A Signal Subspace Tracking Algorithm for Microphone Array Processing of Speech

Sofiène Affes, *Member, IEEE,* and Yves Grenier, *Member, IEEE*

***Abstract*— This paper presents a method of adaptive microphone array beamforming using matched filters with signal subspace tracking. Our objective is to enhance near-field speech signals by reducing multipath and reverberation. In real applications such as speech acquisition in acoustic environments, sources do not propagate along known and direct paths. Particularly in hands-free telephony, we have to deal with undesired propagation phenomena such as reflections and reverberation. Prior methods developed adaptive microphone arrays for noise reduction after a time delay compensation of the direct path. This simple synchronization is insufficient to produce an acceptable speech quality, and makes adaptive beamforming unsuitable. In this contribution, we prove the identification of source-to-array impulse responses to be possible by subspace tracking. We consequently show the advantage of treating synchronization as a matched filtering step. Speech quality is indeed enhanced at the output by the suppression of reflections and reverberation (i.e., dereverberation), and efficient adaptive beamforming for noise reduction is applied without risk of signal cancellation. Evaluations confirm the performance achieved by the proposed algorithm under real conditions.**

***Index Terms*—Adaptive beamforming, dereverberation, identification, matched filtering, microphone arrays, speech enhancement, subspace tracking, voice activity detection.**

I. Introduction

THERE IS increasing interest in speech acquisition in adverse acoustic environments with regard to voice control and hands-free telephone communications. For speech recognition controlled devices as well as for speech transmission, efficient acquisition systems need to reduce noise. But they should also suppress undesired multipath propagation phenomena such as reflections and reverberation of speech (i.e., dereverberation). Microphone arrays seem appropriate to achieve these tasks, but adjusting them to fit the sound field remains so far a major matter of investigation [1], [2]. We shall show in this contribution that the identification and the matched filtering of source-to-array impulse responses are necessary to release microphone arrays from this constraint. Upon this statement, the subspace-tracking-based algorithm we propose achieves the above requirements and outperforms previous methods.

In array processing techniques such as beamforming [3], input data is classically synchronized at sensors by a simple time delay compensation (TDC[1]) of the direct source propagation path, before applying the beamformer's coefficients for noise reduction. This preprocessing step, called *steering*, is justified by the fact that sources are usually modeled or approximated to propagate along planar or spherical waves. In real applications of speech acquisition in acoustic environments, sensors are however acoustic microphones with unknown directivity patterns. In addition, reflections and reverberation can no longer be neglected by the processing stage (i.e., beamformer). If not suppressed, they will make extracted speech sound unpleasant at the output. Besides, early reflections can be considered as coherent jammers and may cancel the speech signal in adaptive beamforming [4]. TDC becomes insufficient to fit the sound field, and noise reduction is also affected.

Many adaptive microphone arrays were proposed for speech enhancement in quite friendly acoustic environments [5]–[8]. Unfortunately, most of them turn down the first stage of steering (i.e., synchronization) and put the emphasis on noise reduction alone. In [2], we evaluated these methods for speech acquisition in cars, and precisely noticed their poor performance in noise reduction in the tested environment.

Kaneda and Ohga [5] assume the location of the speaker to be known and fixed. They measure the corresponding impulse responses (IR's), then use them to train the beamformer with recorded noise. This requires stationary conditions difficult to reach with a mobile speaker and nonstationary signals. To improve noise reduction, they allow some distortion of the desired source. Sondhi and Elko [6] adopt a similar structure but consider TDC of the direct path. To further improve noise reduction, they introduce a soft constraint on signal modulus allowing an amount of distortion. Zelinski [7] also considers TDC of the direct path. He, however, assumes the noise to be diffuse and uncorrelated, then applies a delay-sum (DS) beamformer [3] by summing the inputs after steering. To enhance noise reduction, he proposes a Wiener postfilter. Simmer *et al.* [8], [10] improve this filter and implement a unit for adaptive TDC of the direct path [9]. Gierl [11] combines TDC with multidimensional spectral subtraction.

Manuscript received July 31, 1995; revised January 13, 1997. This work follows up studies partly funded by the EEC under the European contract ESPRIT project 6166 FREETEL, Enhancement of Hands-Free Telephony. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

Y. Grenier is with the Département Signal, Ecole Nationale Supérieure des Télécommunications, 75634 Paris Cedex 13, France.

S. Affes was with the Département Signal, Ecole Nationale Supérieure des Télécommunications. He is currently with the Institut National de la Recherche Scientifique–Télécommunications, Verdun, P.Q. Canada H3E 1H6.

Publisher Item Identifier S 1063-6676(97)06387-6.

[1] In this paper, TDC is strictly used to denote time delay compensation with only single tap filters.

Although not tested in [2], and contrary to previous methods, Van Compernolle [12], [13] and Nordholm *et al.* [14] propose adaptive beamformers with a generalized sidelobe canceler (GSC) structure [15] updated during silence. Adaptive beamforming is more efficient for noise reduction, but suffers from severe speech cancellation in the presence of steering errors [4]. To further minimize this effect, Van Compernolle proposes a unit for adaptive TDC, updated during speech activity to avoid deviations to noise sources. Nordholm *et al.* assume TDC of a spread source in the near field, and introduce a linear constraint on superresolution to cover the emitting area. All the methods above propose suboptimal beamformers for noise reduction, and introduce an amount of speech cancellation or distortion depending on whether processing is adaptive or not.

To really achieve satisfactory results, we underlined in [2] and [16] our conclusion that steering should be definitely seen as a matched filtering step or an inversion of IR's rather than TDC of the direct path, and that multichannel identification of acoustic paths is necessary. We also proved in [2] the advantage of matched filtering over time TDC achievable by beamforming in terms of producing a very natural quality of speech and a higher intelligibility at the output (i.e., dereverberation). Several acoustic beamformers propose the inversion of IR's by deconvolution in the steering stage, but suffer from the fact that acoustic room impulses often are not minimum phase and not invertible [17]. Indeed, deconvolution implies that one is attempting to invert the transfer function, which is very problematic for nonminimum-phase systems. Rather, the system response is just being conjugated here, which is conventionally known as matched filtering. We hence avoid the inversion problems encountered in deconvolution. Flanagan *et al.* [18] recently applied matched-filter processing to microphone arrays and reported its dereverberation capacity. However, they used a very large number of microphones with a suboptimal DS beamforming structure for noise reduction. They also calculated fixed IR's from the room geometry or measured them in actual rooms as in [5], without addressing the tracking of nonstationary acoustic paths. In this contribution, we adaptively identify the IR's and respectively adjust the matched filters to them. We also apply a GSC beamformer for an efficient noise reduction with a small number of microphones.

This work follows up former studies referenced in this paper. After preliminary studies made in [2] and [16], we proposed in [19] a robust wideband adaptive beamformer based on source-subspace tracking of propagation vectors in an array manifold (i.e., IR identification) [20][2]. We studied the algorithm with a simple manifold of far-field sources as a particular case of a more general array characterization. With this flexible formulation, a possible adaptation to acoustic environments can be viewed. In addition, the high performance of the algorithm and its low complexity observed in that simple case offer a significant perspective for further implementation in real applications.

[2] We refer here to the underlying method in [20] as the adaptive source-subspace extraction and tracking (ASSET) technique.

In this paper, we adapt [19] to speech acquisition in a banker market trading room. In Section II, we first make an acoustic characterization of the array to possibly find the underlying features of IR's. We will notice the total energy of any frequency component to be quite constant for emitter locations around a central speaker position. From this key observation, we introduce significant constraints characterizing the array. To reliably identify IR's in Section III, we adapt the tracking procedure to the studied environment and introduce a voice activity detector for the tracking activation inspired from [9]. We also apply a GSC structure [15] for speech acquisition and noise reduction and replace its classical DS branch by matched filters. Evaluation results under real conditions, described in Section IV, show a very good quality of speech after dereverberation and an efficient noise reduction. The proposed algorithm outperforms the GSC structure combined with TDC suggested in [12] and [13]. In addition, the method is even able to cancel a strong echo emitted from a close loudspeaker without any knowledge of its reference signal. We finally give our conclusion and perspectives in Section V.

