Speech Dereverberation by Blind Adaptive MIMO Filtering Exploiting Nongaussianity, Nonwhiteness, and Nonstationarity

Herbert Buchner[1] *and Walter Kellermann*[2]

[1] Deutsche Telekom Laboratories, Berlin University of Technology, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

[2] Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany
E-Mail: herbert.buchner@telekom.de, wk@LNT.de

Abstract

In this paper, we present a class of novel algorithms for blind dereverberation of speech signals based on TRINICON, a general framework for broadband adaptive MIMO signal processing. In order to exploit all fundamental stochastic signal properties of speech for the dereverberation/deconvolution process and to avoid any whitening artifacts known from previous approaches, we propose the incorporation of a specially designed signal model based on an expansion using multivariate Chebyshev-Hermite polynomials. The multivariate model also inherently includes linear prediction which is known to be related directly to the human vocal tract model. The framework is applicable to both single-speaker scenarios and also to multiple simultaneously active speakers. In the latter case it also includes blind source separation in addition to the dereverberation.

1 Introduction

In broadband signal acquisition by sensor arrays, such as in hands-free speech communication systems, the original source signals $s_q(n)$, $q = 1,\ldots,Q$ are filtered by a linear multiple input and multiple output (MIMO) system, e.g., the reverberant room, before they are captured as sensor signals $x_p(n)$, $p = 1,\ldots,P$. In this paper, we describe this MIMO mixing system by FIR filter models, where $h_{qp,\kappa}$, $\kappa = 0,\ldots,M-1$ denote the coefficients of the model from the q-th source signal $s_q(n)$ to the p-th sensor signal $x_p(n)$ according to Fig. 1. Moreover, we assume throughout this paper that $Q \leq P$ which are known as the *overdetermined* and *determined* cases, respectively. Note that in general, the sources

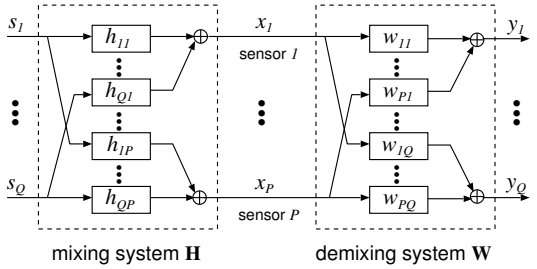

Figure 1: Setup for blind MIMO signal processing.

$s_q(n)$ may or may not be all simultaneously active at a particular instant of time. We are interested in finding an optimum length-L FIR "demixing" system with coefficients $w_{pq,\kappa}$ by adaptive signal processing to yield the output signals $y_q(n)$. Since in practical acoustic scenarios, neither the original source signals $s_q(n)$ nor the mixing system are directly accessible, the adaptation of the demixing system has to be performed *blindly*.

Based on this MIMO structure, we may identify the following blind signal processing problems for enhancing the output signals $y_q(n)$:
(a) Signal separation (noise/interferers): Cancel out all cross-channels of the cascaded mixing-demixing system.
(b) Deconvolution/Dereverberation: In addition to the separation, acquire "dry" sources up to a delay and a scaling factor.

In contrast to the separation problem (a), which is closely related to blind system identification [2], the by far more difficult problem (b) also requires an *inversion* of the mixing system (which typically contains nonminimum-phase impulse responses). It can be shown that an ideal signal-independent broadband separation solution exists for $Q \leq P$ [2], while for the ideal solution of the inverse problem (b) the overdetermined case $Q < P$ is required [4]. This paper focuses on the dereverberation of speech signals (b) using TRINICON ('TRIple-N ICA for CONvolutive mixtures'), a generic concept for broadband adaptive MIMO filtering [7, 5] based on the technique of independent component analysis (ICA), e.g., [1].

2 Adaptive MIMO Signal Processing based on TRINICON

In this section we first give a brief overview of the essential elements of TRINICON for the coefficient adaptation. Thereby, we restrict the presentation here to simple gradient-based coefficient updates in the time domain.

