Singular Block Toeplitz Matrix Approximation and Application to Multi-Microphone Speech Dereverberation

Samir-Mohamad Omar, Dirk T.M. Slock
Mobile Communications Department, EURECOM
2229 Route des Crêtes BP 193, 06904 Sophia Antipolis Cedex, France
{omar,slock}@eurecom.fr

***Abstract*—We consider the blind multichannel dereverberation problem for a single source. We have shown before [5] that the single-input multi-output (SIMO) reverberation filter can be equalized blindly by applying MIMO Linear Prediction (LP) to its output (after SISO input pre-whitening). In this paper, we investigate the LP-based dereverberation in a noisy environment, and/or under acoustic channel length underestimation. Considering ambient noise and late reverberation as additive noises, we propose to introduce a postfilter that transforms the MIMO prediction filter into a somewhat longer equalizer. The postfilter allows to equalize to non-zero delay. Both MMSE-ZF and MMSE design criteria are considered here for the postfilter. We also focus here on computationally efficient (FFT based) block Toeplitz covariance matrix enhancement that enforces the SIMO filtered source plus white noise structure before applying MIMO LP. A second suggested refinement is an iterative refinement between SISO and MIMO LP. Simulations show that the proposed scheme is robust in noisy environments, and performs better compared to the classic Delay-&-Predict equalizer and the Delay-&-Sum beamformer.**

I. Introduction

Blind dereverberation is the process of removing the effect of reverberation from an observed reverberant signal. Reducing the distortion caused by reverberation is a difficult blind deconvolution problem, due to the colored and non-stationary nature of speech and the length of the equivalent impulse response from the speaker's mouth to the microphone(s). Consider a clean speech signal, s_k, produced in a reverberant room. The reverberant speech signal observed on M distinct microphones can be written as:

$$\mathbf{y}_k = \mathbf{h}(q)\, s_k \tag{1}$$

where $\mathbf{y}_k = [y_{1,k} \cdots y_{M,k}]^T$ is the reverberant speech signal, $\mathbf{h}(z) = [h_1(z) \cdots h_M(z)]^T = \sum_{i=0}^{L_h-1} \mathbf{h}_i z^{-i}$ is the SIMO FIR channel transfer function, L_h is the channel length. The introduction of q, where q^{-1} is the one sample time delay operator: $q^{-1}s_k = s_{k-1}$, allows to introduce the compact notation of transfer functions in the time domain (whereas z in the z-transform is a complex number).

Blind dereverberation faces the channel/speech source identifiability problem. In fact, for any invertible scalar filter $\alpha(q)$, $(\alpha(q)\mathbf{h}(q),\ (1/\alpha(q))\, s_k)$ is also an acceptable solution for (1). In [1], the authors compute a multichannel FIR equalizer using a subspace based method. The identifiability problem is solved using accurate information of the "source" (or "noise") subspace dimension. The validity of the technique hinges critically on the true channel impulse response being of strictly finite duration, and its successful identification requires knowledge of the channel length [2]. For the acoustic case, the true channel impulse response length is generally unknown and/or ill-defined. This is a major limitation to the practical applicability of the subspace based methods to speech dereverberation.

In contrast, the alternative Linear Prediction (LP) based technique (proposed and refined by Slock et al. [3], [4]) proved to be consistent in the presence of channel order error. This makes the LP equalizer one of the more attractive solutions for blind speech dereverberation, as proposed in [5]. One tricky issue though is that in order for the LP to perform zero delay channel equalization, the source should be white, otherwise LP will perform both channel equalization and source whitening. Hence, in the case of speech dereverberation, some additional processing is required. In [5], [6], the speech correlation gets compensated via a SISO pre-whitening at the LP equalizer input (microphone signals). Next, the multivariate LP can be computed, and applied to the reverberant microphone signals $\mathbf{y}_k$:

$$\underbrace{\mathbf{u}_k}_{M\times 1} = \underbrace{A(q)}_{M\times M}\ \underbrace{\mathbf{y}_k}_{M\times 1} = \underbrace{\mathbf{h}_0}_{M\times 1}\ \underbrace{s_k}_{1\times 1} \quad \text{since} \quad A(z)\,\mathbf{h}(z) = \mathbf{h}_0 \tag{2}$$

where $A(q)$ is the MIMO linear prediction error filter, and $\mathbf{h}_0 = \mathbf{h}(z = +\infty)$ is the multichannel precursor coefficient. The LP equalizer is obtained by performing Maximum Ratio Combining (MRC) ($\mathbf{h}_0^T$) on the prediction error signal $\mathbf{u}_k$ components.

In [8] a somewhat related approach has been proposed, in which only the first microphone signal (assumed to have the shortest delay) is predicted in terms of the past samples on all microphones (MISO prediction). Compared to MIMO prediction, MISO prediction loses the MRC advantage. Since the MISO prediction is applied directly to $\mathbf{y}_k$, a dereverberated but also whitened source signal gets produced. Now, multivariate channel prediction assumes that the individual microphone channel transfer functions $h_i(z)$ $(i = 1, \ldots, M)$ have no SISO transfer function factor in common. If such a common factor exists, or equivalently if the source is colored, the multivariate LP will model this factor with an all-pole filter and the LP filter will contain a scalar transfer function factor that is the inverse of the all-pole model. This scalar factor can be determined as the common roots of the M MISO LP component transfer function polynomials or, as in [8], as the eigenvalues of a large matrix of which the MISO LP coefficients constitute one column. Postfiltering of the MISO LP residual with the inverse of the extracted factor then allows to recover in principle the unwhitened source. This common root extraction approach is prone to ill-conditioning, as the results in [8] tend to confirm. Indeed, due to the tapered off behavior of the late reverberation on all microphones, the $h_i(z)$ tend to have zeros that cluster near the origin and hence that are close or (almost) in common. This is not a big problem for the MIMO LP approach in [5], [6] where the effect is that the reverberation tail will not get equalized, but it is small anyway. For the purpose of the determination of the source color as in [8] on the other hand, the effect of such ill-conditioning is more severe.

