MICROPHONE ARRAY RESPONSE TO SPEAKER MOVEMENTS

Yves GRENIER

ENST - Département Signal
46 rue Barrault, 75634, Paris 13, France.
e-mail: `grenier@sig.enst.fr`

Sofiène AFFES

INRS-Télécommunications
Ile des Soeurs, Verdun, H3E 1H6, Canada.
e-mail: `affes@inrs-telecom.uquebec.ca`

ABSTRACT

Matched filtering and adaptive beamforming are both necessary for efficient speech dereverberation and noise reduction by microphone arrays. This can be achieved by the identification of impulse responses. In this contribution, we show that adaptive microphone arrays are sensitive to identification errors of impulse responses, particularly due to speaker movements. We prove that *adjusted* matched-filtering and permanent tracking of impulse responses are also necessary. The proposed microphone array responds well to these requirements under realistic conditions.

1. INTRODUCTION

Successful microphone array processing of speech should achieve speech dereverberation and efficient noise reduction, and should also show a high adaptation capacity to speaker movements [1]. So far, these issues have been addressed separately. In [6],[7] we implemented these requirements and proved the efficiency of the proposed microphone array in noise reduction and speech dereverberation. We herein assess its robustness and response to speaker movements.

In [1],[2], the problem of speech dereverberation is addressed. The dereverberation capacity of matched-filter processing proposed therein is reported. However, fixed impulse responses are either measured or calculated from the room geometry, and the speaker movements are not tracked. An adaptive identification procedure of impulse responses would allow a permanent tracking (*i.e. adjusted* matched-filtering). Besides, a very large number of microphones is used with an underlying Delay-Sum beamforming structure. This structure is suboptimal for noise reduction and its performance increases only with a larger number of microphones. A more efficient beamformer would require a smaller number of microphones.

In [3],[4], the problem of noise reduction is addressed. The noise reduction capacity of adaptive beamforming such as the GSC structure [5] is reported therein. However, speech dereverberation is not implemented and speech is assumed to propagate in free space. It is known though, that adaptive beamforming is very sensitive to propagation modeling errors and to speaker location uncertainties and that speech cancelation may occur. To reduce this phenomenon, suboptimal GSC structures for noise reduction are finally proposed in [3],[4]. Although the geometrical location of the speaker is particularly tracked in [4], the multipath signals due to reflections and reverberation do not allow for an optimal implementation of the GSC structure for noise reduction and speech remains reverberated. A combination with an *adjusted* matched-filter processing would achieve better performance.

Contrary to previous methods [1]-[4], we addressed in [6],[7] all the above issues together and proposed an efficient microphone array relying on adaptive identification of impulse responses. Based on this characterization, we dereverberate speech by *adjusted* matched filtering as in [1],[2], and optimally reduce noise by GSC beamforming [5] as in [3],[4]. In [6],[7], we assessed the capacity of the proposed microphone array in noise reduction and speech dereverberation. In this contribution, we specifically evaluate its sensitivity and its response to local speaker movements and show its capacity to track them in real situations.

2. CONFIGURATION AND ALGORITHM

We consider for our application an array of $m = 12$ microphones located around the screen of a computer workstation in a large banker market trading room of 30 m length by 20 m width and 3 m height[1]. In this large room, we measure at different locations a quite constant reverberation time of about 1.7 s. As shown in Fig. 1, six microphones are linearly placed along the top edge of the workstation, and six others are placed on both the left and right edges. The spacing between each pair of adjacent sensors is 0.07 m. This array feeds the front-end receiver of a hands-free telephone installed on an operator desk. The loudspeaker is fixed to the keyboard. Three nominal positions of the operator are considered for the study (*i.e.* center, left and right, see Fig. 1).

We model the signals received from the array of $m = 12$ microphones in the frequency domain as follows:

$$X_{f,n} = G_{f,n} s_{f,n} + N_{f,n} \ . \qquad (1)$$

The subscripts $f = 0, \cdots, 2L-1$ and n respectively denote in (1) the FFT over $2L$ snapshots of the indexed quantity at the frequency bin f and the block of input data number n. Notice that signals are received at a sampling frequency of 8 kHz. In (1), $X_{f,n}$ denotes the m-dimensional observation vector, $s_{f,n}$ is the speech signal emitted from the operator and $N_{f,n}$ is the noise vector. Noise particularly contains cocktail party speech, double talk, and possibly a strong

[1] Recordings were made by ENST, France, and PAGE Iberica, Spain, in Banesto in Madrid, Spain.

