A SOURCE SUBSPACE TRACKING ARRAY OF MICROPHONES FOR DOUBLE TALK SITUATIONS

Sofiène AFFES [1,2]

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

Yves GRENIER [2]

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

ABSTRACT

We propose in this paper an array of microphones which achieves good performance in double talk situations. By a subspace tracking procedure, we jointly identify the acoustic paths from the speech and echo sources. With an adaptive beamformer constrained over these paths, we properly recover the original speech, cancell the echo and reduce the background noise. Simulations made with real data confirm the efficiency of this array.

1. INTRODUCTION

Microphone array systems are today the subject of growing developments for acoustic applications, especially in hands-free telecommunication [1]. For a reliable acquisition of the user, the major functions these systems need to implement in hands-free operations are: the enhancement of speech degraded by reflections and reverberations, and noise reduction in general, from which we explicitly mention the specific problem of acoustic echo cancellation (AEC). So far, these aspects have been studied separately, while microphone arrays have been usually considered for noise reduction. Actually, their combination remains an open area of research, particularly in double talk situations where both speech and echo are simultaneously present with background noise. We herein address the problem of double talk, and propose an array of microphones which properly achieves the above requirements.

The advances made in AEC are mostly related to adaptive filtering, and multi-channel AEC (M-C AEC) was not studied until recently [2]. However, their formulation always reduces to a separate AEC from a single or multiple loudspeakers at each microphone input. The combination of these inputs for the reduction of speech reverberations and noise by array processing is not of their scope, and double situations are not well studied yet.

Kellerman [3] recently proposed to feed the M-C AEC outputs to the inputs of an array beamformer after time-delay synchronization. He also proposed a single AEC filter applied on the beamformer output, and noticed for both structures their advantage to reduce the AEC filter lengths. In either cases, double talk first disturbs AEC due to the speech, or beamforming due to the echo. Previously, Xu proposed in [4] an attractive beamformer orthogonal to the acoustic echo paths by spatial filtering, but considered a geometrical source propagation as usually assumed by microphone array systems. In reverberant environments, this approximation severely reduces the efficiency of these systems in speech acquisition and noise reduction [6].

In this contribution, we propose along the lines mentioned above an array of microphones which is proved to be efficient for double talk situations. We actually merge AEC in a multi-source scheme of subspace tracking adapted from [5] and [6], to simultaneously estimate the acoustic paths of both the speech and echo sources. This scheme guarantees isolation between the speech and echo by constrained beamforming, and provides stability of the array processor during double talk. We particularly consider for the application a hands-free telephone of 6 microphones and a single loudspeaker as shown in figure 1. This system is to be used by operators in a banker market trading room. We thus run our simulations with data recorded in real conditions, although other applications combining M-C AEC with multiple loudspeakers can be viewed.

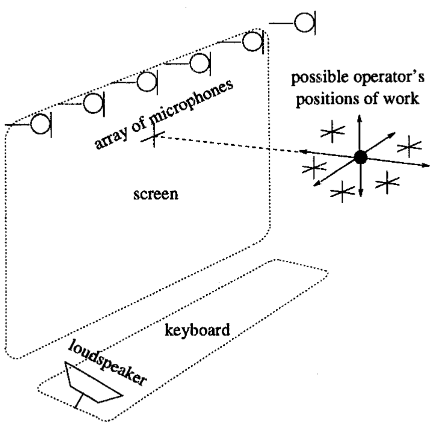

Figure 1: Configuration of a hands-free telephone in a banker market trading room.