In [9],[10] the so-called TRINICON method was introduced for blind separation of acoustic sources. One of the main characteristics of the objective function optimized by the TRINICON method, which is also based completely on second-order statistics (SOS), is that the extracted sources at the output of a MIMO FIR filter are as jointly decorrelated as possible, apart from intra-source correlations. In other words, the MIMO FIR demixing filter tries to produce source estimates with as little inter-source correlation as possible. As a result the cascade of the MIMO demixing and mixing filters will tend to a diagonal MIMO filter (apart from source permutations) and hence the sources may appear in a filtered fashion. Hence the problem solved is not so much that of dereverberation but of source separation. Also, the method is only applicable starting with at least two sources. And in spite of being SOS based, the objective function is not quadratic and requires an iterative (natural gradient based) solution.

Dereverberation techniques are generally introduced in a noiseless environment (the problem is already quite difficult even under these ideal conditions). In this paper, we propose a robust scheme for dereverberation in the presence of noise. This noise may be either additive acoustic noise or residual late reverberation due to underestimation of the reverberation delay spread (for computational complexity reasons or for estimation considerations in non-stationary environments). We investigate the resulting dereverberation performance in a noisy environment.

We next summarize the basic D-&-P equalization technique from [5], [6], [7]. At first $\mathbf{y}_k$ gets replaced by $D(q)\,\mathbf{y}_k$, a microphone-wise delayed version of the microphone signals so that the source signal arrives with the same delay at all microphones. We shall denote the aligned version of $\mathbf{y}_k$ still by $\mathbf{y}_k$. Next, a SISO source LP filter $A_s(z)$ gets determined by performing LP on the $y_{i,k}$ SOS averaged over the M microphones. We then obtain $\mathbf{x}_k = A_s(q)\,\mathbf{y}_k = \mathbf{h}(q)\,\widetilde{s}_k$ where $\widetilde{s}_k = A_s(q)\,s_k$ is the whitened source signal. MIMO LP on $\mathbf{x}_k$ yields a prediction error

$$\widetilde{\mathbf{x}}_k = A_{\mathbf{x}}(q)\,\mathbf{x}_k = \mathbf{h}_0\,\widetilde{s}_k \quad \text{with} \quad A_{\mathbf{x}}(z)\,\mathbf{h}(z) = \mathbf{h}_0\,. \tag{3}$$

Finally, the dereverberated source gets estimated as $\widehat{s}_k = \mathbf{h}_0^T\,A_{\mathbf{x}}(q)\,\mathbf{y}_k$.

II. Robust Delay-&-Predict Equalization in Noisy Environments

In a noisy environment, the microphone signals can be written as

$$\mathbf{y}_k = \mathbf{h}(q)\,s_k + \mathbf{v}_k \tag{4}$$

where the noise $\mathbf{v}_k$ represents acoustic noise and/or the effect of modeling error in $\mathbf{h}(z)$. We shall model $\mathbf{v}_k$ as spatiotemporally white noise (with spectrum $S_{\mathbf{v}}(z) = \sigma_v^2\,I_M$), independent of s_k. Such noise, for given noise power, is the worst case noise. In any case, at medium to high SNR, the correlation of the noise is a secondary effect compared to accounting for the noise power. The SISO and MIMO LP problems in the dereverberation approach considered here should still be formulated for the noise-free signals, even in the noisy case. However, since the LP problems only involve SOS, the noiseless SOS can easily be obtained from the noisy SOS in the white noise hypothesis, especially in the multichannel configuration considered here in which signal and noise subspaces arise. The simplest SOS denoising would be to subtract the noise covariance matrix ($\sigma_v^2 I$) from the covariance matrix $R_{\mathbf{y}}$ of $\mathbf{y}_k$ by estimating σ_v^2 from the noise subspace eigenvalue(s) of $R_{\mathbf{y}}$. Various degrees of sophistication are possible, some of which will be evoked later. Applying the (noiseless) MIMO LP to the noisy microphone signals, we get

$$\mathbf{u}_k = A_{\mathbf{x}}(q)\,\mathbf{y}_k = \mathbf{h}_0\,s_k + A_{\mathbf{x}}(q)\,\mathbf{v}_k\,. \tag{5}$$

The robustified D&P equalizer then gets constructed as

$$F_{D-\&-P}(q) = \mathbf{w}(q)\,A_{\mathbf{x}}(q)\,,\quad \widehat{s}_k = F_{D-\&-P}(q)\,\mathbf{y}_k = \mathbf{w}(q)\,\mathbf{u}_k \tag{6}$$

whereas the basic D&P equalizer uses $\mathbf{w}(q) = \mathbf{h}_0^T$, which maximizes the power of the desired signal part but not necessarily the output SNR. In [7], we have proposed the postfilter $\mathbf{w}(q)$ with a MMSE-ZF design using explicitely the white noise hypothesis (in a multichannel configuration, there is an infinity of zero-forcing designs, one of which will be MMSE). The filter length of $\mathbf{w}(q)$ allows the design of non-zero-delay equalizers. Here we shall consider the design of the postfilter using the MMSE-ZF and MMSE criteria, without a white noise hypothesis.

