the array.

3.3 Objective Measurements

The noise reduction - the word "noise" includes all the perturbations (echo, reverberation, ..) - is measured by means of a two step procedure as described in [8]. First, the postfilter is computed from the microphones signal $\{x_i(n)\}$ and stored at each frame. Then, the whole system (array and stored filter transfer functions) is applied to the noise only. This enables to measure the real noise reduction factor of the array, $NR_a(f)$ and of the postfilter, $NR_p(f)$. To compute $NR_a(f)$, the psd of the noise at the input of the array is estimated as the noise energy averaged over all the sensors. This enables to avoid biased estimation due to the zeros of the room transfer functions.

The beamforming and the stored filters are also applied to the signals obtained by filtering the anechoic signal, $s(n)$, by the direct path of the measured impulse responses. The distortion is given by the time variation of the cepstral distance [9], $d_c(t)$, between the reference signal and the corresponding output signal, $z(n)$. The reference signal is defined as the filtering of $s(n)$ by the direct path associated to the center microphone. $d_c(t)$ is measured over consecutive time segments of 16 ms duration.

4 RESULTS

4.1 Reverberation Reduction

Let us first consider the case that only local talker is active. The anechoic signal we have used is a french sentence of duration 2.5 s: "Les deux camions se sont heurtés de face" pronounced by a male speaker. Table 1 contains the measured values of the noise reduction factors per octave bands for the controlled array structure. The performance of the array is better than the one indicated by the directivity factor. This is due to the influence of the first reflections of the room impulse responses which are severely reduced by the array.

Octave band	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	[0-7 kHz]
Q(f) (dB)	1.37	2.76	3.85	3.91	3.68	4.41	3.87
NRa(f) (dB)	1.67	5.92	8.89	5.51	6.41	7.75	5.52
NRp(f) (dB)	-0.33	1.21	2.55	0.88	1.02	1.83	0.58

TABL. 1. *System with controlled array: directivity factor Q(f), $NR_a(f)$ and $NR_p(f)$ measured on the reverberation.*

On the other hand, the extra reduction of the reverberation yielded by the postfilter does not exceed 2.55 dB and is somewhat disappointing. This poor performance can be explained by the existing correlation between the desired signal (the direct path) and the noise (the reverberation). This correlation results from the properties of speech. The original pitch structure is preserved by the filtering by the room impulse responses. Thus, the desired signal and the noise overlap in the frequency domain and cannot be discriminated by the postfilter. Taking advantage of the time variations of the pitch would require a too high frequency resolution which is not compatible with the implementation requirements. The time variation of $NR_p(f)$ is not constant. It is much higher in the time segments where only the reverberation is present. These segments correspond to low values of *SNR(f)* [8]. The results is that the perceived reduction of the reverberation is quite significant and is higher than the one indicated by $NR_p(f)$.

It should be noted that due to the above mentioned correlation, the theoretical predictions of section 2 are not valid since $s(n)$ and $n_i(n)$ were assumed to be uncorrelated. The analysis of the system in the case of a reflection [5] provides some indications about the expected performance. It is shown that the postfilter behaviour is similar to the one of the array and that $s(n)$ is attenuated at the frequencies the array cancels the reflection. However, in the current example, unacceptable distortions of the signal have not been observed. The cepstral distance $d_c(t)$ is lower than 0.25 (see figure 3) except in the case of "sharp" non stationarities of the speech signal. It is currently admitted that if $d_c(t)$ is lower than 0.5, the distortion is not audible. This is confirmed by the listening we have performed.

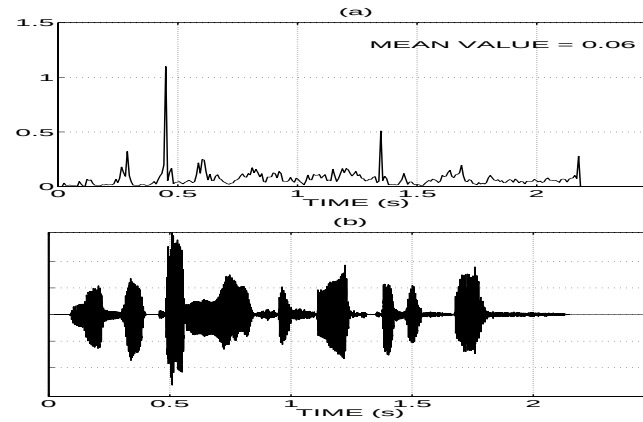

FIG. 3. *(a): Cepstral distance, $d_c(t)$ for the controlled array (local talk only); (b): anechoic signal, s(t).*