2. FORMULATION AND BACKGROUND

We model the signals received from the array of $m = 6$ microphones in the frequency domain as follows:

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

0-7803-3192-3/96 $5.00©1996 IEEE

The subscripts $f = 0, \cdots, 2L-1$ and n respectively denote in (1) the FFT (Fast Fourier Transform) over $2L = 512$ snapshots at 8 kHz of the indexed quantity at the frequency bin f and the block of input data number n. $X_{f,n}$ denotes the m-dimensional observation vector. $s_{f,n}$ is the speech signal emitted from the operator, $e_{f,n}$ is the available echo signal emitted from the loudspeaker, and $N_{f,n}$ is the noise vector. G_f and H_f denote the m-dimensional vectors of IRs (impulse responses) to the microphones respectively from the operator's mouth and the loudspeaker. We actually assume their variations to be very slow and practically constant, and hence assume $G_{f,n} \simeq G_f$ and $H_{f,n} \simeq H_f$ for simplicity although it is acquired that time-variations can be tracked. Notice that the formulation of equation (1) still holds for a M-C AEC. The scalar $e_{f,n}$ could be replaced by $E_{f,n}$, a $p \times 1$ vector of echo sources emitted from $p < m$ loudspeakers, whereas H_f would be a $m \times p$ matrix.

In (1), we observe ambiguities due to the multiplicative factors between G_f and $s_{f,n}$, and between H_f and $e_{f,n}$. If we define the mean energies $\beta_f^2 = \frac{\|G_f\|^2}{m}$ and $\gamma_f^2 = \frac{\|H_f\|^2}{m}$, we can reformulate (1) by:

$$X_{f,n} = \underline{G}_f \underline{s}_{f,n} + \underline{H}_f \underline{e}_{f,n} + N_{f,n} \ , \tag{2}$$

where $\underline{G}_f = \frac{G_f}{\beta_f}$ and $\underline{H}_f = \frac{H_f}{\gamma_f}$ normalized to $\sqrt{m}$, can be seen as propagation vectors of the modulated narrowband speech and echo sources $\underline{s}_{f,n} = \beta_f s_{f,n}$ and $\underline{e}_{f,n} = \gamma_f e_{f,n}$, and where β_f and γ_f are the modulation factors.

If the sources propagate along spherical or planar wavefronts, we can apply the subspace-based algorithm in [5] to directly track the steering vectors in the array manifold. The forcing projection of steering vectors in the array manifold guarantees their convergence. Like previous microphone array systems, we can synchronize $X_{f,n}$ along the corresponding time-delay estimates of the direct path prior to an optimal beamforming. However, we noticed in [6] that time delay compensation is inappropriate in adverse acoustic environments because reflections and reverberations are not negligible. We rather underlined our conclusion that IRs should be fully identified and compensated to really achieve satisfactory results. Along this line, we adapted [5] to a specific application of speech acquisition and noise reduction (*i.e.* no echo: $e_{f,n} = 0$) to directly track $\underline{G}_f$ although it is not assigned in an array manifold [6].

Actually, we first noticed from an acoustic characterization that β_f is constant for operator's work positions around a central location (see figure 1) and measured it. We also applied from the block processing scheme a linear convolution constraint on the estimates of $G_f = \beta_f \underline{G}_f$. This step which forces them in a particular structure instead of an array manifold still helps them to convergence particularly at a low SNR (Signal to Noise Ratio) [6].

In double talk situations where the echo level is very strong, the acoustic characterization still holds, but the structure forcing step is no longer sufficient to avoid possible deviations of the estimates of G_f to H_f. To guarantee isolation, we merge in this contribution AEC in a multi-source scheme of source-subspace tracking [5], using an available reference of the echo $e_{f,n}$. This known reference which generates the echo source-subspace defined by the vector $\underline{H}_f$ can provide the echo-free hyperplane where the speech source can be estimated again with [6]. In a 2-D generalization of [6], we jointly and directly track the speech and echo source-subspaces defined by $\underline{G}_f$ and $\underline{H}_f$ as shown in the following section.