II. Acoustic Characterization and Model

In this section, we first describe the configuration then mention the drawbacks of TDC in the studied environment. We show indeed that TDC entails speech cancellation in adaptive beamforming, and a low quality of speech due to sound reflections and reverberation. Identification and matched filtering of IR's avoid these phenomena and can be implemented along the lines given below at the end of the section.

A. Configuration

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 $\times$ 20 m width $\times$ 3 m height.[3] Six microphones are linearly placed along the top edge, and six others are placed on both the left and right edges as shown in Fig. 1. 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. We can now model the signals received from the microphone array at time t as follows:

$$\mathcal{X}(t) = \mathcal{G}(t) \otimes s(t) + \mathcal{N}(t) \tag{1}$$

where $\mathcal{X}(t)$ denotes the m-dimensional observation vector and where $s(t)$ is the emitted speech signal uttered from the operator; $\mathcal{G}(t) \triangleq [g_1(t), \cdots, g_m(t)]^T$ is the m-dimensional vector of IR's, $\mathcal{N}(t)$ is the noise vector, and $\otimes$ denotes time convolution. All the quantities considered in (1) are real.

Note that all signals are wideband and nonstationary. Noise particularly contains cocktail party speech, double talk, and possibly a strong echo emitted from the loudspeaker. Although its spectral characteristics are similar to desired speech, we

[3] The room environment data was recorded by ENST and PAGE Iberica in a banker market trading room of Banesto, Madrid, Spain.

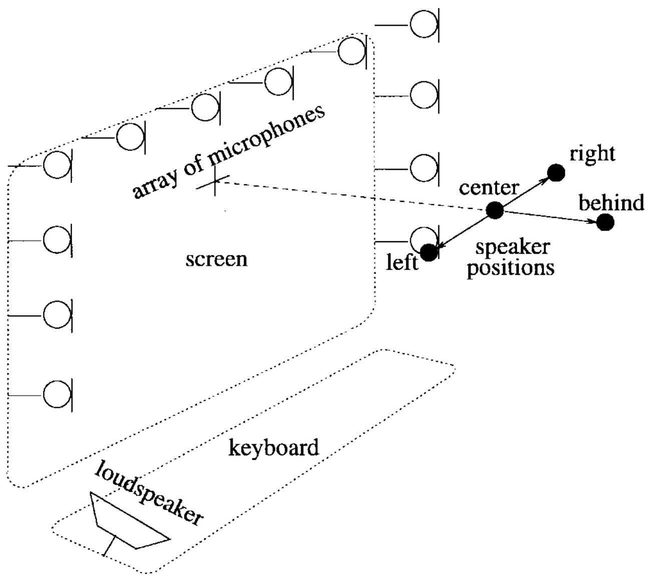


Fig. 1. Configuration of microphone array in a banker market trading room.

assume that $\mathcal{N}(t)$ and $s(t)$ are uncorrelated. Also, we do not assume a parametric model of $\mathcal{G}(t)$ characterizing the sound field. We do not neglect the mobility of the speaker, although it is assumed to be local around a central position. In fact, we reasonably assume that $\mathcal{G}(t)$ is slowly varying and locally constant in time.

To characterize acoustic features specific to the studied environment, we measure IR's over 8192 coefficients at a sampling frequency of 8 kHz, at four selected nominal positions of the speaker's mouth (center, right, left, and behind as shown in Fig. 1). The central position is located at 0.90 m, perpendicular to the array centroid. Two other positions are located on each side 0.15 m away from center, and a last position is located 0.20 m behind. To measure the IR's from each position, we send Golay codes to a loudspeaker placed at the corresponding location and record the signals from the microphones simultaneously [22]. The Golay codes are generated from a remote PC and sent to the loudspeaker through a D/A converter. The IR's are finally estimated by circular convolution of the excitation sequence with the received signals [22].

Other IR's were actually measured at different locations to the right of the operator. The positions were selected at larger variations up to a distance covering the two next operators at 4 m from central position. These IR's were measured to evaluate the room conditions. They particularly show a quite constant reverberation time over positions of around 1.7 s [21], [22], and illustrate the reverberation effect of the large trading room at various positions of the speaker.

B. TDC versus Identification/Matched Filtering of IR's

In the studied environment, TDC is unsuitable for adaptive beamforming and speech cancellation may occur, while identification and matched filtering of IR's avoid this effect. This can be confirmed from the simple observation of IR's. In Fig. 2(a), we plot the sixth IR of the central position over the first 1024 coefficients and clearly notice strong reflections and reverberation. Reflections are the early impulses reflected by large surfaces such as walls, furniture, etc. They are depicted by the segment of the curve from 10 to 16 ms. Reverberation is a complex mixture of multiply reflected and diffracted waves without a macroscopic or predictable structure. They are illustrated by the tail of the curve. Due to the presence of close and disturbing reflections, a simple synchronization over the direct path cannot be guaranteed.

Even if TDC can be properly achieved, adaptive beamforming would cancel speech from uncompensated reflections and reverberation [4]. For instance, Van Compernolle used a TDC unit similar to [9] based on cross-correlation [12]. He, however, replaced this unit by adaptive filters in [13] to improve the accuracy of time delay estimates. Nevertheless, he reported with both schemes predictable signal cancellation phenomena at a positive signal-to-noise ratio (SNR) [4]. Fig. 2(a) shows that reflections and reverberation are too strong to be approximated by simple time delays. One way to efficiently suppress reflections and reverberation is to identify IR's for matched filtering in the steering stage. Simulation later will confirm the advantage of this scheme over TDC proposed in [6]–[14].

There is another drawback of TDC in the studied environment. Reflections and reverberation of speech are simply delayed with TDC, and would be noticeable after processing in the listening. On the other hand, identification and matched filtering of IR's recovers a natural quality of speech. This can be assessed by quantitative measurements. To do so, we define the energy decay curve (EDC) [21], [22] of the ith IR $g_i(t)$ for $i = 1, \cdots, m$ as follows:

$$E_{g_i}(t) \triangleq \sum_{\tau=t}^{\infty} g_i^2(\tau). \tag{2}$$

In Fig. 2(b), the solid line plots the normalized EDC in dB of the sixth IR, which defines the amount of energy left in the response at time t. Notice that the decay slope changes abruptly at an instant $T_d = 16$ ms, called *total duration.* It corresponds to the contribution of the direct path and early reflections. At that point of the EDC, we define the clarity index in dB [21], [22] by

$$C(g_i) \triangleq 10 \log_{10}\left(\frac{E_{g_i}(0)}{E_{g_i}(T_d)}\right). \tag{3}$$

This index, which specifies the quality of an acoustic channel for speech transmission, is the ratio of the total energy of the associated IR to the energy contained in its late reverberation part. The quality of speech transmitted is considered good when this index exceeds 12 dB. The normalized curve of E_{g_6} plotted in Fig. 2(b) shows a relatively low clarity index of 9.7 dB. A consequence is that the speech picked up by microphones will not sound pleasant to the listener. The classical delay-sum (DS) beamforming cannot significantly improve this index at output after TDC of the IR's (i.e., 12.7 dB on the curve plotted as a semi-dashed line), while IR identification and matched filtering

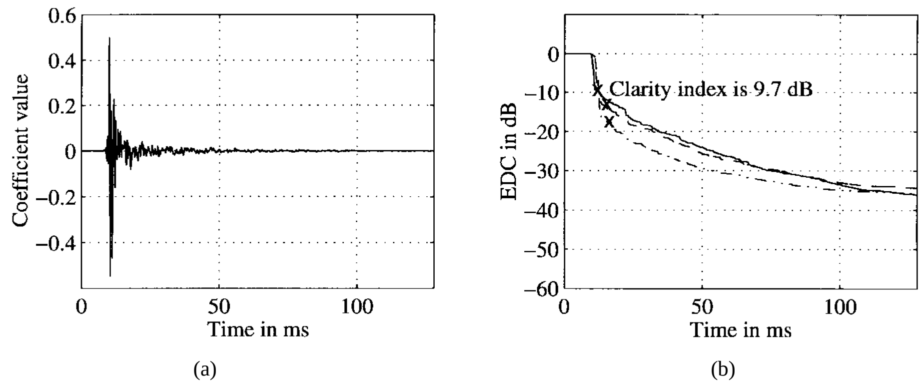


Fig. 2. IR coefficients and normalized energy decay curves for speaker position at center. (a) IR of sixth channel. (b) EDC of sixth channel (solid), of total IR with TDC and DS (dashed), and of total IR with matched filtering (semi-dashed).

over 256 coefficients offers a potential[4] clarity index of 18 dB (plotted as a dashed line). Simulations will show that the proposed algorithm reaches this index. In this case, the identification of IR's can be reasonably made over $L = 256$ coefficients.