A. MMSE-ZF Design

For a given filter length L_w and an equalization delay $0 \leq d \leq (L_w{-}1)$, the weighting filters are optimized by maximizing the output SNR (under the d-delay zero-forcing constraint), i.e.

$$\begin{cases} \mathbf{w} = \arg\max\limits_{\mathbf{w}} \dfrac{\sigma_s^2}{\displaystyle\oint \mathbf{w}(z) S_{\mathbf{u}}(z) \mathbf{w}^{\dagger}(z) \frac{dz}{2\pi j z} \; - \sigma_s^2} \\ \mathbf{w}(z)\,\mathbf{h}_0 = z^{-d} \end{cases} \tag{7}$$

where $\mathbf{w}^{\dagger}(z)$ denotes the paraconjugate (matched filter, Hermitian transpose in the Fourier domain, Hermitian transpose and time reversal of the coefficients in the time domain) of $\mathbf{w}(z)$, and $S_{\mathbf{u}}(z) = A_{\mathbf{x}}(z)\, S_{\mathbf{y}}(z)\, A_{\mathbf{x}}^{\dagger}(z)$ is the matrix spectrum of $\mathbf{u}_k$. For a time domain formulation, let $\underline{\mathbf{w}} = [\mathbf{w}_0 \cdots \mathbf{w}_{L_w-1}]$, $\mathbf{U}_k = [\mathbf{u}_k^T \cdots \mathbf{u}_{k-L_w+1}^T]^T$, $\mathbf{H}_0 = I_{L_w} \otimes \mathbf{h}_0$ and $\mathbf{e}_d = [0 \ldots 0\; 1\; 0 \ldots 0]$ with a 1 in position $d{+}1$. Hence $\widehat{s}_k = \underline{\mathbf{w}}\, \mathbf{U}_k$. The optimization in (7) becomes

$$\left\{ \begin{array}{c} \underline{\mathbf{w}}_{L_w,d}^{zf} = \arg\min\limits_{\underline{\mathbf{w}}} \underline{\mathbf{w}}\, \mathbf{R}_{\mathbf{U}} \underline{\mathbf{w}}^T \\ \underline{\mathbf{w}}\, \mathbf{H}_0 = \mathbf{e}_d \end{array} \right. \tag{8}$$

where $\mathbf{R}_{\mathbf{U}}$ is short for $\mathbf{R}_{\mathbf{U}\mathbf{U}} = E\, \mathbf{U}_k \mathbf{U}_k^T$, the covariance matrix of $\mathbf{u}_k$ of (block) size L_w. The optimal postfilter is

$$\underline{\mathbf{w}}_{L_w,d}^{zf} = \mathbf{e}_d \left(\mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1} \mathbf{H}_0\right)^{-1} \mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1} \tag{9}$$

with corresponding optimal

$$\mathrm{SNR}_{L_w,d}^{zf} = \frac{\sigma_s^2}{\mathbf{e}_d \left(\mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1} \mathbf{H}_0\right)^{-1} \mathbf{e}_d^T \;\; - \sigma_s^2} \;\; . \tag{10}$$

The optimal delay (maximum SNR) corresponds to the position of the smallest diagonal element of $\left(\mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1} \mathbf{H}_0\right)^{-1}$.

B. MMSE Design

The MMSE design corresponds to $\underline{\mathbf{w}}_{L_w,d}^{mmse} = \mathbf{R}_{q^{-d_s}\mathbf{U}} \mathbf{R}_{\mathbf{U}\mathbf{U}}^{-1}$. Now $\mathbf{R}_{q^{-d_s}\mathbf{U}} = \mathbf{e}_d\, R_{SS}\, \mathbf{H}_0^T$ where R_{SS} is the source covariance matrix of size L_w, to be constructed using an AR model using the SISO LP filter. Note that $\mathbf{e}_d\, R_{SS}$ means that only row $d{+}1$ of R_{SS} needs to be computed. Hence $\underline{\mathbf{w}}_{L_w,d}^{mmse} = \mathbf{e}_d\, R_{SS}\, \mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1}$ and

$$\mathrm{SNR}_{L_w,d}^{mmse} = \frac{\sigma_s^2}{\mathbf{e}_d\, R_{SS}\, \mathbf{H}_0^T \mathbf{R}_{\mathbf{U}}^{-1} \mathbf{H}_0\, R_{SS}\, \mathbf{e}_d^T} - 1 \;\; . \tag{11}$$

III. Enhancements

Before elaborating on the enhancements, let's consider the details of the basic signal denoising operations.