0-8186-7919-0/97 $10.00

echo emitted from the loudspeaker. The m-dimensional vector $G_{f,n}$ denotes the impulse responses from the operator's mouth to the microphones in the frequency domain.

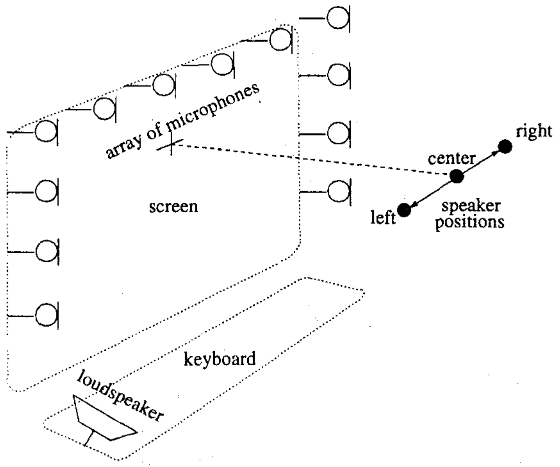

Figure 1: Configuration of microphone array in a banker market trading room.

Let us define the mean energy of impulse responses in the frequency domain by $\beta_f^2 = \|G_f\|^2/m$. In [6],[7], we observed that β_f^2 is constant for any location of the speaker around the central position shown in Fig. 1, and that it can be measured once for all. This property solves the ambiguity of the multiplicative factor between $G_{f,n}$ and $s_{f,n}$. We hence rewrite equation (1) as follows:

$$X_{f,n} = \alpha_{f,n}\, U_{f,n} + N_{f,n}\ , \qquad (2)$$

where the complex vector $U_{f,n} = G_{f,n}/\beta_f$ is the propagation or signal subspace basis vector with norm $\sqrt{m}$, and where the complex scalar $\alpha_{f,n} = \beta_f\ s_{f,n}$ is the signal parameter. In the following, we view this equation as an identification problem in the narrowband case. We hence apply a beamforming and subspace-tracking method to extract the speech $s_{f,n}$ (*i.e.* $\alpha_{f,n}$) and to fully identify the impulse responses $G_{f,n}$ (*i.e.* $U_{f,n}$) as shown below.

Actually, the main structure of the algorithm for speech acquisition and noise reduction is well described in [6]. It can be also inferred from the generalization to echo cancelation and speech acquisition in double talk situations in [7]. Here we briefly mention the major steps of the algorithm.

We first assume that an estimate of $U_{f,n}$ at iteration n say $\hat{U}_{f,n}$ is available. Then we estimate the signal parameter and efficiently reduce the noise by a GSC beamformer [5]:

$$\hat{\alpha}_{f,n} = \hat{y}_{f,n} - W_{f,n}^H P_{f,n}^H Y_{f,n}\ , \qquad (3)$$

where $\hat{y}_{f,n} = \hat{U}_{f,n}^H X_{f,n}/m$ is the output of an *adjusted* matched-filter with a Delay-Sum structure. This filter compensates the impulse responses and dereverberates speech. On the other hand, the adaptive filter $W_{f,n}$ of the GSC side-structure optimally reduces colored noise from the synchronized inputs $Y_{f,n} = \mathrm{diag}[\hat{U}_{f,n}^H]X_{f,n}$ after projection by the blocking matrix $P_{f,n}$ (*i.e.* $P_{f,n}^H \hat{U}_{f,n} = 0$) [5].

We secondly identify $U_{f,n}$ by a subspace-tracking procedure [6],[7]:

$$\tilde{U}_{f,n+1} = \hat{U}_{f,n} + \delta_n \mu_{f,n} (X_{f,n} - \hat{U}_{f,n}\hat{y}_{f,n})\ \hat{y}_{f,n}^H\ , \qquad (4)$$

where $\mu_{f,n}$ is the step-size of the LMS-like tracking equation possibly including a normalization factor, and where δ_n is a voice activity detector equal to 1 during speech activity and 0 otherwise [6],[7]. Notice that $\tilde{U}_{f,n+1}$ denotes an unconstrained estimate of $U_{f,n+1}$. In a block processing scheme, we actually have to force the structure of the resulting impulse responses in the time domain to correspond to a linear convolution. This step which amounts to setting the last half of impulse response coefficients in the time domain to zero provides the constrained estimate $\hat{U}_{f,n+1}$.