3. THE PROPOSED ALGORITHM

We implement the algorithm in 4 steps:

3.1. Beamforming

We first assume that estimates of $\underline{G}_f$ and $\underline{H}_f$ at iteration n say $\hat{\underline{G}}_{f,n}$ and $\hat{\underline{H}}_{f,n}$ are available and near convergence. In the echo-free observation-hyperplane, we can extract the speech source-subspace component with a distortionless beamformer $U_{f,n}$ (*i.e.* $U_{f,n}^H \hat{\underline{G}}_{f,n} = 1$) constrained to cancelling the echo (*i.e.* $\hat{U}_{f,n}^H \hat{\underline{H}}_{f,n} = 0$). For $f = 0, \cdots, L$ we have:

$$\begin{aligned} \tilde{\underline{s}}_{f,n} &= U_{f,n}^H X_{f,n} \ , \\ U_{f,n} &= \frac{\left\|\hat{\underline{H}}_{f,n}\right\|^2 \hat{\underline{G}}_{f,n} - \left(\hat{\underline{G}}_{f,n}^H \hat{\underline{H}}_{f,n}\right) \hat{\underline{H}}_{f,n}}{\left\|\hat{\underline{G}}_{f,n}\right\|^2 \left\|\hat{\underline{H}}_{f,n}\right\|^2 - \left|\hat{\underline{G}}_{f,n}^H \hat{\underline{H}}_{f,n}\right|^2} \ . \end{aligned} \tag{3}$$

With a $m \times p$ matrix $H_{f,n}$, we can take in general $U_{f,n}$ as the first column beamformer of $\hat{A}_{f,n}(\hat{A}_{f,n}^H \hat{A}_{f,n})^{-1}$, the pseudoinverse of $\hat{A}_{f,n} = [\hat{\underline{G}}_{f,n}, \hat{\underline{H}}_{f,n}]$. Normally, the remaining column beamformers would have been used to provide estimates of the echo signals, if references were not available in the studied context of AEC (see [5,9]).

The conventional beamforming structure in (3) provides immunity against the echo interference and is optimal for white noise reduction, but further improvements can be proposed for an optimal processing of correlated noise. We can also extract the $m-2$ noise subspace components with the $m \times (m-2)$ blocking matrix $P_{f,n}$ by:

$$Y_{f,n} = P_{f,n}^H X_{f,n} \ , \tag{4}$$

such that $P_{f,n}^H[\hat{\underline{G}}_{f,n}, \hat{\underline{H}}_{f,n}] = [0_{m-2}, 0_{m-2}]$. A noise filter $W_{f,n}$ can be trained from these components to further reduce the residual noise still present in $\tilde{\underline{s}}_{f,n}$ in an optimal GSC (Generalized Sidelobe Canceller) structure [7] by:

$$\begin{aligned} \hat{\underline{s}}_{f,n} &= \tilde{\underline{s}}_{f,n} - W_{f,n}^H Y_{f,n}, \\ W_{f,n+1} &= W_{f,n} + \eta_{f,n} Y_{f,n} \hat{\underline{s}}_{f,n}^H, \end{aligned} \tag{5}$$

where $\eta_{f,n}$ is an adaptation step-size of $W_{f,n}$ possibly including a normalization factor. In the general case of M-C AEC, iterative implementations of (3) and (4) can be easily adapted from [9] to reduce complexity in (5).

Unlike the method in [3] which implements an echo canceller at the array output, we cancell the echo by spatial filtering in (3). We thus provide isolation from the echo within the array processor, to efficiently deal with double talk situations. A structure similar to ours was proposed in [4], but input signals in (3) are simply time-delayed therein.

Prior methods which compensate time delays would observe speech cancellation in (5) due to steering errors and speech leaks in $Y_{f,n}$ [8]. For this reason, they either avoid adaptive beamforming or use suboptimal structures (see references in [6]). On the other hand, we avoid signal cancellation in speech estimation and efficiently reduce both the echo and noise by adjusted and constrained beamforming. Besides, we noticed in [5,9] that the underlying multi-source structure is robust to source coherence. In the studied context, this useful feature amounts to robustness to a possible correlation between the speech and echo signals.