A. Basic Denoising Operations

First we start with the sample correlations on the basis of the (delay aligned) signal $\{\mathbf{y}_k \,,\; k = 1, \ldots, N_{\mathbf{y}}\}$

$$\begin{array}{l} \widehat{r}_{\mathbf{y}}(n) = \frac{1}{N_{\mathbf{y}}-L_{\mathbf{x}}-L_s} \sum_{k=1}^{N_{\mathbf{y}}-L_{\mathbf{x}}-L_s} \mathbf{y}_{k+n} \mathbf{y}_k^T \;\; , \\ n = 0, 1, \ldots, L_{\mathbf{x}}{+}L_s \end{array} \tag{12}$$

where $L_{\mathbf{x}}$ is the desired order of the MIMO prediction error filter $A_{\mathbf{x}}(q)$, and L_s is the order of the source whitening filter $A_s(q)$. Several variations on (12) are possible, including $\frac{1}{N_{\mathbf{y}}-n} \sum_{k=1}^{N_{\mathbf{y}}-n}$ and the biased estimate $\frac{1}{N_{\mathbf{y}}} \sum_{k=1}^{N_{\mathbf{y}}-n}$. This last choice (called "pre- and post-windowed") guarantees positive semidefiniteness when the sample correlations are put in a block Toeplitz symmetric covariance matrix $\widehat{R}_{\mathbf{Y}}$ of size $N_n \times N_n$ with $[\widehat{r}_{\mathbf{y}}(0)\;\; \widehat{r}_{\mathbf{y}}^T(1) \cdots \widehat{r}_{\mathbf{y}}^T(N_n{-}1)]$ as first block row and $N_n = L_{\mathbf{x}}{+}L_s{+}1$ (note that $\widehat{r}_{\mathbf{y}}(-n) = \widehat{r}_{\mathbf{y}}^T(n)$). Alternatively, $\widehat{R}_{\mathbf{Y}}$ can be obtained directly as

$$\widehat{R}_{\mathbf{Y}} = \frac{1}{N_{\mathbf{y}} - L_{\mathbf{x}} - L_s} \sum_{k=1}^{N_{\mathbf{y}}-L_{\mathbf{x}}-L_s} \mathbf{Y}_{N_n,k}\, \mathbf{Y}_{N_n,k}^T \tag{13}$$

with $\mathbf{Y}_{N_n,k} = [\mathbf{y}_{k-N_n+1}^T \cdots \mathbf{y}_{k-1}^T\;\; \mathbf{y}_k^T]^T$. $\widehat{R}_{\mathbf{Y}}$ can optionally be made block Toeplitz by averaging along block diagonals ("block Toeplitzification").

The sample covariance matrix can be denoised as follows

$$\widehat{R}_{\mathbf{Y}} \;\leftarrow\; \lfloor \widehat{R}_{\mathbf{Y}} - \widehat{\sigma}_v^2 I \rfloor_+ \tag{14}$$

where $\lfloor R \rfloor_+$ denotes the positive semidefinite part of symmetric matrix R, which can be computed by setting either the negative eigenvalues or diagonal values to zero in resp. the eigen decomposition or the LDU decomposition of R. One possible noise variance estimate is $\widehat{\sigma}_v^2 = \lambda_{min}(\widehat{R}_{\mathbf{Y}})$, the minimum eigenvalue of $\widehat{R}_{\mathbf{Y}}$. It is computationally not too complex, only affects $\widehat{r}_{\mathbf{y}}(0)$ in case of block Toeplitz $\widehat{R}_{\mathbf{Y}}$, and allows to avoid the use of $\lfloor . \rfloor_+$, but it underestimates σ_v^2. The better (spatial ML) estimate is to take the arithmetic average of the noise subspace eigenvalues [4], which has dimension $M N_n - (N_n{+}L_h)$ (> 0 assumed). The $\lfloor . \rfloor_+$ operation destroys the block Toeplitz character and hence block Toeplitzification may be performed if a denoised correlation sequence $\widehat{r}_{\mathbf{y}}(n)$ is desired.

The source correlation sequence can now be obtained as

$$\widehat{r}_s(n) = \mathrm{tr}\,\{\widehat{r}_{\mathbf{y}}(n)\} \;,\;\; n = 0, 1, \ldots, L_s \tag{15}$$

where $\mathrm{tr}\,\{.\}$ denotes trace (sum of diagonal elements). From which the (scalar) source prediction error (whitening) filter $A_s(z)$ can be obtained. The source whitened and denoised matrix correlations can then be obtained as

$$\widehat{r}_{\mathbf{x}}(n) = A_s(q)\; A_s^{\dagger}(q)\; \widehat{r}_{\mathbf{y}}(n) \;,\; n = -L_{\mathbf{x}}, \ldots, L_{\mathbf{x}}. \tag{16}$$

The $\widehat{r}_{\mathbf{x}}(n)$ can be put in a block Toeplitz matrix $\widehat{R}_{\mathbf{X}}$. Alternatively, $\widehat{R}_{\mathbf{X}}$ can be obtained from $\widehat{R}_{\mathbf{Y}}$ by filtering left and right with $A_s(z)$. $\widehat{R}_{\mathbf{X}}$ allows the computation of the MIMO prediction error filter $A_{\mathbf{x}}(z)$ and the $M \times M$ prediction error covariance matrix $\Sigma_{\widetilde{\mathbf{x}}}$. Since $\widehat{R}_{\mathbf{X}}$ should be essentially of rank $L_{\mathbf{x}}{+}L_h$, the solution $A_{\mathbf{x}}(z)$ of the normal equations can be obtained either by finding the prediction filter order recursively by the multichannel Levinson algorithm and stop at the order where things get singular. Or regularize before solving: $\widehat{R}_{\mathbf{X}} \leftarrow \widehat{R}_{\mathbf{X}} + \delta\, I$ where $\delta = 10^{-n}\, \lambda_{max}(\widehat{R}_{\mathbf{X}})$ with $n \in [4, 10]$. Finally, $\mathbf{h}_0 = V_{max}(\Sigma_{\widetilde{\mathbf{x}}})$ where $V_{max}(.)$ denotes the eigen vector corresponding to the maximum eigen value.