2.1 Optimization Criterion

Various approaches exist to estimate the demixing matrix $\mathbf{W}$ by utilizing the following fundamental source signal properties [1] which were all combined in TRINICON:
(i) Nongaussianity is exploited by using higher-order statistics for ICA. The minimization of the mutual information (MMI) among the output channels can be regarded as the most general approach to separation problems [1]. To obtain an estimator that is also suitable for the inverse problem, we use the Kullback-Leibler divergence (KLD) [9] between a certain *desired* joint pdf (essentially representing a hypothesized stochastic source model as shown below) and the joint pdf of the actually estimated output signals.
(ii) Nonwhiteness is exploited by simultaneous minimization of output cross-relations over multiple time-lags. We therefore consider multivariate pdfs, i.e., 'densities covering D time-lags'.
(iii) Nonstationarity is exploited by simultaneous minimization of output cross-relations at different time-instants. We assume ergodicity within blocks of length N so that the ensemble average is replaced by time averages over these blocks.

Throughout this section, we formulate the framework for $Q = P$ without loss of generality. In practice, the current number of simultaneously active sources is allowed to vary throughout the application and only the conditions $Q \leq P$ (for separation only) and $Q < P$ (for deconvolution), respectively, have to be fulfilled.

To introduce an algorithm for broadband processing of convolutive mixtures, we first formulate the convolution of the FIR demixing system of length L in the following matrix form [5]:

$$\mathbf{y}(n) = \mathbf{W}^{\mathrm{T}}\mathbf{x}(n), \tag{1}$$

where n denotes the time index, and

$$\mathbf{x}(n) = [\mathbf{x}_1^{\mathrm{T}}(n),\ldots,\mathbf{x}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{2}$$

$$\mathbf{y}(n) = [\mathbf{y}_1^{\mathrm{T}}(n),\ldots,\mathbf{y}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, \tag{3}$$

$$\mathbf{x}_p(n) = [x_p(n),\ldots,x_p(n-2L+1)]^{\mathrm{T}}, \tag{4}$$

$$\mathbf{y}_q(n) = [y_q(n),\ldots,y_q(n-D+1)]^{\mathrm{T}}. \tag{5}$$

The parameter D in (5), $1 \leq D < L$, denotes the number of time lags taken into account to exploit the nonwhiteness of the source signals as shown below. $\mathbf{W}_{pq}$, $p = 1,\ldots,P$, $q = 1,\ldots,P$ denote $2L \times D$ *Sylvester matrices* that contain all coefficients

of the respective filters in each column by successive shifting, i.e., the first column reads $\left[\mathbf{w}_{pq}^{\mathrm{T}},0,\ldots,0\right]^{\mathrm{T}}$, the second column $\left[0,\mathbf{w}_{pq}^{\mathrm{T}},0,\ldots,0\right]^{\mathrm{T}}$, etc. Finally, the $2PL\times PD$ matrix $\mathbf{W}$ combines all Sylvester matrices $\mathbf{W}_{pq}$.

Based on the KLD, the following cost function was introduced in [5] taking into account all three fundamental signal properties (i)-(iii):

$$\mathcal{J}(m,\mathbf{W}) \;=\; -\sum_{i=0}^{\infty}\beta(i,m)\frac{1}{N} \cdot \sum_{j=iN_L}^{iN_L+N-1}\left\{\log(\hat{p}_{s,PD}(\mathbf{y}(j)))-\log(\hat{p}_{y,PD}(\mathbf{y}(j)))\right\}, \qquad (6)$$

where $\hat{p}_{s,PD}(\cdot)$ and $\hat{p}_{y,PD}(\cdot)$ are assumed or estimated PD-variate source model (i.e., desired) pdf and output pdf, respectively. The index m denotes the block time index for a block of N output samples shifted by L samples relatively to the previous block. Furthermore, β is a window function allowing for online, offline, or block-online algorithms [7].