Table 2 contains the results for the non directivity controlled array. They are similar to the one of table 1 which can be explained by the likeness of the two directivity factors. On the other hand, the distortion of the desired speech signal by the postfilter is much higher (see figure 4) and is quite perceptible. It is partly due to the impact of the grating lobes which results in a spectral shaping of the output signal. But the main cause of this distortion is related to the steering misadjustment. Due to the narrowness of the main beam of the non directivity controlled array at high frequencies, steering errors involve a filtering of the desired signal by the array. This spectral shaping is increased by the postfilter [5]. To illustrate this feature, steering mismatches equal to 10° for the controlled array system and to 2° for the non controlled one have been simulated. The corresponding cepstral distances are shown in figure 5. The distortion in the case of the non controlled array is clearly much higher. Listenings indicate that the distortion is very annoying (severe low pass filtering) whereas with the controlled array, in spite of the 10° mismatch, the perceived quality remains acceptable.

Octave band	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	[0-7 kHz]
Q(f) (dB)	1.37	2.80	4.49	4.68	5.04	5.00	4.59
NRa(f) (dB)	2.10	6.03	8.22	5.78	5.90	5.30	5.18
NRp(f) (dB)	-0.33	1.25	2.46	0.93	0.84	1.47	0.13

TABL. 2. *Same as table 1 for the system with non-controlled array.*

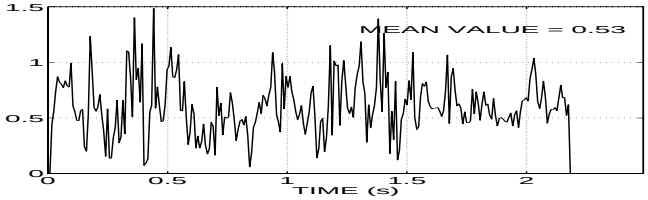

FIG. 4. $d_c(t)$ for the non controlled array (local talk only).

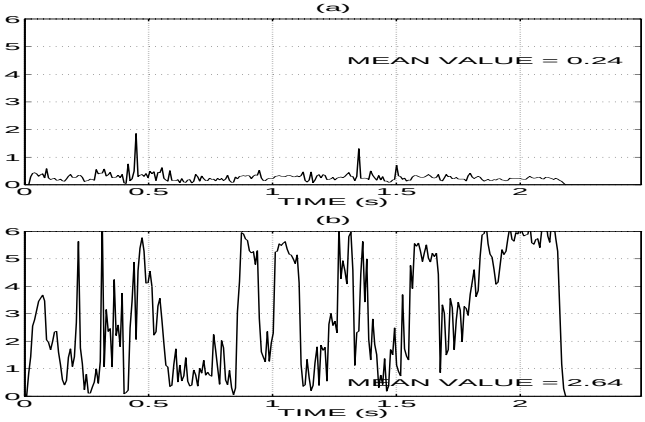

FIG. 5. *(a): $d_c(t)$ for the controlled array, 10° steering mismatch; (b): non-controlled array, 2° steering mismatch.*

A localisation system has to take into account the talker movements and will hardly deliver a steering angle with a precision better than 1°. It must be noted that in the current experiment, a mismatch of 2° corresponds to a displacement of 6 cm at the talker position. Thereby, the use of a directivity controlled array is absolutely necessary and is a requirement for achieving a good quality.

4.2 Acoustic Echo

To analyse the behaviour of the system in case of acoustic echo, we have considered the double talk configuration. The local talk is active as in the previous sub-section. The acoustic echo signal is a french sentence ("Là bas, il y a de mauvaises vagues très hautes") of duration 2.5 s pronounced by a female speaker. Only the directivity controlled array is studied.