C. Frequency Domain Identification

We identify IR's in the frequency domain. This implementation offers an attractive structure paralleling existing narrowband identification procedures for each frequency component. It requires, however, an adaptation to the studied environment.

We first take the fast Fourier transform (FFT) of (1) over $2L = 512$ snapshots each $K = 16 \leq L$ sampling periods according to the scheme of Fig. 3. For $f = 0, \cdots, 2L-1$ we have

$$X_{f,n} = G_f s_{f,n} + N_{f,n}, \tag{4}$$

where the subscripts f and n in (4) denote, respectively, the FFT of the indexed quantity in (1) at frequency bin f and the $(m \times 2L)$-block of input data, numbered as n. We previously assumed time variations of $\mathcal{G}(t)$ to be very slow and practically constant in comparison to the variations of $s(t)$ and $\mathcal{N}(t)$. We, hence, approximate $G_{f,n} \simeq G_f$ for simplicity, although it is understood that time variations can be tracked. By virtue of the Hermitian symmetry of the model, note in the following that all the processing in the frequency domain will be performed over the first $L+1$ frequency bins instead of the $2L$ available components.

Equation (4) shows that G_f and $s_{f,n}$ can be estimated only within a multiplicative factor [i.e., $(G_f/k) \times (ks_{f,n}) = G_f s_{f,n}$]. However, this ambiguity can be removed. Indeed, we show below that the modulus of G_f can be estimated *a priori*. In Fig. 4(a), we plot $|g_{6,f}|^2$ for the four selected positions of the speaker where $g_{i,f}$ is the FFT of the ith IR $g_i(t)$. The curves show relatively high variations of IR's from one position to another. On the other hand, the average curves

$$\beta_f^2 \triangleq \frac{\sum_{i=1}^{m} |g_{i,f}|^2}{m} = \frac{||G_f||^2}{m} \tag{5}$$

[4] We use 256 coefficients of each measured IR for perfect identification.

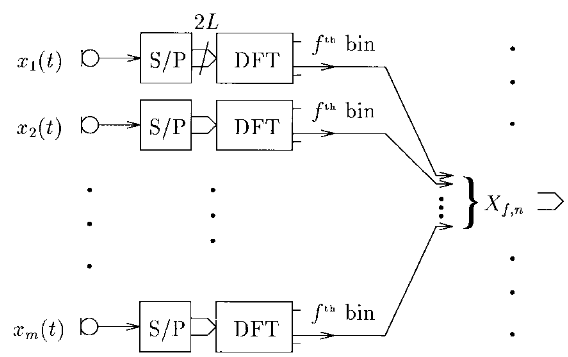


Fig. 3. Serial to parallel and transform to the frequency domain of observation signals.

plotted for the same four positions in Fig. 4(b) show small variations. Their standard deviation is actually smaller than 10% of the mean value at any frequency component. In this case, we can assume that the mean energy β_f^2 is constant for any location of the speaker around the central position. This constant can be measured as a weighted combination of the curves plotted in Fig. 4(b). For instance, we can take the average if we *a priori* assume a uniform probability distribution over the speaker positions. Actually, this observation was also confirmed in a different context of hands-free telephony in cars [1], which proves the assumption to be quite realistic for different acoustic environments. Intuitively, some kind of "local energy conservation principle" gives support to this feature, which underlines the acoustic characterization of our IR's.

Now that the problem of ambiguity is solved, we can reformulate the problem in a way that better introduces our algorithm. To do so, we rewrite (4) as follows:

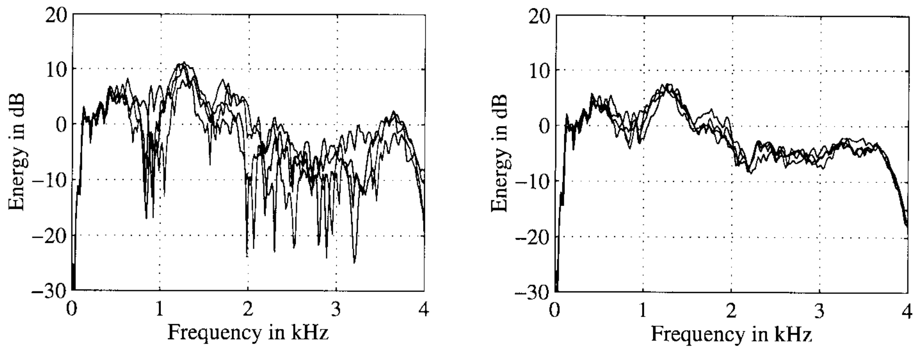
$$X_{f,n} = \alpha_{f,n} U_f + N_{f,n} \tag{6}$$

where the complex vector

$$U_f \triangleq \frac{1}{\beta_f} G_f \tag{7}$$

is the signal subspace basis vector with norm $\sqrt{m}$, and where the complex scalar

$$\alpha_{f,n} \triangleq \beta_f s_{f,n} \tag{8}$$



(a) (b)

Fig. 4. Energy characterization of IR's. (a) Energy of sixth IR for four positions. (b) Mean energy of IR's for four positions.

is the signal parameter. Note here that $||U_f||^2 = U_f^H U_f = m$, and that m will be used for normalization in the following. If it is possible to track the signal subspace properly, the idea is to recover the signal parameter $\alpha_{f,n}$ and consequently estimate $s_{f,n}$ by an adequate distortionless beamformer V_f (i.e., $V_f^H U_f = 1, \hat{\alpha}_{f,n} = V_f^H X_{f,n}$). For instance, the matched filtering beamformer $V_f = U_f/m$, which has the simple structure of DS, conjugates the propagation vector U_f or equivalently the IR's regardless of the noise structure and optimally reduces uncorrelated white noise. We shall show in the next section how to combine it with a GSC structure to efficiently reduce colored noise, but first the propagation vectors U_f have to be identified.

The system identification problem in (6) is commonly studied in the narrowband case by localization methods in the electromagnetic far field or near field. Eigensubspace-based algorithms particularly estimate the location parameter or equivalently U_f corresponding to the propagation along the direct path. However, they often assume the wavefront to be planar or spherical and the noise to be white and uncorrelated (see references in [24]). These assumptions are unrealistic in the studied context. On the other hand, we successfully derived in [19] and [20] a source subspace tracking procedure of U_f in an array manifold in general, and tested its efficiency for speech acquisition with real data. Using this technique in audio acoustics, we shall show in the next section how to avoid sound field modeling when identifying U_f by subspace tracking.

III. The Proposed Algorithm

We describe in this section the different steps of the algorithm. We first explain the adaptive GSC structure when adapted to the matched filtering of identified IR's in the steering stage. We secondly introduce the IR identification procedure, relate it to existing techniques, then prove its convergence. We show that its performance is enhanced when estimated propagation vectors are constrained to fit with *a priori* acoustic features. It is also improved by a voice activity detector blocking the identification procedure during silence. Finally, we briefly explain speech reconstruction.

A. Matched Filtering and GSC Beamforming

With identified IR's, we can combine matched filtering with adaptive beamforming for both optimal speech acquisition and noise reduction without speech cancellation. Let us assume that an estimation of the signal subspace basis U_f at iteration n, say $\hat{U}_{f,n}$, is available and near convergence. We can immediately estimate $\alpha_{f,n}$ using the matched filtering beamformer described earlier by $\hat{\alpha}_{f,n} = \hat{U}_{f,n}^H X_{f,n}/m$. This step, which has the structure of a classical DS beamformer, amounts to replacing TDC by matched filtering, where the usual steering vector of simple time delays is replaced by $\hat{U}_{f,n}$. Contrary to TDC in [6]–[14], the matched filtering compensates speech distortion due to multipath propagation by conjugating the IR's. However, its output denoted in the following by $\hat{y}_{f,n}$ is not optimal unless the noise is uncorrelated and diffuse. To better estimate the signal parameter $\alpha_{f,n}$, unlike [18], we further reduce the residual noise still present in $\hat{y}_{f,n}$ from the noise references defined in the noise subspace orthogonal to $\hat{U}_{f,n}$. The identification of U_f provides noise references free from speech leaks. This prevents speech cancellation.