2.2 Gradient-Based Coefficient Update

For brevity and simplicity we concentrate in this subsection on iterative Euclidean gradient-based block-online coefficient updates which can be written in the general form

$$\breve{\mathbf{W}}^{0}(m) \;:=\; \breve{\mathbf{W}}(m-1), \qquad (7a)$$

$$\breve{\mathbf{W}}^{\ell}(m) \;=\; \breve{\mathbf{W}}^{\ell-1}(m)-\mu\Delta\breve{\mathbf{W}}^{\ell}(m),\ \ell=1,\ldots,\ell_{\max}, \qquad (7b)$$

$$\breve{\mathbf{W}}(m) \;:=\; \breve{\mathbf{W}}^{\ell_{\max}}(m), \qquad (7c)$$

where μ is a stepsize parameter, and the superscript index ℓ denotes an iteration parameter to allow for multiple iterations ($\ell=1,\ldots,\ell_{\max}$) within each block m. The downwards pointing hat symbol on top of $\mathbf{W}$ in (7) serves to distinguish the *condensed* $PL\times Q$ demixing coefficient matrix $\breve{\mathbf{W}}$ to be optimized, from the corresponding larger Sylvester matrix $\mathbf{W}$ in the cost function. The matrix $\breve{\mathbf{W}}$ consists of the first column of each submatrix $\mathbf{W}_{pq}$ without the L zeros.

Obviously, when calculating the gradient of $\mathcal{J}(m,\mathbf{W})$ w.r.t. $\breve{\mathbf{W}}$ explicitly, we are confronted with the problem of the different matrix formulations $\mathbf{W}$ and $\breve{\mathbf{W}}$. The larger dimensions of $\mathbf{W}$ are a direct consequence of taking into account the nonwhiteness signal property by choosing $D>1$. The rigorous distinction between these different matrix structures is also an essential aspect of the general TRINICON framework and leads to an important building block whose actual implementation is fundamental to the properties of the resulting algorithm, the so-called *Sylvester constraint* ($\mathcal{SC}$) on the coefficient update, formally introduced in [7]. Using the Sylvester constraint operator the gradient descent update can be written as

$$\Delta\breve{\mathbf{W}}^{\ell}(m)=\mathcal{SC}\left\{\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})\right\}|_{\mathbf{W}=\mathbf{W}^{\ell}(m)}. \qquad (8)$$

Depending on the particular realization of ($\mathcal{SC}$), we are able to select both, well known and also novel improved adaptation algorithms [10]. In [2] an explicit formulation of a *generic* Sylvester constraint was derived to further formalize and clarify this concept.

It can be shown [3] that by taking the gradient of $\mathcal{J}(m)$ with respect to the demixing filter matrix $\breve{\mathbf{W}}(m)$ according to (8), we obtain the following generic gradient descent-based TRINICON update rule:

$$\Delta\breve{\mathbf{W}}^{\ell}(m)=\frac{1}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\left[\mathbf{x}(j)\Phi_{s,PD}^{\mathrm{T}}(\mathbf{y}(j))-\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\right)^{+}\right]\right\}, \qquad (9a)$$

with $\cdot^{+}$ denoting the pseudoinverse of a matrix, and with the generalized score function

$$\Phi_{s,PD}(\mathbf{y}(j)) \;=\; -\frac{\partial\log\hat{p}_{s,PD}(\mathbf{y}(j))}{\partial\mathbf{y}(j)} -\frac{1}{N}\sum_{r}\sum_{i_1,i_2,\ldots}\frac{\partial\mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}}{\partial\mathbf{y}}\sum_{j=iN_L}^{iN_L+N-1}\frac{\partial\hat{p}_{s,PD}}{\partial\mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}} \qquad (9b)$$