3.2. Source-subspace tracking

From the observation vector $X_{f,n}$, the estimated speech $\hat{\underline{\mathrm{s}}}_{f,n}$ and the echo $e_{f,n}$, we can jointly track $\underline{\mathrm{G}}_{f,n}$ and $\underline{\mathrm{H}}_{f,n}$ in an input/output identification-like procedure as in [6]. We actually generalize [6] along the multi-source scheme described in [5] to separately identify the speech and echo source-subspaces. $e_{f,n}$ is available and H_f can be equivalently tracked instead of $\underline{\mathrm{H}}_f$. We also use $\tilde{\underline{\mathrm{s}}}_{f,n}$ instead of $\hat{\underline{\mathrm{s}}}_{f,n}$ for a better stability as noticed in [6]. If we define the noise vector estimate by:

$$\hat{N}_{f,n} = X_{f,n} - \left[\hat{\underline{\mathrm{G}}}_{f,n}, \hat{H}_{f,n}\right] \left[\begin{array}{c} \tilde{\underline{\mathrm{s}}}_{f,n} \\ e_{f,n} \end{array} \right] , \tag{6}$$

we can separately state the tracking equations as follows for $f = 0, \cdots, L$:

$$\begin{aligned} \tilde{\underline{\mathrm{G}}}_{f,n+1} &= \hat{\underline{\mathrm{G}}}_{f,n} + \delta^s_n \mu^s_{f,n} \; \hat{N}_{f,n} \; \tilde{\underline{\mathrm{s}}}^H_{f,n} \, , & (7) \\ \tilde{H}_{f,n+1} &= \hat{H}_{f,n} + \delta^e_n \mu^e_{f,n} \; \hat{N}_{f,n} \; e^H_{f,n} \, . & (8) \end{aligned}$$

$\tilde{\underline{\mathrm{G}}}_{f,n+1}$ and $\tilde{H}_{f,n+1}$ denote at present unconstrained estimates until structure forcing within an acoustic characterization in the next subsection. The adaptation step-sizes $\mu^s_{f,n}$ and $\mu^e_{f,n}$ are possibly normalized. δ^s_n is the speech activity detector of [6] equal to 0 during speaker's silence and 1 otherwise, whereas δ^e_n is the echo activity detector simply activated over the signal energy of the loudspeaker.

Notice that (8) still holds in the case of M-C AEC where $e_{f,n}$ could be replaced by a $p \times 1$ vector $E_{f,n}$ with a $m \times p$ matrix $H_{f,n}$. Actually, this equation can be merged in (7) in a multi-source scheme of array processing in [5], where the loudspeakers can be seen as multiple sound sources with known reference signals. Alternatively, equation (7) can be merged in (8) in a M-C AEC scheme, where the speech source can be seen as a "virtual loudspeaker" with an unknown reference signal. In any case, the convergence is guaranteed by a joint subspace tracking of both the speech and echo.

The key point of subspace tracking is the noise vector estimate in (6) free from speech and echo, which guarantees a better stability of the LMS-type tracking equations (7) and (8) during double talk. The LMS-like tracking equation proposed in [6] is similar to (7), but the error term $\hat{N}_{f,n}$ is not cleaned from the echo, and $\tilde{\underline{\mathrm{s}}}_{f,n}$ is not estimated in the echo-free hyperplane. Besides, equation (8) is similar to an exact LMS implementation of AEC or even M-C AEC [2]. But contrarily to these methods, the error term $N_{f,n}$ in (8) is free from speech. Equation (8) identifies H_f and the echo-free hyperplane. This guarantees convergence to G_f and s_f by (7), (3) and (5) as in [6]. Simultaneously, $\hat{N}_{f,n}$ is better estimated and the tracking perturbations are significantly reduced in double talk situations.

3.3. Acoustic characterization

In equations (7) and (8), notice that $\tilde{\underline{\mathrm{G}}}_{f,n+1}$ and $\tilde{H}_{f,n+1}$ denote unconstrained estimates of $\underline{\mathrm{G}}_f$ and H_f. As in [6], we assign their structure in the time domain to correspond to a linear convolution in a block processing scheme. This step amounts to setting the last half of IR coefficients to zero as shown below.