As shown in Fig. 5, we use a GSC structure [15] for $f = 0, \cdots, L$ as follows:

$$\begin{aligned}\hat{y}_{f,n} &\triangleq \frac{\hat{U}_{f,n}^H X_{f,n}}{m}\\ Z_{f,n} &\triangleq P_{f,n}^H X_{f,n}\\ \hat{\alpha}_{f,n} &= \hat{y}_{f,n} - W_{f,n}^H Z_{f,n}\\ W_{f,n+1} &= W_{f,n} + \eta_{f,n} Z_{f,n} \hat{\alpha}_{f,n}^H\end{aligned} \tag{9}$$

where $P_{f,n}$ is a $m \times (m-1)$ signal blocking matrix projecting $X_{f,n}$ on the noise subspace orthogonal to $\hat{U}_{f,n}$ to obtain $Z_{f,n}$ [15]; the superscript H denotes conjugate transpose, and $\eta_{f,n}$ is the stepsize of the GSC, possibly including a normalization factor (see [26] for more details, e.g., $\eta_{f,n} \triangleq \eta_0/||Z_{f,n}||^2$). The GSC filter $W_{f,n}$ is an $(m-1)$-dimensional vector initially set to zero and implemented in a least mean squares (LMS) structure [26]. To show the advantage of the algorithm over previous methods, we start the algorithm with

$$\hat{U}_{f,0} = \left[e^{-j2\pi\hat{\tau}_1(f/2L)}, \cdots, e^{-j2\pi\hat{\tau}_m(f/2L)}\right]^T \tag{10}$$

where $\hat{\tau}_i (i = 1, \cdots, m)$ are time delay estimates of the direct path, as made in [6]–[14].

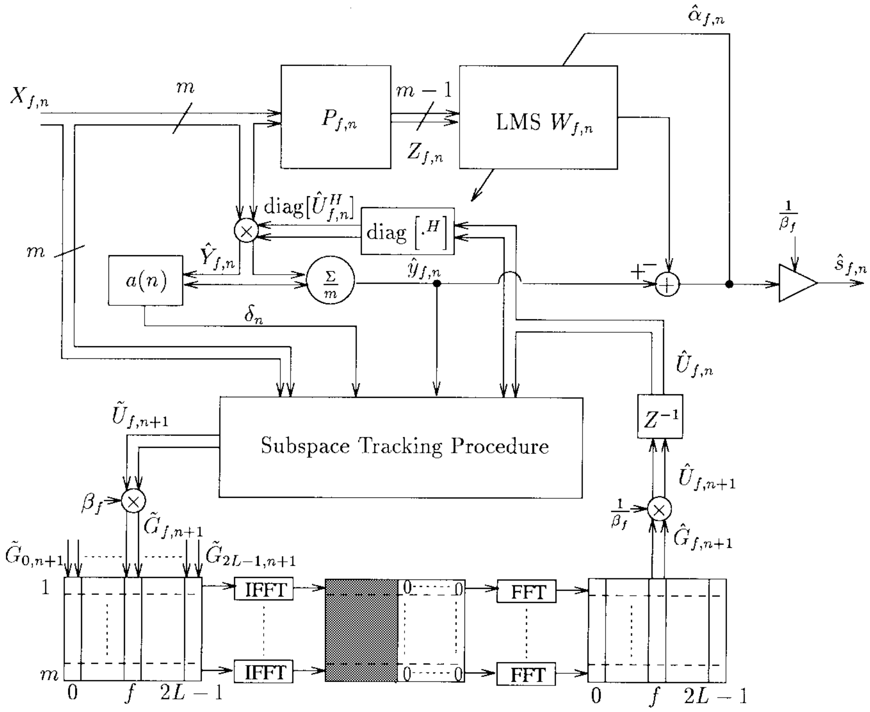


Fig. 5. Block diagram of the algorithm at frequency f.

B. Channel Identification

The input-output IR identification scheme we propose relies on a general framework of subspace tracking and structure forcing of propagation vectors. It can be related to existing techniques, and its convergence can be guaranteed in the studied environment.

In the same way as in [19], we correct $\hat{U}_{f,n}$ and track the basis vectors U_f at each frequency bin f. With vectors $X_{f,n}$ of input data and any estimated output sequence of the signal parameter $\hat{\alpha}_{f,n} = V_{f,n}^H X_{f,n}$, where $V_{f,n}$ is a given distortionless beamformer (i.e., $V_{f,n}^H \hat{U}_{f,n} = 1$), we apply the following general equation for identification:

$$\tilde{U}_{f,n+1} = \hat{U}_{f,n} + \mu_{f,n}(X_{f,n} - \hat{U}_{f,n}\hat{\alpha}_{f,n})\hat{\alpha}_{f,n}^H \tag{11}$$

where $\mu_{f,n}$ is the stepsize of the LMS-like tracking equation (11), possibly including a normalization factor. Note that (11) indeed corresponds to a gradient-type solution of an identification problem if $\hat{\alpha}_{f,n}$ is a known reference sequence of speech signal parameter [26]. Contrary to $\hat{U}_{f,n}$, the unconstrained estimate of $U_{f,n+1}$ is denoted at present by $\tilde{U}_{f,n+1}$ in (11). We will show in the next subsection how to constrain it with respect to an acoustic characterization of the IR's to have $\hat{U}_{f,n+1}$.

Actually, the gradient-type equation (11) derives from minimization of the cost function $E[\|(I_m - U_{f,n}V_{f,n}^H)X_{f,n}\|^2]$ where $U_{f,n}$ is assigned to check some acoustic features (e.g., lying in an array manifold if it exists), and where the beamformer $V_{f,n}$ is defined such that $V_{f,n}^H U_{f,n} = 1$. In a recent work [20], [25], we generalized this criterion to a multisource tracking equation (i.e., $U_{f,n}$ is a matrix whose columns are "structure-fitted" propagation vectors and $V_{f,n}^H U_{f,n} = I$). This criterion can be related to other methods referenced in [24], but contrary to them, it proposes a direct estimation of propagation vectors by simultaneous subspace tracking and structure fitting. For instance, if we select the DS beamformer $V_{f,n} = U_{f,n}/m$ in the one-dimensional (1-D) case, we obtain the simplified neuron model proposed by Oja [23] as a principal component analyzer. It minimizes the cost function $E[\|(I_m - U_{f,n}U_{f,n}^H/m)X_{f,n}\|^2]$, but converges to the eigenvector with the highest eigenvalue. In [24], Yang generalized this criterion to the case of a multidimensional signal subspace (i.e., $U_{f,n}$ is a matrix) and applied it to image processing. Although $U_{f,n}$ is proved to converge to the orthonormal eigensubspace basis corresponding to the highest eigenvalues, its column vectors do not correspond to propagation vectors.

In (11), we can use the GSC beamformer output $\hat{\alpha}_{f,n}$ estimated in the previous subsection with $V_{f,n} = \hat{U}_{f,n}/m - P_{f,n}W_{f,n}$. However, we observed that the tracking procedure is less stable and slower when applied to nonstationary signals. This is due to the perturbations and the additional convergence time of the side structure implemented by $P_{f,n}$ and $W_{f,n}$. Instead, we use the DS output in (9) as follows (see Fig. 5):

$$\tilde{U}_{f,n+1} = \hat{U}_{f,n} + \mu_{f,n}(X_{f,n} - \hat{U}_{f,n}\hat{y}_{f,n})\hat{y}_{f,n}^H. \tag{12}$$

In [24], Yang particularly proves that (12) converges to $\tilde{U}_{f,\infty}$ with norm $\sqrt{m}$, the basis vector of the 1-D signal subspace with the highest energy. At a reasonable SNR, when desired speech is the loudest among present sources as assumed in [13], this equation reasonably converges to any solution of the form $\tilde{U}_{f,\infty} \simeq e^{j\phi_f}U_f$ where ϕ_f is a phase shift. Although the human auditory system is not very sensitive to phase distortion [6], our experience is that $\phi_f \simeq 2\pi\tau_\infty f/(2L)$

Select $\hat{U}_{f,0} = [e^{-j2\pi\hat{\tau}_1 \frac{f}{2L}}, \cdots, e^{-j2\pi\hat{\tau}_m \frac{f}{2L}}]^T$, $a(0) = 0$, $W_{f,0} = 0_{m-1}$.