With blocks shifted each K snapshots, we finally estimate the speech signal at the block $n{+}1$ in an OLS (overlap-save) scheme by:

$$[\hat{s}\,(K(n+1))\,,\cdots,\hat{s}\,(K(n+1)+2L-1)] = \\ \mathrm{Re}\left\{\mathrm{IFFT}\left(\left[\tfrac{\hat{\alpha}_{0,n}}{\beta_0},\cdots,\tfrac{\hat{\alpha}_{2L-1,n}}{\beta_{2L-1}}\right]\right)\right\}. \qquad (5)$$

As blocks overlap over $2L - K$ samples, we only keep the segment containing the first K samples. We further reduce the residual noise by spectral subtraction [6].

In comparison to the above algorithm, matched filtering was proposed in [1],[2] for speech dereverberation, but fixed impulse responses are measured or computed therein and optimal beamforming was not considered for noise reduction. On the other hand, GSC beamforming was used in [3] and [4]; but without matched filtering speech cancelation and/or suboptimal noise reduction are observed and speech remains reverberated. Contrary to these methods, we dereverberate speech and optimally reduce noise. The effectiveness of our method relies on an adaptive identification of impulse responses.

3. EVALUATION

As we increase the degrees of adaptivity of the microphone array, it becomes more sensitive to nonstationarities. In this section, we specifically evaluate the sensitivity of the proposed array to identification errors of impulse responses and assess its response to speaker movements with quantitative measurements.

To do so, we shall need to synthesize simulated data so as to access these measurements. We actually take special care to make our experiments with simulated data very close to reality. Indeed, we record in an anechoic room a clean speech signal uttered from a speaker to simulate the original speech of the operator. We then convolve it with the impulse responses measured inside the trading room from any selected nominal position of the speaker to the array of microphones (see Fig. 1). This convolution faithfully reproduces the reverberation effect of the large banker market trading room. The convolved signals are finally corrupted at a mean SNR of 7 dB by a real background noise recorded separately at work time in the trading room. The background noise contains cocktail party speech due to the large number of operators present in the trading room, the noise of keyboards, the noise of the workstation fans, etc$\cdots$, and makes the experiment very close to reality.

3.1. Sensitivity and response to identification errors from initialization

We first assess the sensitivity of the array to identification errors when the operator speaks from central position and when impulse responses are started with a geometrical propagation [6]. We skip the tracking step of equation (4) (*i.e.* $\mu_{f,n} = 0$). This amounts to the simple TDC (time delay compensation) usually employed [3],[4].

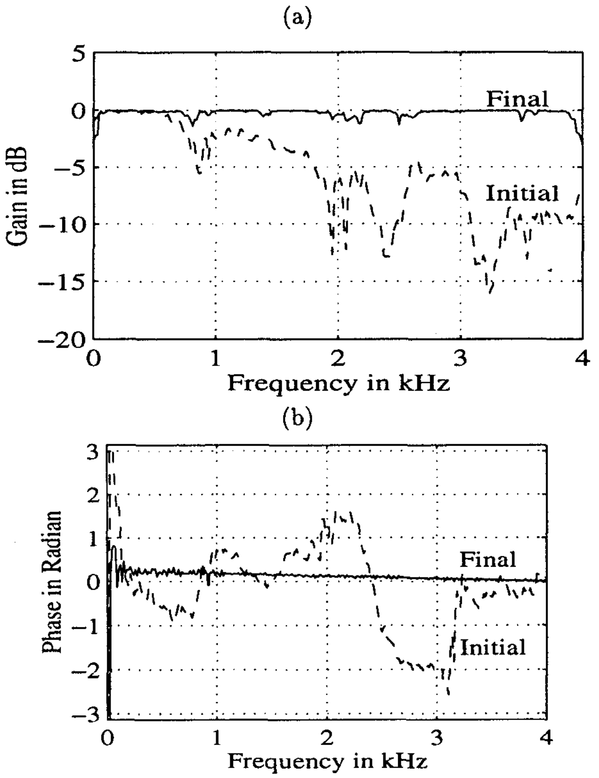

Figure 2: Total response $(\hat{U}_{f,n}/m - P_{f,n}W_{f,n})^H G_{f,n}/\beta_{f,n}$ before (*i.e.* initial) and after 1 s (*i.e.* final) of speech activity. (a) Gain in dB. (b) Phase in Radian.