resulting from the hypothesized source model $\hat{p}_{s,PD}=\hat{p}_{s,PD}(\mathbf{y},\boldsymbol{\mathcal{Q}}_s^{(1)},\boldsymbol{\mathcal{Q}}_s^{(2)},\ldots)$ with certain stochastic model parameters $\boldsymbol{\mathcal{Q}}_s^{(r)}$, $r=1,2,\ldots$ (the calligraphic symbols denote multidimensional arrays) given by their elements $\mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}$ in the generic form $\mathcal{Q}_{s,i_1,i_2,\ldots}^{(r)}(i)=\frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}(\mathbf{y}(j))\right\}$ with certain nonlinear functions $\mathcal{G}_{s,i_1,i_2,\ldots}^{(r)}(\mathbf{y})$, $r=1,2,\ldots$. A well known special case of such a parameterization is the estimate of the correlation matrix $\mathbf{R}_{\mathbf{yy}}(i)=\frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\mathbf{y}(j)\mathbf{y}^{\mathrm{T}}(j)\right\}$. The filter coefficients and the stochastic model parameters are estimated in an alternating way.

3 Application of TRINICON to Blind Dereverberation

The hypothesized source model $\hat{p}_{s,PD}(\cdot)$ in (9b) is chosen according to the class of signal processing problem to be solved. For instance, a factorization of $\hat{p}_{s,PD}(\cdot)$ among the sources yields BSS [7, 5], i.e.,

$$\hat{p}_{s,PD}(\mathbf{y}(j)) \overset{\text{(BSS)}}{=} \prod_{q=1}^{P}\hat{p}_{y_q,D}(\mathbf{y}_q(j)), \qquad (10)$$

while a complete factorization w.r.t. the PD dimensions of the multivariate pdf $\hat{p}_{s,PD}(\mathbf{y}(j))$ leads to the traditional so-called multichannel blind deconvolution (MCBD) approach, i.e., traditionally, ICA-based MCBD algorithms assume *i.i.d. source models*, e.g., [6]. In other words, in addition to the separation of statistically independent sources, MCBD algorithms also temporally *whiten* the output signals, so that this approach is not directly suitable for speech and audio signals.

Signal sources which are non i.i.d. should not become i.i.d. at the output of the blind adaptive filtering stage. Therefore, their statistical dependencies should be preserved. In other words, the adaptation algorithm has to distinguish between the statistical dependencies within the source signals, and the statistical dependencies introduced by the mixing system $\breve{\mathbf{H}}$, i.e., the reverberant room, so that only the influence of the room acoustics is minimized. We denote the corresponding generalization of the traditional MCBD technique as *MultiChannel Blind Partial Deconvolution* (MCBPD) [5, 3]. Equations (9) inherently contain a statistical source model (signal properties (i)-(iii) in Sect. 2), expressed by the multivariate densities, and thus provide all necessary requirements for the MCBPD approach.

For the distinction between the production system of the source signals and the room acoustics we can again exploit all three fundamental signal properties already mentioned in Sect. 2:

(i) Nonwhiteness. The auto-correlation structure of the speech signals can be taken into account. While the room acoustics influences all off-diagonals of the $PD\times PD$ output correlation matrix $\mathbf{R}_{\mathbf{yy}}$, the effect of the vocal tract is concentrated in the first few off-diagonals around the main diagonal. In the simplest case, these first off-diagonals of $\mathbf{R}_{\mathbf{yy}}$ are now taken over into the *banded* desired correlation matrix as suggested in [5]. Note that there is a close link to linear prediction techniques which gives guidelines for the number of lags to be preserved [3].

(ii) Nonstationarity. The speech production system and the room acoustics also differ in their time-variance after [5]. While the room acoustics is assumed to be constant during the adaptation process, the speech signal is only short-time stationary [12], modeled by the time-varying speech production

model. Typically, the duration of the stationarity intervals is assumed to be approximately 20ms [12]. We therefore adjust the block length N and in practice preferably also the block shift N_L in the criterion (6) with the model parameter estimates $\mathcal{Q}_s^{(r)}(i)$ and in the corresponding coefficient updates (9) to the assumed duration of the stationarity interval.