For $n = 1, 2, \cdots$ do:

$$\hat{Y}_{f,n} = \mathrm{diag}[\hat{U}_{f,n}^H] X_{f,n} \ ,$$
$$a(n+1) = (1-\gamma)a(n) + \gamma \ \frac{\sum_{f \in \Phi}\{\frac{2}{m(m-1)}\sum_{j=1}^{m-1}\sum_{k=j+1}^{m} \mathrm{Re}(\hat{Y}_{j,f,n}\hat{Y}_{k,f,n}^H)\}}{\sum_{f\in\Phi}\left|\frac{1}{m}\sum_{l=1}^{m}\hat{Y}_{l,f,n}\right|^2} \ ,$$
$$a(n) \rightarrow \delta_n \rightarrow \bar{\mu}_{f,n} \ ,$$

$$\hat{y}_{f,n} = \frac{\hat{U}_{f,n}^H \ X_{f,n}}{m} \ ,$$
$$Z_{f,n} = P_{f,n}^H \ X_{f,n} \ ,$$
$$\hat{\alpha}_{f,n} = y_{f,n} - W_{f,n}^H \ Z_{f,n} \ ,$$
$$W_{f,n+1} = W_{f,n} + \eta_{f,n} \ Z_{f,n} \ \hat{\alpha}_{f,n}^H \ ,$$

$$\tilde{U}_{f,n+1} = \hat{U}_{f,n} + \bar{\mu}_{f,n}(X_{f,n} - \hat{U}_{f,n}\hat{y}_{f,n}) \ \hat{y}_{f,n}^H \ ,$$

$$\tilde{U}_{f,n+1} \rightarrow \text{linear convolution constraint} \rightarrow \hat{U}_{f,n+1} \ ,$$

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

Fig. 6. Proposed algorithm for speech subspace tracking, matched filter beamforming, and GSC noise reduction.

is close to a linear phase where τ_∞ is a short time delay. The delay τ_∞ is actually positive and corresponds to a causal delay. Hence, the effect of ϕ_f on speech quality is not significant and the IR's are properly identified.[5]

C. Channel Characterization

We now show how to incorporate acoustic features to guarantee convergence even at low SNR's. In the previous subsection, we separately estimated normalized propagation vectors at each frequency, regardless of the fact that they are related to estimated IR's within a multiplicative factor β_f. In addition, the underlying fast convolution in the frequency domain between these IR estimates and speech should be constrained to be linear due to the block processing scheme [27]. This constraint implies setting a part of each IR to zero in the time domain. In this case, we should fit the estimated propagation vectors to a particular structure of IR's as shown in Fig. 5.

To do so, we incorporate the *a priori* information obtained in Section II-C stating that the mean energy of IR's at each frequency is constant and equal to β_f^2. We actually form the matrix $\tilde{G}_{n+1} \triangleq [\beta_0\tilde{U}_{0,n+1}, \cdots, \beta_L\tilde{U}_{L,n+1}, \cdots, \beta_{2L-1}\tilde{U}_{2L-1,n+1}]$, which approximates the row-by-row FFT of the unconstrained IR estimates. To apply the linear convolution constraint, we compute the matrix $\tilde{\mathcal{G}}_{n+1}$ of unconstrained IR estimates in the time domain as the row-by-row inverse FFT (IFFT) of $\tilde{G}_{n+1}$. Then, we set its $m \times L$ right half part to zero to have constrained IR estimates $\hat{\mathcal{G}}_{n+1}$ in the time domain. It is this step that guarantees the linear convolution constraint. We again take the row-by-row FFT of $\hat{\mathcal{G}}_{n+1}$ to estimate the constrained IR estimates $\hat{G}_{n+1}$ in the frequency domain. We finally have $\hat{U}_{f,n+1}$ for $f = 0, 1, \cdots, L$ from the first $L+1$ column vectors of $\hat{G}_{n+1}$. More details can be found in [27]–[29] about constrained adaptive filtering in the frequency domain and fast linear convolution.

This characterization is likely to limit any deviation of the tracking procedure from the true IR's, even at reasonably low SNR's and when desired speech is not the loudest. It seems difficult to provide theoretical arguments to support this intuitive expectation, but simulations do confirm that the linear convolution constraint improves convergence in highly adverse conditions. However, this constraint can be omitted under better conditions to save computation.

D. Speech Activity Detection

When the SNR is very low, particularly during periods of silence, (12) is likely to track noise sources. It would be better then to stop the adaptation of the algorithm so as to keep the estimates of U_f from being attracted in the noise subspace. To do so, we first define the steered input signals by $\hat{Y}_{f,n} \triangleq \mathrm{diag}[\hat{U}_{f,n}^H]X_{f,n}$ where $\mathrm{diag}[A]$ is a diagonal matrix with the elements of vector A on the main diagonal. Note here that $\hat{Y}_{f,n}$ yields the matched filtering output $\hat{y}_{f,n}$ of (9) when its elements are averaged. This is a preliminary step that guarantees the selectivity of the speech activity detection in the direction of the operator by spatial filtering. We then introduce a modified version of the voice activity detector presented in [9], as follows:

$$a(n+1) = (1-\gamma)a(n) + \gamma \cdot \frac{\sum_{f\in\Phi}\left\{\frac{2}{m(m-1)}\sum_{j=1}^{m-1}\sum_{k=j+1}^{m}\mathrm{Re}(\hat{Y}_{j,f,n}\hat{Y}_{k,f,n}^H)\right\}}{\sum_{f\in\Phi}\left|\frac{1}{m}\sum_{l=1}^{m}\hat{Y}_{l,f,n}\right|^2} \quad (13)$$

[5] It could be advantageous to extract the speaker position from the estimated IR's after convergence as required for camera pointing in some teleconference applications.

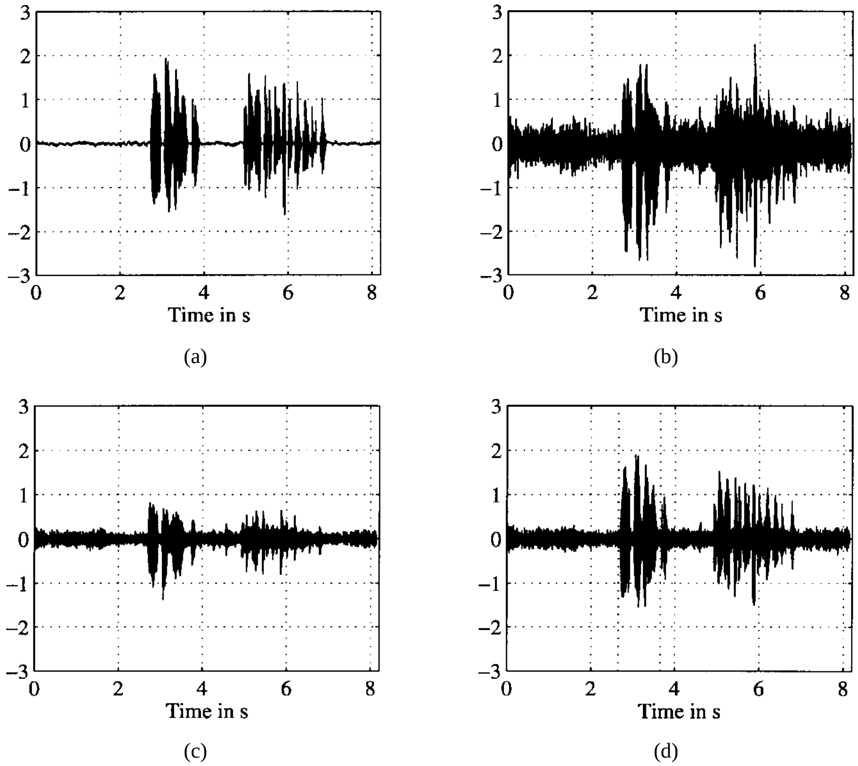


Fig. 7. Speech signal at different stages. (a) Original speech. (b) Speech received at sixth microphone. (c) Speech estimated without tracking. (d) Speech estimated with tracking.

where speech activity $a(n)$ is given by a smoothed ratio of the sum of the cross-spectrum components at a selected set of frequencies Φ, over the sum of the autospectrum components at the same frequencies; γ is a smoothing factor, $\mathrm{Re}(.)$ denotes the real part of a complex number, and $\hat{Y}_{j,f,n}$ is the jth component of $\hat{Y}_{f,n}$. We found it also better in (13) to select ten frequencies around 1.5 kHz and 2.8 kHz rather than defining $\Phi = \{0, 1, \cdots, L/2\}$ as proposed in [9] (i.e., the low frequency region going up to 2 kHz). We noted indeed that speech activity can be better discriminated from noise in these frequency regions. To test the presence of speech or silence, speech activity $a(n)$ is simply compared to a given threshold $a_{\min}$ as follows:

$$\delta_n \triangleq \begin{cases} 1, & \text{if } a(n) \geq a_{\min}, \\ 0, & \text{otherwise (silence).} \end{cases} \tag{14}$$

We then replace the stepsize of the tracking equation in (12) by $\overline{\mu}_{f,n} \triangleq \delta_n \mu_{f,n}$ to block adaptation during silence as shown in Fig. 5.