B. Block Toeplitz Covariance Matrix Enhancement

Here we go back to sample covariance refinements suggested by Cadzow in the eighties [13]. The idea is to iteratively reinforce several structural properties, the reinforcement of which consists of a projection onto a convex set. The iterations then converge to the joint reinforcement of all properties. Theoretically, the matrix valued vector signal spectrum is of the form

$$S_{\mathbf{y}\mathbf{y}}(z) = \mathbf{h}(z)\, S_{ss}(z)\, \mathbf{h}^{\dagger}(z) + S_{\mathbf{v}\mathbf{v}}(z) \tag{17}$$

where $.^{\dagger}$ denotes paraconjugate, and $S_{\mathbf{v}\mathbf{v}}(z) = \sigma_v^2\, I$ is the white noise spectrum. The signal part of the spectrum, $\mathbf{h}(z)\, S_{ss}(z)\, \mathbf{h}^{\dagger}(z)$ is singular, not because of spectral poverty as in the SISO case, but because of limited rank in the matrix dimension. In the SISO case, a stationary signal covariance matrix can only be singular if the signal consists of a number of (complex) sinusoids, with their number being smaller than the covariance matrix dimension. Singularity in the MIMO case has nothing to do with spectral poverty but with matrix singularity of the matrix spectrum at every frequency.

Inspired by [13], (17) suggests the following procedure. First construct a $N \times N$ (blocks) block Toeplitz sample covariance matrix $\widehat{R}_{\mathbf{Y}}$. This can be done in principle either by stacking sample correlation estimates $\widehat{r}_{\mathbf{y}}(n)$ in a symmetric block Toeplitz matrix (simplest) or by block-Toeplitzification of an appropriately sized sample covariance matrix $\widehat{R}_{\mathbf{Y}}$ (more complex). The size N should exceed the sum of L_h and the memory of the source correlations, and should in any case be much larger than L_h (possibly attained by some supplementary zero padding). Then, we would like to use FFT techniques for computational efficiency, which require block circulant matrices. Rather then approximating the block Toeplitz covariance matrix of given dimension by a block circulant matrix, we propose (riminiscent of overlap-save techniques) to embed the $N \times N$ block Toeplitz covariance matrix into a $2N \times 2N$ block circulant matrix of double size, of which the upper-left quarter submatrix is the unmodified block Toeplitz covariance matrix. So we obtain the $2N \times 2N$ block circulant matrix $\overline{\widehat{R}}_{\mathbf{Y}}$ with first block row $[\widehat{r}_{\mathbf{y}}(0)\ \widehat{r}^T_{\mathbf{y}}(1) \cdots \widehat{r}^T_{\mathbf{y}}(N{-}1)\ 0\ \widehat{r}_{\mathbf{y}}(N{-}1) \cdots \widehat{r}_{\mathbf{y}}(1)]$. A block circulant matrix can be block diagonalized by (block) DFT/FFT

$$(F_{2N} \otimes I_M)\, \overline{\widehat{R}}_{\mathbf{Y}}\, (F_{2N}^{-1} \otimes I_M) = \begin{bmatrix} \widehat{S}_{\mathbf{y}}(z_0) & 0 & \cdots & 0 \\ 0 & \widehat{S}_{\mathbf{y}}(z_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{S}_{\mathbf{y}}(z_{2N-1}) \end{bmatrix}$$

where F_{2N} is the DFT matrix of size $2N$, $F_{2N}^{-1} = \frac{1}{2N} F_{2N}^*$ (complex conjugate). This leads directly to

$$\begin{bmatrix} \widehat{S}_{\mathbf{y}}(z_0) \\ \widehat{S}_{\mathbf{y}}(z_1) \\ \vdots \\ \widehat{S}_{\mathbf{y}}(z_{2N-1}) \end{bmatrix} = (F_{2N} \otimes I_M) \begin{bmatrix} \widehat{r}_{\mathbf{y}}(0) \\ \widehat{r}_{\mathbf{y}}(1) \\ \vdots \\ \widehat{r}^T_{\mathbf{y}}(1) \end{bmatrix} \tag{18}$$

where at each FFT frequency bin we get a sample matrix spectrum $\widehat{S}_{\mathbf{y}}(z_n)$, $z_n = e^{j2\pi n/2N}$, $n = 0, \ldots, 2N{-}1$, with the following properties: $\widehat{S}^{\dagger}_{\mathbf{y}}(z_n) = \widehat{S}^H_{\mathbf{y}}(z_n)$ (Hermitian transpose), $\widehat{S}_{\mathbf{y}}(z_n) = \widehat{S}^T_{\mathbf{y}}(z_{2N-n})\,, n = 1, \ldots, N$. Note that the FFTs in (18) can be carried out efficiently in Matlab by reshaping the $2N \times 1$ vectors of $M \times M$ blocks into $2N \times M^2$ matrices.