With the measured modulation factors β_f and the unconstrained estimates of the normalized steering vectors $\tilde{\underline{\mathrm{G}}}_{f,n+1}$, we first form the following matrix:

$$\begin{aligned} \tilde{\mathcal{G}}_{n+1} &= \left[\beta_0 \tilde{\underline{\mathrm{G}}}_{0,n+1}, \cdots, \beta_{2L-1} \tilde{\underline{\mathrm{G}}}_{2L-1,n+1}\right] & (9) \\ &= \left[\tilde{G}_{0,n+1}, \cdots, \tilde{G}_{2L-1,n+1}\right] , \end{aligned}$$

whose rows approximate the FFT of IRs. The terms indexed from $L+1$ to $2L-1$ are obtained by a Hermitian symmetry. We secondly compute the row by row IFFT (Inverse FFT) of $\tilde{\mathcal{G}}_{n+1}$ whose rows approximate IRs in the time domain, then set its $m \times L$ right half part to 0. We again take the row by row FFT of this constrained matrix to finally estimate $\hat{\mathcal{G}}_{n+1}$ or equivalently have $\hat{G}_{f,n+1} = \beta_f \hat{\underline{\mathrm{G}}}_{f,n+1}$. In the same way, we compute $\hat{\mathcal{H}}_{n+1}$ from $\tilde{\mathcal{H}}_{n+1}$ to have $\hat{H}_{f,n+1}$.

These linear convolution constraints improve convergence as noted in [6], but can be skipped at a high SNR to save computations. Indeed, isolation between the speech and echo source-subspaces can be sufficiently guaranteed in that case by the adjusted and constrained beamforming in subsection 3.1.

3.4. Speech estimation

In the model equation (1), input signals are transformed to the frequency domain in an analysis/synthesis scheme. Data blocks are actually shifted each $K = 16$ snapshots. This oversampling is shown [6] to improve the convergence behavior of the tracking equations (7) and (8). We thus estimate the speech signal at the block $n+1$ in the time domain by:

$$\begin{aligned} &\left[\hat{s}\left(K(n+1)\right), \cdots, \hat{s}\left(K(n+1)+2L-1\right)\right] \triangleq \\ &\qquad \mathrm{Re}\left\{ \mathrm{IFFT}\left(\left[\frac{\hat{\underline{\mathrm{s}}}_{0,n}}{\beta_0}, \cdots, \frac{\hat{\underline{\mathrm{s}}}_{2L-1,n}}{\beta_{2L-1}} \right] \right) \right\} , \end{aligned} \tag{10}$$

where Re{.} denotes the real part of a complex number. As blocks overlap over $2L - K$ samples, we only keep the following segment of length K for subsampling:

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

Unlike previous microphone arrays which simply synchronize reverberated speech along time delay estimates of the direct path, we compensate reflections and reverberations and recover a natural quality of speech. We also clean the speech from the echo and the background noise by spatial filtering. AEC methods properly cancell the echo, but are not as efficient as microphone array systems in noise reduction, and dot not remove speech reverberations.

If required however in some other applications, speech can be "spatialized" in the remote room at $p < m$ loudspeakers symmetrical to p microphones in the local room (*e.g.* $p = 2$ for a stereo effect). For instance, we can use the corresponding p components of $\underline{\hat{\mathrm{G}}}_{f,n}\underline{\hat{\mathrm{s}}}_{f,n}$ in (10) to approximately reproduce the output of classical M-C AEC without background noise. But contrarily to these methods, we can also use elaborate "spatialization" techniques [10] of the estimated speech in (10) when the local and remote rooms have arbitrary acoustics and/or configurations, or when special acoustic effects are required.

4. SIMULATION RESULTS

We consider 2 original signals of 2 speech sentences each, respectively uttered from a female and a male speakers and recorded in anechoic room. The female speech source in figure 2-a is convolved with measured IRs from the array to a nominal operator's position of work to simulate the desired speech, whereas the male speech source is convolved with IRs from the array to the loudspeaker to simulate the echo. Both signals are added to simulate a double talk situation. They are also corrupted by background noise containing cocktail party speech of other operators, the noise of keyboards, the noise of workstations, etc⋯. The echo is twice stronger than the desired speech and covers it each time as shown in figure 2-b. Actually, the definition of simple measurements for an objective evaluation in double talk situations are not clearly defined at present. We however measure a rough value of a SNER ("Signal to Noise plus Echo Ratio") of -3 dB in average at each microphone.