It should be noted here that the GSC structure of (9) is not blocked, contrary to [12] and [13]. The continuous processing of the GSC, which provides an efficient noise reduction even during speech activity, is now possible because we discarded the risk of signal cancellation. Notice also that δ_n simultaneously rules the adaptation of (12) at any frequency f, though it can be split into multiple control regions of speech activity over frequency sets other than Φ. Finally, we should recall that speech activity is observed in both frequency and space. The analysis of the frequency content alone would detect all speechlike signals, but the spatial selectivity through the steered inputs $\hat{Y}_{f,n}$ mentioned earlier restricts them to the speech uttered only from the desired operator. The acoustic characterization of IR's described in the previous subsection maintains this spatial selectivity even at low SNR's. This prevents the voice activity detector from responding to undesired speech signals.

E. Signal Recovery and Synthesis

Using the relation $\hat{s}_{f,n} = \hat{\alpha}_{f,n}/\beta_{f,n}$, we now recover the speech signal at the block $n+1$ in an overlap-save (OLS) [27] analysis/synthesis scheme by

$$\begin{aligned} &[\hat{s}(K(n+1)), \cdots, \hat{s}(K(n+1)+2L-1)] \\ &\quad \triangleq \mathrm{Re}\left\{ \mathrm{IFFT}\left(\left[\frac{\hat{\alpha}_{0,n}}{\beta_0}, \cdots, \frac{\hat{\alpha}_{2L-1,n}}{\beta_{2L-1}} \right] \right) \right\}. \end{aligned} \tag{15}$$

With blocks shifted each $K < L$ samples, input data is oversampled at a rate higher than required to update (12) more frequently. This is shown [28], [29] to improve the tracking performance of the algorithm. As blocks overlap over $2L - K$ samples, we only keep the following segment of length K:

$$[\hat{s}(K(n+1)+L), \cdots, \hat{s}(K(n+1)+L+K-1)].$$

We finally summarize the different steps of the algorithm presented in the previous subsections in Fig. 6.

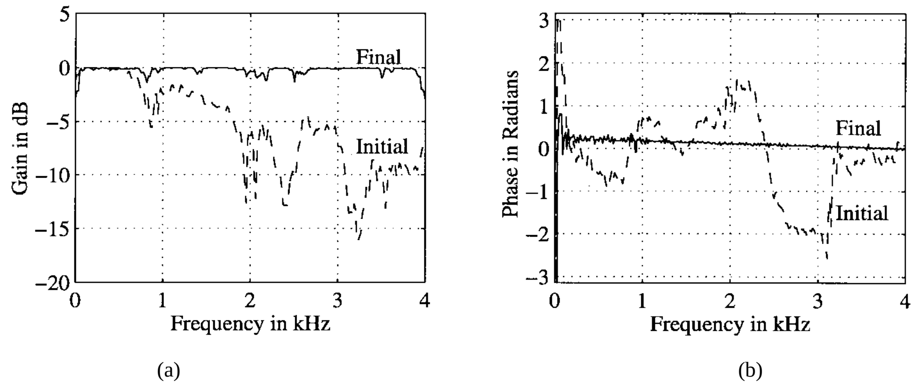


Fig. 8. Total response of proposed system $V_{f,n}^H G_{f,n}/\beta_{f,n} = (\hat{U}_{f,n}/m - P_{f,n}W_{f,n})^H G_{f,n}/\beta_{f,n}$ at $t = 2.7$ s (i.e., initial) and $t = 3.7$ s (i.e., final). (a) Gain of proposed system in dB. (b) Phase of proposed system in radians.

IV. Evaluation and Perspectives

In this section, we assess the performance of the studied algorithm for speech acquisition and noise reduction. We first want to compare it to prior methods based on simple TDC. For this reason, we start the proposed scheme with (10) as stated in Section III, although other experiments following below successfully test other initializations. We also want to evaluate the proposed method and its tracking behavior with quantitative measurements. To do so, we shall need to synthesize simulated data so as to access these measurements. Later, we resume our evaluation with experiments under real conditions before we draw out our perspectives.

A. Experiments

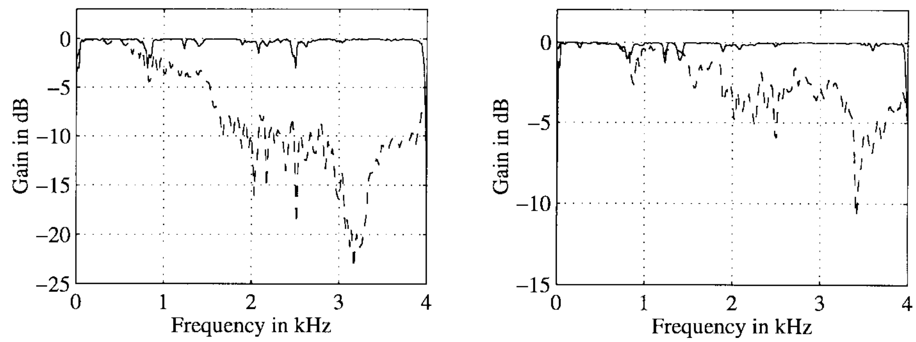
We take special care to make our first set of experiments with simulated data very close to reality. Indeed, we record a clean signal of two speech sentences uttered from a female speaker in an anechoic room to simulate the original speech of the operator. We then convolve the original waveform plotted in Fig. 7(a) with the IR's measured from the nominal central position of the speaker to the array of microphones (see Fig. 1, Section II-A). 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 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., and makes the experiment very close to reality. In Fig. 7(b), we plot one of the synthesized signals simulating the noisy speech received at the sixth microphone.

To make our comparison, we first skip the tracking step illustrated by (12) (i.e., $\mu_{f,n} = 0$). This amounts to the simple TDC usually employed [6]–[14]. In this case, we clearly observe in Fig. 7(c) the cancellation of speech signal as reported in [12] and [13]. On the other hand, the proposed algorithm avoids this phenomena as shown in Fig. 7(d), and proves the efficiency of the subspace tracking procedure of (12). Desired speech is properly recovered with a satisfying noise reduction. In Fig. 8(a), we plot the gain of the total response from the central position of the speaker to the processor output (i.e., $|V_{f,n}^H G_{f,n}/\beta_{f,n}|^2 = |(\hat{U}_{f,n}/m - P_{f,n}W_{f,n})^H G_{f,n}/\beta_{f,n}|^2$). The initial curve corresponds to TDC, and shows the usual approximation [6]–[14] to be inadequate beyond a small low-frequency region. The final curve corresponds to the identified IR's after convergence of (12) 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. 8(a) and Fig. 8(b), respectively, the audible quality of the output speech sounds very natural while point jammers are significantly reduced. This experiment shows a large capacity of the algorithm in speech dereverberation and noise reduction in adverse conditions.

To provide quantitative measurements, we compute the clarity index and the SNR at the output. We actually measure at the output the potential clarity index of 18 dB given in Section II-B. It is higher than the commonly accepted 12 dB threshold for speech quality. The SNR is empirically[6] computed as the ratio $(E_{s+n} - E_n)/E_n$, where the mean energies E_{s+n} and E_n are computed from the output signal during speech activity and silence, respectively. This does not take account a speech quality enhancement of 8 dB in clarity due to reduction of reflections and reverberation. The measured SNR gain of approximately 7 dB is less than the optimal 10.8 dB reduction of spatially diffuse noise (i.e., $10\log_{10}(m) \simeq 10.8$). To further improve the SNR gain performance, we propose a postprocessing stage of the residual noise as suggested in [6]–[8], [10], and [11]. We use, however, a spectral subtraction method developed by Ephraim and Malah [30], and measure an additional gain of 5 dB at an output SNR as high as 19 dB.