Now, at each frequency bin n, $S_{\mathbf{y}}(z_n)$ is of the form

$$\begin{aligned} S_{\mathbf{y}}(z_n) &= S_{\mathbf{y},\mathcal{S}}(z_n) + S_{\mathbf{y},\mathcal{N}}(z_n) \\ &= \mathbf{h}(z_n)\, S_s(z_n)\, \mathbf{h}^{\dagger}(z_n) + \sigma_v^2\, I_M \\ &= V_{max,n}\, (\lambda_{max,n} - \sigma_v^2)\, V^H_{max,n} + \sigma_v^2\, I_M \end{aligned} \tag{19}$$

where $S_{\mathbf{y},\mathcal{S}}(z_n)$, $S_{\mathbf{y},\mathcal{N}}(z_n)$ are the signal and noise components of $S_{\mathbf{y}}(z_n)$, and $\lambda_{max,n}$ and $V_{max,n}$ are its maximum eigenvalue and corresponding eigenvector. Now, the $\widehat{S}_{\mathbf{y}}(z_n)$ can be forced to the closest (in Frobenius norm) matrix of the form in (19) by computing its spatial eigen decomposition. Let $\widehat{\lambda}_{1,n} \geq \widehat{\lambda}_{2,n} \geq \cdots \geq \widehat{\lambda}_{M,n}$ be its eigenvalues, hence $\widehat{\lambda}_{max,n} = \widehat{\lambda}_{1,n}$, $\widehat{V}_{max,n} = \widehat{V}_{1,n}$. Then we get $\widehat{S}_{\mathbf{y}}(z_n) = \widehat{S}_{\mathbf{y},\mathcal{S}}(z_n) + \widehat{S}_{\mathbf{y},\mathcal{N}}(z_n) = \widehat{V}_{max,n}\, (\widehat{\lambda}_{max,n} - \widehat{\sigma}_v^2)\, \widehat{V}^H_{max,n} + \widehat{\sigma}_v^2\, I_M$ with $\widehat{\sigma}_v^2 = \frac{1}{2N(M-1)} \sum_{n=0}^{2N-1} \sum_{i=2}^{M} \widehat{\lambda}_{i,n}$ due to the spatiotemporal white noise assumption. Note that in fact at every frequency bin only $\lambda_{max,n}$ and $V_{max,n}$ need to be computed since $\sum_{i=2}^{M} \widehat{\lambda}_{i,n} = \text{tr}\,\{\widehat{S}_{\mathbf{y}}(z_n)\} - \widehat{\lambda}_{max,n}$. Since the noise spectrum $\widehat{S}_{\mathbf{y},\mathcal{N}}(z_n) = \widehat{\sigma}_v^2\, I_M$ is fairly simple, there is no further structure to be imposed. The signal spectrum $\widehat{S}_{\mathbf{y},\mathcal{S}}(z_n) = \widehat{V}_{max,n}\, (\widehat{\lambda}_{max,n} - \widehat{\sigma}_v^2)\, \widehat{V}^H_{max,n}$ on the other hand is supposed to be spectrum of a FIR correlation sequence. This FIR character can be imposed by windowing in the time domain. However, before imposing this FIR character on correlations corresponding to the acoustic channel, the effect of source correlations has to be removed by source linear prediction. The resulting source whitened signal spectrum $\widehat{S}_{\mathbf{y},\mathcal{S}}(z_n)$ then undergoes IFFT to obtain the corresponding matrix correlation sequence. The frequency-wise rank structure enforcement will have destroyed the FIR character of the correlation sequence, which can then simply be enforced in the time domain by proper windowing (without forgetting the symmetry structure of the first block column of the block circulant matrix). The operations of eigen structure enforecement in frequency domain and FIR structure enforcement in the time domain can then be iterated untill convergence. Typically a few iterations suffice. We are now ready to state the following iterative process:

1) Compute the matrix spectrum

$$\begin{bmatrix} \widehat{S}_{\mathbf{y}}(z_0) \\ \widehat{S}_{\mathbf{y}}(z_1) \\ \vdots \\ \widehat{S}_{\mathbf{y}}(z_{2N-1}) \end{bmatrix} = (F_{2N} \otimes I_M) \begin{bmatrix} \widehat{r}_{\mathbf{y}}(0) \\ \widehat{r}_{\mathbf{y}}(1) \\ \vdots \\ \widehat{r}^T_{\mathbf{y}}(1) \end{bmatrix} \tag{20}$$