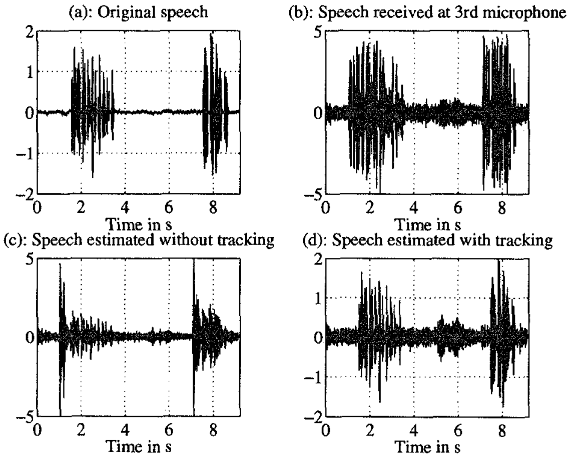

Figure 2: Speech signal at different stages.

We start the algorithm with time-delay IRs corresponding to a geometrical propagation from the source positions. Without adaptation (*i.e.* skip equations (7) and (8)), we show in figure 2-c that speech cancellation is effective, while the proposed algorithm cancells the echo, reduces the background noise, and avoids the speech cancellation as shown in figure 2-d. We actually measure a rough value of 10 dB of SNER at the output with a gain of 13 dB. We also recover a natural quality of speech at the output, and do confirm at the listening the efficiency of this method in double talk situations.

In the near future, this array should be evaluated in a M-C AEC context. Successful tests without AEC were already made to deal with the case of multiple users in a mini-teleconference mode. This configuration should be also assessed with AEC in a full hands-free context. The oversampling rate should be finally improved by considering other adaptive filtering versions in the subspace tracking equations.

5. ACKNOWLEDGMENT

The authors wish to thank all the partners of the ESPRIT Project 6166 FREETEL on hands-free telephony for providing the data for simulations. Recordings were made by ENST and PAGE Iberica in a banker market trading room of Banesto, Madrid, Spain.

6. REFERENCES

[1] G.W. Elko, "Microphone array systems for hands-free telecommunication", *Proc. of the IEEE International Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 31-38, June 1995.

[2] A. Gilloire, "Recent advances in adaptive filtering algorithms for acoustic echo cancellation", *Proc. of the IEEE International Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 115-134, June 1995.

[3] W. Kellerman, "Some properties of echo path impulse responses of microphone arrays and consequences for acoustic echo cancellation", *Proc. of the IEEE International Workshop on Acoustic Echo and Noise Control*, Røros, Norway, pp. 39-43, June 1995.

[4] M. Xu, *Antenne acoustique adaptative pour la prise de son*, PhD Dissertation, ENST, Paris, France, December 1988.

[5] S. Affes, S. Gazor, and Y. Grenier, "An algorithm for multi-source beamforming and multi-target tracking", *To appear in the IEEE Trans. on Signal Processing*, submitted January 1995, accepted September 1995.

[6] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for speech acquisition and noise reduction with a microphone array", *Proc. of the IEEE/IEE Workshop on Signal Processing Methods in Multipath Environments*, Glasgow, UK, pp. 64-73, April 20-21, 1995.

[7] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27-34, January 1982.

[8] B. Widrow, K.M. Duvall, P.R. Gooch, and W.C. Newmann, "Signal cancellation phenomena in adaptive antennas: causes and cures", *IEEE Trans. on ASSP*, vol. 30, no. 3, pp. 469-478, May 1982.

[9] S. Affes, *Formation de voie adaptative en milieux réverbérants*, PhD Dissertation, ENST, Paris, France, October 1995.

[10] J.M.Jot, *Etude et réalisation d'un spatialisateur de sons par modèles physiques et perceptifs*, PhD Dissertation, ENST, Paris, France, September 1992.