The values of $NR_a(f)$ measured on the echo are quite significant (table 3). This is a consequence of the relative positions of the loudspeaker and of the array in the room. The low value at the octave centred around 500 Hz is related to the coincidence between harmonic of the echo signal and zeros of the echo path transfer function for some microphones.

Octave band	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	[0-7 kHz]
NRa(f) (dB)	6.53	11.56	3.80	8.47	5.48	6.83	9.68
NRp(f) (dB)	3.06	12.62	7.23	8.07	5.55	4.59	8.87

TABL. 3. *$NR_a(f)$ and $NR_p(f)$ measured on the echo (controlled array).*

The extra reduction of the echo yielded by the postfilter is also quite appreciable. The frequency variation of $NR_p(f)$ measured on the echo is similar to the ones of $NR_a(f)$ which confirms the theoretical prediction relative to the link between the array and postfilter behaviours. The non-correlation between the desired signal and the echo enables an optimal performance of the postfilter (non overlapping spectral domains) provided that the frequency resolution is sufficient. This is confirmed by the fact that the values of $NR_p(f)$ measured on the reverberation and of $d_c(t)$ are similar to the ones obtained in the case of local talk only. The averaged measured cepstral distance is lower than 0.5 (see figure 6).

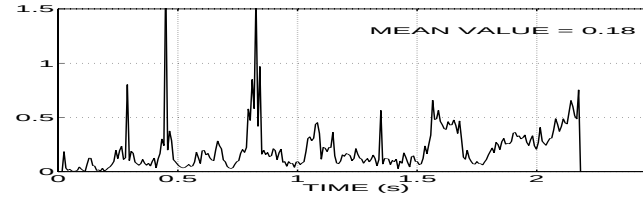

FIG. 6. *$d_c(t)$, double talk case (controlled array).*

A similar experiment has been conducted in the case of echo only configuration. The echo reduction provided by the postfilter is equal to 15.1 dB. The one of the array is the same as for double talk which yield a total reduction of the acoustic echo about 25 dB.

5 CONCLUSION

The results we have presented in this paper show that the postfilter significantly improves the whole performance of the microphone array in a videoconferencing context. From the point of view of the reverberation reduction only, this improvement is limited. This is partly related to the strong correlation between the desired speech signal and the reverberant components. On the other hand, the postfilter yields an appreciable extra reduction of the acoustic echo even in a double talk configuration. This paper also confirms that the performance of the postfilter is closely dependent on the underlying microphone array characteristics. It follows that the array should be carefully designed. First a control of its directivity is required for a sufficient robustness of the whole system. Then the position of the array in the room is very critical. It is the whole sound processing system including the loudspeakers and the microphone array which must be adapted to the videoconferencing room. Finally, this study should be completed by a similar analysis (and an adaptation) of the array combined with postfilter in other environments such as mobiles radio or multimedia workstations.

REFERENCES

[1] G. Elko, "Microphone array systems", *Proc. Int. Workshop. on Acoustic Echo Control*, pp. 31-38, Röros, Norway, 1995.

[2] J.B. Allen et al, "Multimicrophone signal processing technique to remove reverberation from speech signals", *JASA* , vol. 62, N° 4, pp. 912-915, 1977.

[3] R. Zelinski, "A microphone array with adaptive postfiltering for noise reduction in reverberant rooms", *Proc. ICASSP'88*, pp. 2578-2581, New York, USA, 1988.

[4] K.U. Simmer et al, "Suppression of coherent and incoherent noise using a microphone array", *Annales des Télécom.*, tome 49, n° 7-8, pp. 439-446, July 1994.

[5] C. Marro, Y. Mahieux, "Analysis of dereverberation and noise reduction techniques based ..", *submitted to IEEE Transaction on Speech and Audio*.

[6] J.L. Flanagan et al, "Autodirective microphone systems", *Acustica*, vol. 73, pp 58-71, 1991.

[7] Y. Mahieux et al, "A microphone array for multimedia workstations", *Journal of the AES*, May 1996.

[8] Y. Mahieux, C. Marro, "Comparison of dereverberation techniques for videoconferencing applications", *100th AES convention*, Copenhagen, 1996

[9] R.F Kubichek, "Standards and technology issues in objective voice quality assessment", *Digital Signal Processing*, vol. 1, pp. 38-44, 1991.