2) Compute the eigendecomposition of the spectrum $\widehat{S}_{\mathbf{y}}(z_n)$ at each frequency bin $n = 0, 1, \ldots, N$. Determine the noise variance $\widehat{\sigma}_v^2 = \frac{1}{2N(M-1)} \sum_{n=0}^{2N-1} \sum_{i=2}^{M} \widehat{\lambda}_{i,n}$ and the signal part of the spectrum $\widehat{S}_{\mathbf{y},\mathcal{S}}(z_n) = \widehat{V}_{max,n}\, (\widehat{\lambda}_{max,n} - \widehat{\sigma}_v^2)\, \widehat{V}^H_{max,n}$.
3) Determine the source spectrum $\widehat{S}_s(z_n) = \text{tr}\,\{\widehat{S}_{\mathbf{y},\mathcal{S}}(z_n)\}$. Find the source correlations $\widehat{r}_s(n)\,,\ n = 0, 1, \ldots, L_s$ by IFFT. Determine the source AR model $\widehat{A}_s(q)$ (of order L_s) by linear prediction on the source correlations and determine its DFT $\widehat{A}_s(z_n)$ of size $2N$ (via zero padding) Find the source whitened signal spectrum $\widehat{S}_{\mathbf{x}}(z_n) = \widehat{S}_{\mathbf{y}}(z_n)\, |\widehat{A}_s(z_n)|^2$.
4) Compute the acoustic channel correlations

$$\begin{bmatrix} \widehat{r}_{\mathbf{x}}(0) \\ \widehat{r}_{\mathbf{x}}(1) \\ \vdots \\ \widehat{r}^T_{\mathbf{x}}(1) \end{bmatrix} = \frac{1}{2N} (F^*_{2N} \otimes I_M) \begin{bmatrix} \widehat{S}_{\mathbf{x}}(z_0) \\ \widehat{S}_{\mathbf{x}}(z_1) \\ \vdots \\ \widehat{S}_{\mathbf{x}}(z_{2N-1}) \end{bmatrix} \tag{21}$$

Put the correlations outside the range $n \in \{0, 1, \ldots, L_h{-}1\}$ to zero to obtain the transpose of the following block row $[\widehat{r}_{\mathbf{x}}(0)\, \widehat{r}_{\mathbf{x}}^T(1) \cdots \widehat{r}_{\mathbf{x}}^T(L_h{-}1)\; 0\; \cdots\; 0\; \widehat{r}_{\mathbf{x}}(L_h{-}1) \cdots \widehat{r}_{\mathbf{x}}(1)]$.

5) Compute the spectrum of the thus windowed correlation sequence

$$\begin{bmatrix} \widehat{S}_{\mathbf{x}}(z_0) \\ \widehat{S}_{\mathbf{x}}(z_1) \\ \vdots \\ \widehat{S}_{\mathbf{x}}(z_{2N-1}) \end{bmatrix} = (F_{2N} \otimes I_M) \begin{bmatrix} \widehat{r}_{\mathbf{x}}(0) \\ \widehat{r}_{\mathbf{x}}(1) \\ \vdots \\ \widehat{r}_{\mathbf{x}}^T(1) \end{bmatrix} \tag{22}$$

Then reconstruct the total signal spectrum as

$$\widehat{S}_{\mathbf{y}}(z_n) = \left\lfloor \frac{\widehat{S}_{\mathbf{x}}(z_n)}{|\widehat{A}_s(z_n)|^2} + \widehat{\sigma}_v^2 I_M \right\rfloor_+ \tag{23}$$

Go back to step 2 untill convergence.

After convergence, the MIMO linear predictor $A_{\mathbf{x}}(q)$ can be determined from the source whitened correlations $\widehat{r}_{\mathbf{x}}(n)$, $n = 0, 1, \ldots, L_x$. We recommend to use the minimal prediction order $L_{\mathbf{x}} = \left\lceil \frac{L_h}{M-} \right\rceil$, and to introduce a minimum of regularization in the normal equations (ideally the MIMO linear predictor should be solved by the multichannel version of the Levinson algorithm so that the proper order (singularity of the MIMO prediction error covariance) can be detected). The refinements of this section are implemented for the D&P Equalizer approaches mentioned in the simulations below.

C. Iterative LP Refinement

So far the source spectrum has been recovered from the output spectrum by assuming that the SIMO reverberation filter is approximately a paraunitary filter. This hypothesis can be taken as an initialization for an iterative process. Using a (MMSE-)ZF design, the source statistics can be reconstructed at the output of the D&P equalizer with denoised measured signal statistics as input. SISO LP can be performed on the resulting source correlations, and then the MIMO LP can be reiterated. This refinement has not been implemented yet in the simulations below.

IV. Experimental Results

MMSE-ZF postfiltering for robust dereverberation in noisy environment

We illustrate the behavior of zero-forcing post-processing, and we provide a comparison with the classic Delay-&-Predict equalizer. We consider a rectangular room with dimensions $L_x = 8\ m$, $L_y = 10\ m$ and $L_z = 4\ m$, and with wall reflection coefficients $\rho_x = \rho_y = \rho_z = 0.9$ ($T_{60} = 250\ ms$). A speech signal with duration of 8.8s, and sampled at 8 kHz is used as the original source signal. The reverberant speech signal is observed on 2 distinct microphones. A computer implementation (graciously provided by Geert Rombouts while at K.U. Leuven) of the image method as described in [11] is used to generate synthetic room impulse responses for the microphones. We constrain the postfilter length (and hence the equalization delay d) to $L_w \leq 100$ ($d \leq 12.5\ ms$). The optimal delay (maximizing (10)) is selected. Figure 1 plots the Signal-to-Echo+Noise Ratio (SENR $= \dfrac{\sum_k s_k^2}{\min\limits_{\alpha} \sum\limits_k (s_k - \alpha\, \widehat{s}_k)^2}$) as a function of the input Signal-to-Noise Ratio (SNR $= \dfrac{\sum_k \|\mathbf{y}_k - \mathbf{v}_k\|^2}{\sum_k \|\mathbf{v}_k\|^2}$). Note that α is introduced because the source can only be reconstructed up to a scale factor. The curves show that, in all regions, the MMSE-ZF D-&-P performs better than both the classic D-&-P and D-&-S. Particularly in a noisy environment, the postfiltering becomes essential in order to have acceptable enhancement accuracy. On the other hand, one can also remark that the post-processing still has a positive effect even in absence of ambient noise (SNR=60 dB). The reason is that the postfiltering also compensates for the errors in the estimation of the source spectrum (the estimation is done by averaging only two observation spectra ($M = 2$)).