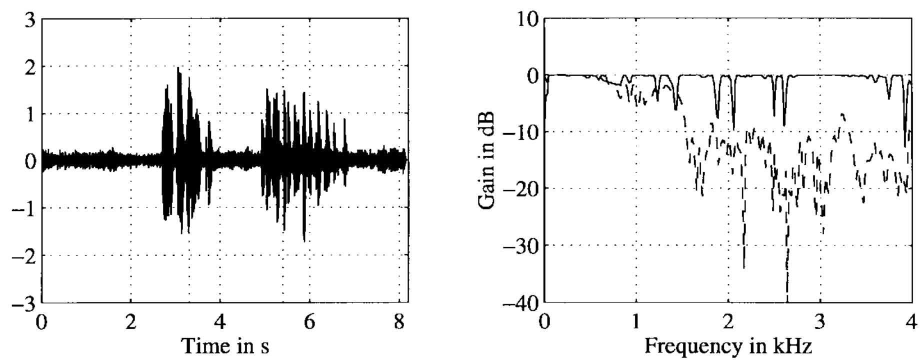
This experiment shows, for a particular configuration, that matched filtering and GSC beamforming are sensitive to identification errors of IR's. The proposed method corrects them. We show next how sensitive they are to these errors and how the algorithm responds to them with other positions of the speaker and other initializations. In Fig. 9, we repeat the experiment with the speaker placed this time at the left-side position. In Fig. 9(a), we first initialize the algorithm with (10) as in Fig. 8. Without tracking, we naturally notice

[6]We used an evaluation tool provided by the Enhancement of Hands-Free Telephony (FREETEL) project to make comparison with former results.



(a) (b)

Fig. 9. Gain in dB of proposed system as in Fig. 4(a) when speech comes now from the left side-position, initial (dashed), final (solid). (a) Initialization with TDC from central position as in Fig. 4(a). (b) Initialization with the IR's obtained after convergence in Fig. 4(a).



(a) (b)

Fig. 10. Tracking behavior of proposed system when the speaker position suddenly changes from the left to the right side position at $t = 3.3$ s. (a) Output speech. (b) Gain in dB, just after movement at $t = 3.3$ s (dashed), and after 1 s of speech activity at $t = 5.4$ s (solid).

that identification errors of IR's are higher than those from simple TDC from the central position. However, the proposed method is still able to correct them in an efficient way. This figure shows the capacity of the algorithm to track IR's from different speaker positions with the same initialization by simple TDC in (10). In Fig. 9(b), we secondly initialize the algorithm with the IR's from central position obtained after convergence in Fig. 8(a). Although identification errors without tracking are smaller, they are still significant to make speech signal cancellation effective as in Fig. 7(c). They illustrate the sensitivity of matched filtering and GSC beamforming to identification errors of IR's from one speaker position to another. On the other hand, the proposed algorithm properly corrects these errors by the subspace-based tracking procedure. This figure shows that the identification of IR's for one speaker position is insufficient, and proves that permanent tracking is necessary to properly follow speaker movements.

We now extend the evaluation of the algorithm to the case of speaker movements and show its capacity to adapt to this situation. To do so, we assess in Fig. 10 its tracking behavior for a sudden change of the speaker position from the left-side to the right-side location (see Fig. 1), in the middle of the first sentence at $t = 3.3$ s. We actually initialize the tracking procedure with the IR's from the left-side position obtained after convergence in Fig. 9. When compared to Fig. 7(a) and Fig. 7(d), the output speech of Fig 10(a) shows the algorithm to behave as well in speech enhancement. After the movement of the speaker at $t = 3.3$ s, 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. In Fig. 10(b), we plot the gain of the proposed system just after the movement of the speaker at $t = 3.3$ s, and after 1 s of speech activity at $t = 5.4$ s. We note that the sudden movement of the speaker from the left to the right-side position instantaneously entails large identification errors. This amounts to a new initialization of the algorithm during speech activity. We also note 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.

Following the previous assessments with simulated data, we now test the algorithm with data completely recorded under real conditions. Cooperative operators sitting at the experimental work desk are asked to utter two sentences. The recordings are all made at work time in the banker market trading room. Since the preliminary results we previously obtained are very satisfying, we use fewer microphones to reduce the cost of the system. We actually keep the six array microphones located at the top edge of the workstation screen for this part of the evaluation with real data. Four tests are run with sentences uttered from both male and female speakers at average SNR's ranging from 0 to 8 dB. The recorded input signals are qualitatively quite similar to the simulated data and do confirm the artificially reproduced conditions of the previous experiments to be very close to reality. These signals, after processing, are again qualitatively similar to the output speech of the previous experiments and show the quantitative measurements of speech enhancement with real data to be in the same range. Indeed, the quality of both the output speech and residual noise still sounds good and natural in terms of speech dereverberation and noise reduction. A significant improvement is evident when compared to the results of [16]. The total gain in SNR ranges from 9 to 12 dB after postprocessing and confirms the efficiency of the proposed method under real conditions.

Other tests proved the algorithm to be able to cancel even a strong echo emitted from a close loudspeaker, without any knowledge of its reference signal and without any degradation to the output speech. The echo is louder than the desired speech, but convergence is not affected. This confirms the efficiency of the linear convolution constraint over IR's and shows the proper functioning of the voice activity detector. The underlying issue of speech enhancement and echo cancellation in double talk situations is addressed in more detail in [32], where an efficient generalization is given.

B. Discussion

The evaluation results show the capacity of the algorithm to enhance near-field speech of a moving speaker in a very practical situation. They prove its efficiency in dereverberation and noise reduction in large rooms under adverse conditions. However, several issues and possible improvements are still left to be discussed for future investigations.

A first question of a practical order is related to the "portability" of the acoustic characterization when the array is moved from one workstation (i.e., work position) to another. So far, the constant energy assumption of β_f has been validated for local variations of the speaker location in the same work position.[7] One either need to precisely measure β_f at each work position or approximate it by a global and optimized measure with some relative errors minimized over each position. Note, however, that all the steps of the algorithm, except the speech recovery and synthesis in (15), are not affected by such errors over β_f. The optional linear convolution constraint may only lose some of its efficiency without seriously degrading the performance in speech dereverberation and noise reduction. In the worst case, we shall notice a small and negligible spectral shaping effect on output speech.

In the studied context of hands-free telephony in a banker market trading room, we could improve the performance of speech dereverberation and noise reduction without a significant cost increase in equipment. Indeed, we could increase the array dimension with the same number of microphones at each workstation, by cross-feeding to the array processor of each work position the microphone inputs of the neighboring workstations. The selection of the neighboring microphones would depend in general on their directivity and their positioning in the trading room.

A general point to address beyond the above generalization 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 of the array. 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 the algorithm to behave normally. These preliminary tests made for a future application excluded specific problems due to the tracking in the far field. A deeper study should follow with a detailed evaluation.

Another issue to discuss is the undesirable spatial selectivity that the large cross-connected arrays proposed above may emphasize in the direction of close jammers. This is again related to the "portability" of the acoustic characterization when using these arrays. In this situation, it is unpractical to measure β_f at each workstation from all the remote microphones of the array, while any approximation with a global measure could involve larger errors. The efficiency of the linear convolution constraint can no longer be guaranteed in this case. Consequently, the convergence to the IR's from the desired speaker could be noticeably disturbed by close jammers. Indeed, one or more neighboring operators can now be present in the near field of a remote subset of microphones, while the desired operator is in their far field. This may disadvantage the acquisition of the desired operator in favor of neighboring operators.

One potential solution to this problem we would like to investigate in the future could be based on subspace tracking with a subarray acoustic characterization. In [31], we proposed a partially blind beamformer based on subspace-tracking and a partial characterization of propagation vectors in a subarray manifold. In some applications in the electromagnetic field, the propagation paths could be unmodeled and unknown from the desired source to a subset of sensors, so that the corresponding subarray inputs might not be exploitable. However, forcing the complementary part of the modeled propagation paths to lie in their subarray manifold is shown to fully identify propagation vectors in [31]. The question to address in the future is whether using this structure with microphone arrays would guarantee the convergence in a similar way. In such a case, one should, for instance, restrict the measurement of β_f and the linear convolution constraint over the subset of IR's from the operator to the microphones of its workstation (i.e., subarray acoustic characterization). Possible spectral shaping effects on output speech may be noticed with this structure. However, the potential enhancement in speech dereverberation and noise reduction that large arrays could achieve motivates our future investigations in this direction.