In Fig. 2-a, we plot the gain of the total response from the central position of the speaker to the processor output (*i.e.* $|(\hat{U}_{f,n}/m - P_{f,n}W_{f,n})^H G_{f,n}/\beta_{f,n}|^2$). The initial curve corresponds to TDC, and shows the usual approximation of classical arrays to be inadequate beyond a small low frequency region. The final curve corresponds to the identified impulse responses after convergence of (4) within 1 s from speech activity start, and shows that signal leakage is quite negligible. Despite the small distortions in amplitude and phase observed in Fig. 2-a and Fig. 2-b respectively, the audible quality of the output speech sounds very natural while point jammers are significantly reduced. We actually measure at output a clarity index of 18 dB [6], which is higher than the commonly accepted 12 dB threshold for speech quality, and an output SNR as high as 19 dB.

This experiment shows, for a particular position of the operator, that matched filtering and GSC beamforming are sensitive to identification errors of impulse responses. The proposed algorithm corrects them and shows a large capacity in noise reduction and speech dereverberation in adverse conditions. It proves that the identification of impulse responses is necessary to achieve efficient noise reduction and speech dereverberation by adaptive beamforming and matched filtering respectively.

3.2. Sensitivity and response to identification errors from one position to another

We show next how sensitive matched filtering and GSC beamforming are to identification errors and how the algorithm responds to them with other positions of the speaker. In Fig. 3, we repeat the experience of Fig. 2 with the operator placed this time at the left position.

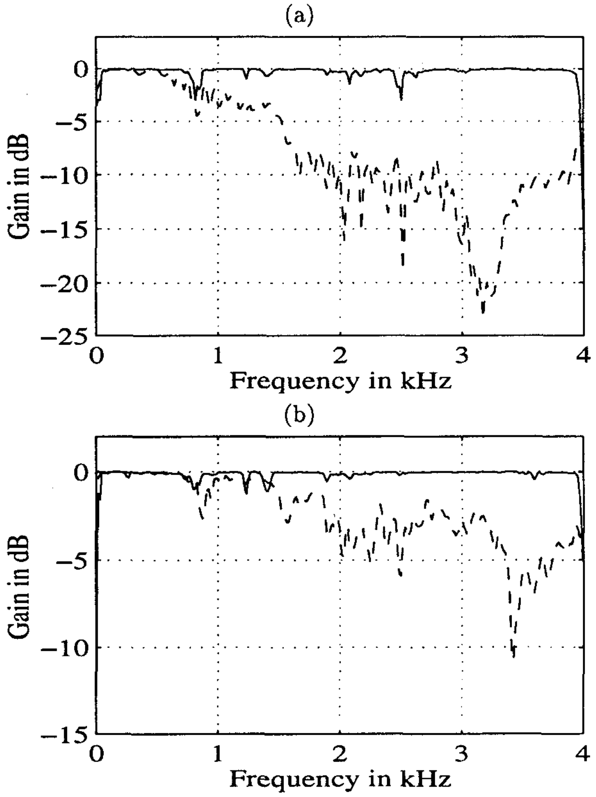

Figure 3: Gain in dB as in Fig. 2 when speech comes now from the left position. (a) Initialization with TDC from central position as in Fig. 2. (b) Initialization with the impulse responses obtained after convergence in Fig. 2.

In Fig. 3-a, we first initialize the algorithm with TDC as in Fig. 2. Without tracking, we naturally notice that identification errors of impulse responses are higher by simple TDC from central position. However, the proposed microphone array is still able to correct these errors in an efficient way. This figure shows the capacity of the algorithm to track impulse responses from different speaker locations with the same and simple initialization by TDC. No further approximations are needed to start the algorithm.

In Fig. 3-b, we secondly initialize the algorithm with the impulse responses from central position obtained after convergence in Fig. 3-a. Although identification errors without tracking are smaller, they are still significant to make speech signal cancelation effective. They illustrate the sensitivity of matched filtering and GSC beamforming to identification errors of impulse responses from one speaker position to another. However, the proposed algorithm properly corrects these errors by the subspace-based tracking procedure of impulse responses in (4).

This figure shows that the identification of impulse responses at one speaker position is insufficient, and proves that permanent tracking is necessary to properly follow speaker movements. It proves that matched filtering cannot allow for an efficient noise reduction with fixed estimates of impulse responses unless they are adjusted in time.