(iii) Nongaussianity. Speech is a well-known example for supergaussian signals. Due to a convolutive sum – describing in our application the filtering by the room acoustics – the pdfs of the recorded microphone signals tend to be somewhat closer to Gaussians. Hence, another strategy is to maximize the nongaussianity of the output signals of the demixing system (as far as possible by the MIMO FIR filters), e.g., [13, 14, 15]. This strategy is addressed, e.g., using the *kurtosis* $\hat{\kappa}_{4,y_q} = \hat{E}\left\{y_q^4\right\} - 3\hat{\sigma}_{y_q}^4$ as a widely-used distance measure of nongaussianity ($\hat{\kappa}_{4,y_q} > 0$ indicates a supergaussian pdf and $\hat{\kappa}_{4,y_q} < 0$ a subgaussian pdf).

The multivariate stochastic speech signal model proposed in the following section precisely takes into account all of these properties.

4 A Multivariate Signal Model for TRINICON-Based Dereverberation

Two different expansions are commonly used to obtain a parameterized representation of probability density functions which only slightly deviate from the Gaussian density (often called *nearly Gaussian densities*): the Edgeworth and the Gram-Charlier expansions, e.g., [1]. They lead to very similar approximations, so we only consider here the Gram-Charlier expansion. These expansions are based on the so-called Chebyshev-Hermite polynomials $P_{\mathrm{H},n}(\cdot)$. An advantage of this representation is that the corresponding expansion coefficients can be related directly to known stochastic quantities based on higher-order cumulants, such as the kurtosis mentioned in the previous section.

To obtain general coefficient update rules based on this representation, we consider a multivariate generalization of the Gram-Charlier expansion. As detailed in [3], it can be expressed as

$$\hat{p}_{y_q,D}(\mathbf{y}_q(j)) = \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i)}} \mathrm{e}^{-\frac{1}{2}\mathbf{y}_q^{\mathrm{T}}(j)\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i)\mathbf{y}_q(j)} \cdot \sum_{n_1=0}^{\infty}\cdots\sum_{n_D=0}^{\infty} a_{n_1\cdots n_D,p} \prod_{d=1}^{D} P_{\mathrm{H},n_d}\left(\left[\mathbf{L}_q^{-1}(i)\mathbf{y}_q(j)\right]_d\right)$$

with the expansion coefficients

$$a_{n_1\cdots n_D,q} = \hat{E}\left\{\prod_{d=1}^{D}\frac{1}{n_d!}P_{\mathrm{H},n_d}\left(\left[\mathbf{L}_q^{-1}(i)\mathbf{y}_q(j)\right]_d\right)\right\},$$

where $\mathbf{L}_q$ is obtained by the Cholesky decomposition $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q} = \mathbf{L}_q^{\mathrm{T}}\mathbf{L}_q$ (note that $\sqrt{\mathbf{y}_q^{\mathrm{T}}\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}\mathbf{y}_q} = \|\mathbf{L}_q^{-1}\mathbf{y}_q\|_2$).

In this paper, we further consider an important special case of this general multivariate model, which is particularly useful for speech processing. In this case, the inverse covariance matrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1} = (\mathbf{L}_q^{\mathrm{T}}\mathbf{L}_q)^{-1}$ is first factorized as [11]

$$\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}^{-1}(i) = \mathbf{A}_q(i)\Sigma_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}^{-1}(i)\mathbf{A}_q^{\mathrm{T}}(i), \tag{12}$$

where $\mathbf{A}_q(i)$ and $\Sigma_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(i)$ denote a $D\times D$ unit lower triangular matrix (i.e., its elements on the main diagonal are equal to 1) and a diagonal matrix, respectively [11]. The $D\times D$ unit lower triangular matrix $\mathbf{A}_q(i)$ can be interpreted as a (time-varying) convolution matrix of a whitening filter. It is therefore convenient for computational reasons to model the signal y_q as an autoregressive (AR) process with time-varying AR coefficients $a_{q,k}(n)$, and residual signal $\tilde{y}_q(n)$. Formally, the above-mentioned exploitation of the nonwhiteness to distinguish between the coloration of the sources and the mixing system is achieved by decoupling the order of the AR process, i.e., the prediction order $0 \le n_A \le D-1$, from the dimension D of the correlation matrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$, i.e.,