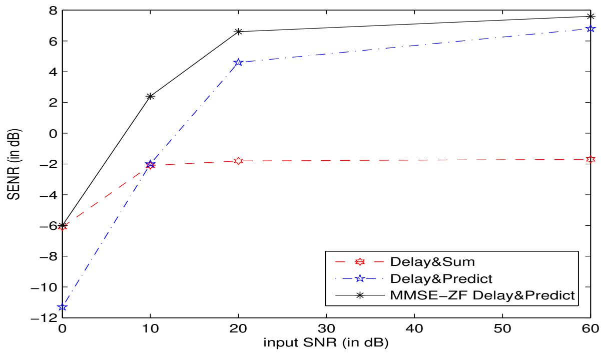

Fig. 1. The SENR function of the input SNR.

V. Conclusions

In this paper, we have introduced robust Delay-&-Predict equalization for blind SIMO dereverberation. We have optimized the transformation of the multivariate prediction filter to a longer equalizer using the MSE criterion. The optimization is performed with or without zero-forcing constraints, leading respectively to MMSE-ZF and MMSE designs. The filter length increase allows for the introduction of some equalization delay, that can also be optimized. Experimental results illustrate that considerable gains can be achieved by allowing for a small equalization delay. It has also been shown that the post-processing is crucial in the low SNR region. In this paper we have also introduced some refinements for the multivariate linear prediction (LP) step, the crucial ingredient in SIMO dereverberation. A first refinement corresponds to a computationally efficient (FFT based) singular block Toeplitz covariance matrix enhancement that enforces the SIMO filtered source plus white noise structure before applying MIMO LP. A second suggested refinement is an iterative refinement between SISO and MIMO LP. In future work we plan to further emphasize the computational efficiency of the first refinement introduced here, and to investigate more the performance enhamcement brought about by the second refinement.

Acknowledgment

Eurecom's research is partially supported by its industrial members: BMW, Bouygues Télécom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, STMicroelectronics, Swisscom, Thales. The work reported herein was also supported by The European FP6 NoE project K-Space.

References

[1] A. Assa-El-Bey, K. Abed-Meraim, and Y. Grenier. "Blind Separation of Audio Sources Convolutive Mixtures Using Parametric Decomposition," *In Proc of IWAENC*, Sept. 2005.

[2] A.J. van der Veen, S. Talwar, A. Paulraj. "A Subspace Approach to Blind Space-Time Signal Processing for Wireless Communication Systems," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.45, pp. 173-190, Jan. 1997.

[3] D.T.M. Slock. "Blind Fractionally-Spaced Equalization, Perfect-Reconstruction Filter Banks and Multichannel Linear Prediction," *In Proc. of IEEE ICASSP*, Apr. 1994.

[4] C.B. Papadias, and D.T.M. Slock. "Fractionally Spaced Equalization of Linear Polyphase Channels and Related Blind Techniques Based on Multichannel Linear Prediction," *IEEE Trans. on Signal Processing*, pp.641-654, Mar. 1999.

[5] M. Triki and D.T.M. Slock. "Blind Dereverberation of Quasi-periodic Sources Based on Multichannel Linear Prediction," *In Proc. of IWAENC*, Sept. 2005.

[6] M. Triki and D.T.M. Slock. "Delay and Predict Equalization For Blind Speech Dereverberation," *In Proc. of IEEE ICASSP*, Vol.5, pp.97-100, May 2006.

[7] M. Triki and D.T.M. Slock. "Multivariate LP Based MMSE-ZF Equalizer Design Considerations and Application to MultiMicrophone Dereverberation," *In Proc. of IEEE ICASSP*, Apr. 2007.

[8] M. Delcroix, T. Hikichi, M. Miyoshi. "Precise Dereverberation Using Multichannel Linear Prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, Feb. 2007.

[9] H. Buchner, R. Aichner, W. Kellermann. "Blind Source Separation for Convolutive Mixtures: a Unified Treatment," in J. Benesty, Y. Huang (Eds.), *Audio Signal Processing for Next-generation Multimedia Communication Systems*, pp. 255-293, Kluwer Academic Publishers, Boston, MA, 2004.

[10] H. Buchner, R. Aichner, W. Kellermann. "A Generalization of Blind Source Separation Algorithms for Convolutive Mixtures based on Second-Order Statistics," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 120-134, Jan. 2005.

[11] P.M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, pp.1527-1529, Nov. 1986.

[12] A. Koul, J.E. Greenberg, "Using Intermicrophone Correlation to Detect Speech in Spatially Separated Noise," *EURASIP Journal on Applied Signal Processing*, Issue 12, 2006.

[13] J. A. Cadzow, "Signal Enhancement – A Composite Property Mapping Algorithm," *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 36, No. 1, Jan. 1988.