3.3. Tracking capacity of sudden speaker movements

We now evaluate the algorithm in the case of sudden speaker movements and show its capacity to adapt to this situation. To do so, we assess in Fig. 4 its tracking behavior for a sudden change of the speaker position from the left-side to the right-side location (see Fig. 1).

We actually initialize the tracking procedure with the impulse responses from the left-side position obtained after convergence in Fig. 3-b. Right after the movement of the speaker, we just notice a small attenuation of the speech signal until the attack of the second sentence. This short duration of speech activity is the time interval that is necessary for the tracking procedure to adapt to the sudden change in speaker position.

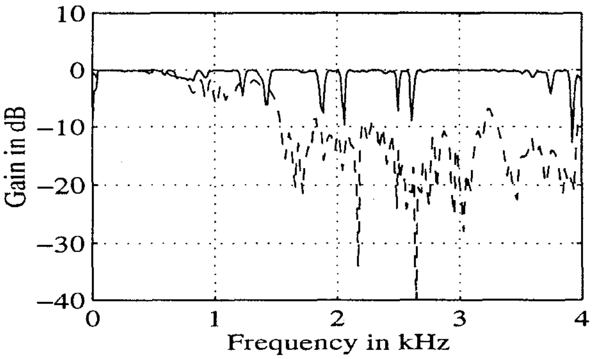

Figure 4: Gain when speaker suddenly moves from left to right, just after the movement (dashed), and after 1 s of speech activity (solid).

In Fig. 4, we plot the gain of the proposed system just after the movement of the speaker (dashed), and after 1 s of speech activity (solid). We notice that the sudden movement of the speaker from the left to the right-side position instantaneously entails large identification errors. Actually, this amounts to a new initialization of the algorithm during speech activity. We also notice that 1 s of speech activity is sufficient for convergence, although small notches at few frequencies still require a further processing time due to larger initial errors in the learning curve.

This experiment proves the tracking capacity of the algorithm to properly adapt to fast speaker movements.

4. CONCLUSION AND PERSPECTIVES

To achieve optimal performance in speech dereverberation and noise reduction by microphone arrays, the identification of impulse responses is necessary [6]-[8]. We herein showed the sensitivity of adaptive microphone arrays to identification errors of impulse responses. We hence proved that permanent tracking of impulse responses (*i.e. adjusted* matched-filtering) is also necessary to adapt to speaker movements. Contrary to previous methods [1]-[4], the microphone array proposed in [6],[7] adaptively identifies impulse responses, simultaneously dereverberates speech and efficiently reduces noise. We herein proved its capacity to respond to speaker movements and to identification errors.

A future point to address is the tracking capacity of the algorithm when the operator is in the far-field of microphones. All the experiments in this paper were indeed made in the near-field. However, recent experiments assessing a mini-teleconference mode with six microphones, all placed in the far-field at about 3 m from speakers moving in a meeting room, proved an adapted version of the algorithm to behave normally [8]. These preliminary tests already made for another application exclude specific problems due to the tracking in the far-field, but a deeper study should follow with a detailed evaluation.

5. REFERENCES

[1] J.L. Flanagan, A.C. Surendran and E.E. Jan, "Spatially selective sound capture for speech and audio processing", *Speech Communication*, vol. 13, pp. 207-222, Oct. 1993.

[2] E.E. Jan and J.L. Flanagan, "Sound capture from spatial volumes: matched-filter processing of microphone arrays having randomly distributed sensors", *in Proc. of ICASSP'96*, vol. II, 1996, pp. 917-920.

[3] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings", *in Proc. of ICASSP'90*, vol. 2, 1990, pp. 833-836.

[4] S. Nordholm, I. Claesson and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars", *IEEE Trans. Vehicular Tech.*, vol. 42, pp. 514-518, Nov. 1993.

[5] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. Antennas Propagat.*, vol. 30, pp. 27-34, Jan. 1982.

[6] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech", Submitted July 1995, to *IEEE Trans. Speech and Audio Processing*, revised July 1996, in review.

[7] S. Affes and Y. Grenier, "A source subspace tracking array of microphones for double talk situations", *in Proc. of ICASSP'96*, vol. II, 1996, pp. 909-912.

[8] S. Affes, *Adaptive Beamforming in Reverberant Environments*, PhD Thesis of ENST (in French), Ref. ENST 95 E 037, ENST, Paris, France, Oct. 1995.