Finally, the algorithm we proposed for hands-free telephony in a banker market trading room leaves out several perspec-

[7] No experiments in the FREETEL project were planned in advance for the proposed method, which was developed later after the recordings were made.

tives regarding its implementation for different applications in other acoustic environments.

V. Conclusion

In this contribution, we proved the identification and matched filtering of IR's to be possible and more advantageous than simple time delay compensation in terms of speech acquisition (i.e., dereverberation) and noise reduction. With respect to this conclusion, the algorithm we developed outperforms previous techniques based on simple synchronization of the direct propagation path. It avoids speech distortion and cancellation, recovers a natural quality of speech, and efficiently reduces noise.

In an acoustic characterization of the environment, we first noted that the total energy of IR's from any location of the speaker close to a nominal central position to be quite constant at any frequency component. From this key observation, we adapted from previous works a signal subspace tracking procedure of propagation vectors to identify IR's in the frequency domain. Propagation vectors are simultaneously constrained to agree with *a priori* acoustic features by structure forcing. This improves the performance of the algorithm. The matched filtering of IR's instead of time delaying in steering avoids speech cancellation when applying adaptive beamforming for optimal speech acquisition and noise reduction.

Among the perspectives we outlined previously, we are at present planning to incorporate the proposed microphone array in a full hands-free telephone system. This system should explicitly use the reference signal provided by the loudspeaker to improve echo cancellation. Techniques developed in [28] and [29] can be combined with the proposed scheme. Now this point is mostly addressed in [32], where an efficient solution is given for double talk situations. This system should also handle a mini-teleconference mode, where not only one but many speakers are free to move around in a room in either the near field or the far field of the array. Although some issues are still under investigation, the first experimental results we obtained are very encouraging.

Acknowledgment

The authors acknowledge all the partners of ENST in the FREETEL project, who authorized them to use their data base. They also thank Dr. O. Cappé for providing a spectral subtraction tool for postprocessing, Dr. D. Morgan, who coordinated the review, and the anonymous reviewers for useful comments on an earlier version of this paper.

References

[1] S. Affes, "Adaptive beamforming in reverberant environments," Ph.D. dissertation, Ref. ENST 95 E 037, ENST, Paris, France, Oct. 1995.

[2] S. Affes and Y. Grenier, "Test of adaptive beamformers for speech acquisition in cars," in *Proc. 5th Int. Conf. Signal Processing Applications and Technology*, vol. I, pp. 154–15, 1994.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 5, pp. 4–24, Apr. 1988.

[4] B. Widrow, K. M. Duvall, P. R. Gooch, and W. C. Newmann, "Signal cancellation phenomena in adaptive antennas: causes and cures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 469–478, May 1982.

[5] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1391–1400, Dec. 1986.

[6] M. M. Sondhi and G. W. Elko, "Adaptive optimization of microphone arrays under a nonlinear constraint," in *Proc. ICASSP'86*, pp. 981–984.

[7] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP'88*, pp. 2578–2581.

[8] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," presented at *2nd Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux-Technopolis, France, Sept. 30–Oct. 2, 1992.

[9] K. U. Simmer, P. Kuczynski, and A. Wasiljeff, "Time delay compensation for adaptive multichannel speech enhancement systems," in *Proc. ISSSE'92*, pp. 660–663.

[10] Z. Yang, K. U. Simmer and A. Wasiljeff, "Improved performance of multi-microphone speech enhancement systems," in *Proc. 14th GRETSI Symp.*, 1993, pp. 479–482.

[11] S. Gierl, "Noise reduction for speech input systems using an adaptive microphone-array," in *Proc. 22nd ISATA*, 1990, pp. 517–524.

[12] D. Van Compernolle, "Adaptive filter structures for enhancing cocktail party speech from multiple microphone recordings," in *Proc. 12th GRETSI Symp.*, 1989, vol. 1, pp. 513–516.

[13] ——, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. ICASSP'90*, vol. 2, pp. 833–836.

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

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

[16] R. Le Bouquin-Jeannès and G. Faucon, Eds., "Advanced solutions for noise reduction," Deliverable 4.123.2, ESPRIT Project 6166 FREETEL, Univ. Rennes, Rennes, France, July 1994.

[17] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, July 1979.

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

[19] S. Affes, S. Gazor, and Y. Grenier, "Wideband robust adaptive beamforming via target tracking," in *Proc. 7th IEEE Signal Processing Workshop on SSAP*, 1994, pp. 141–145.

[20] ——, "An algorithm for multisource beamforming and multitarget tracking," *IEEE Trans. Signal Processing*, vol. 44, pp. 1512–1522, June 1996.

[21] H. Kuttruff, *Room Acoustics*. Applied Science, 1979.

[22] Y. Grenier, Ed., "Characterization of the environments," Deliverable 2.2, ESPRIT Project 6166 FREETEL, ENST-ARECOM, Paris, France, July 1993.

[23] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Biol.*, vol. 15, pp. 267–273, 1982.

[24] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

[25] S. Gazor, S. Affes, and Y. Grenier, "Wideband multi-source beamforming with adaptive array location calibration and direction finding," in *Proc. ICASSP'95*, vol. III, pp. 1904–1907.

[26] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[27] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Mag.*, vol. 9, pp. 14–37, Jan. 1992.

[28] E. Moulines, O. A. Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: structure and convergence analysis," *IEEE Trans. Signal Processing*, vol. 43, pp. 14–28, Jan. 1995.

[29] J. Prado and E. Moulines, "Frequency-domain adaptive filtering with applications to acoustic echo cancellation," *Ann. Télécommun.*, vol. 49, pp. 414–428, July 1994.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[31] S. Affes, S. Gazor, and Y. Grenier, "A subarray manifold revealing projection for partially blind identification and beamforming," *IEEE Signal Processing Lett.*, vol. 3, pp. 187–189, June 1996.

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

Sofiène Affes (SM'94–M'97) was born in Kasserine, Tunisia, on September 19, 1969. He received the Ingénieur degree in electrical engineering in 1992, and the Ph.D. degree in the field of signal and image processing in 1995, both from Ecole Nationale Supérieure des Télécommunications, Paris, France.

During 1991, he was a junior visitor at the Engineering Department, Cambridge University, Cambridge, U.K., working on speech recognition. He is currently a Research Associate at Institut National de la Recherche Scientifique-Télécommunications, Montreal, P.Q., Canada, working on mobile and personal communication systems. His interests are in statistical signal and array processing with applications to acoustics and speech processing, and in digital communications. He has been involved in European ESPRIT projects 2101 ARS during 1991 and 6166 FREETEL from 1993 to 1994. He currently contributes to a project in wireless communications with the Canadian Institute for Telecommunications Research.

Yves Grenier (A'80–M'81) was born in Ham, Somme, France, in 1950. He received the Ingénieur degree from Ecole Centrale de Paris, France, in 1972, the Docteur-Ingénieur degree from Ecole Nationale Supérieure des Télécommunications, Paris, in 1977, and the Doctorat d'Etat es Sciences Physiques from University of Paris-Sud, France, in 1984.

He has been with Ecole Nationale Supérieure des Télécommunications, Paris, since 1977 as Assistant, and since 1984 as Professor. Until 1979, his interests have been in speech recognition, speaker identification, and speaker adaptation of recognition systems. He has then been working on signal modeling, spectral analysis of noisy signals, with applications in speech recognition and synthesis, estimation of nonstationary models, and time–frequency representations. Between 1984 and 1988, he created ARMALIB, a signal processing software library that has been incorporated into SIMPA, the signal processing software proposed by GDR-PRC CNRS ISIS. He is presently interested in array processing, with applications to acoustics (acoustic echo cancellation, noise reduction, and microphone arrays).

Dr. Grenier is a member of the Société Française d'Acoustique. He has been involved in the European ESPRIT projects 2101 ARS from 1989 to 1992 and 6166 FREETEL from 1992 to 1994.