$$y_q(n) = -\sum_{k=1}^{n_A} a_{q,k}(n)y_q(n-k) + \tilde{y}_q(n). \tag{13}$$

The matrices $\mathbf{A}_q$ and $\Sigma_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}$ are then obtained by the D column vectors $\left[1, a_{q,1}(n), a_{q,2}(n), \cdots, a_{q,n_A}(n), 0, \cdots, 0\right]^{\mathrm{T}}$, $\left[0, 1, a_{q,1}(n-1), \cdots, a_{q,n_A-1}(n-1), a_{q,n_A}(n-1), \cdots, 0\right]^{\mathrm{T}}$, etc., and

$$\Sigma_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q} = \mathrm{Diag}\left\{\hat{\sigma}_{\tilde{y}_q}^2(n), \ldots, \hat{\sigma}_{\tilde{y}_q}^2(n-D+1)\right\}. \tag{14}$$

Now, the multivariate stochastic signal model can be rewritten by shifting the *prefiltering matrix* $\mathbf{A}_q$ into the data terms, i.e.,

$$\tilde{\mathbf{y}}_q := \mathbf{A}_q^{\mathrm{T}}\mathbf{y}_q = [\tilde{y}_q(n), \tilde{y}_q(n-1), \ldots, \tilde{y}_q(n-D+1)]^{\mathrm{T}}. \tag{15}$$

Moreover, by assuming the whitened elements of vector $\tilde{\mathbf{y}}_q$ to be i.i.d. (which in practice is a widely used assumption in AR modeling), so that the expansion coefficients $a_{n_1\cdots n_D,q}$ are factorized, we obtain with $\mathbf{L}_q(i) = \mathrm{Diag}\left\{\frac{1}{\hat{\sigma}_{\tilde{y}_q}(j)}, \ldots, \frac{1}{\hat{\sigma}_{\tilde{y}_q}(j-D+1)}\right\}\mathbf{A}_q^{\mathrm{T}}(i)$ and (15) a more compact model representation. The corresponding fourth-order approximation of a zero-mean and nearly Gaussian pdf directly contains the known quantities *skewness* and *kurtosis*, the latter one being the most important higher-order statistical quantity in our context, as mentioned above. Generally, speech signals exhibit supergaussian densities whose third-order cumulants are negligible compared to its fourth-order cumulants. Hence, by considering only the fourth-order term in addition to SOS, we obtain

$$\hat{p}_{y_q,D}(\mathbf{y}_q(j)) = \prod_{d=1}^{D}\frac{1}{\sqrt{2\pi\,\hat{\sigma}_{\tilde{y}_q}^2(j-d+1)}}\mathrm{e}^{-\frac{\tilde{y}_q^2(j-d+1)}{2\hat{\sigma}_{\tilde{y}_q}^2(j-d+1)}} \cdot\left(1+\frac{\hat{\kappa}_{4,\tilde{y}_q}}{4!\sigma_{\tilde{y}_q}^4(j-d+1)}P_{\mathrm{H},n_d}\left(\frac{\tilde{y}_q(j-d+1)}{\hat{\sigma}_{\tilde{y}_q}(j-d+1)}\right)\right).$$

By exploiting the near-gaussianity using the approximation $\log(1+\varepsilon)\approx\varepsilon$ in the logarithmized respresentation of this pdf in (9b), and noting that $P_{\mathrm{H},4}\left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right) = \left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right)^4 - 6\left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right)^2 + 3$, we obtain after a straightforward calculation the corresponding TRINICON coefficient update rule based on (9). An efficient realization which still exploits all three fundamental signal properties (i)-(iii), as discussed above, is obtained using the so-called *correlation method*, i.e., by assuming a global nonstationarity of the source signals but short-time stationarity in each block as known from linear prediction [12]. Using the explicit formulation of the generic Sylvester constraint after [2], these steps finally lead to the MIMO coefficient update rule [3]

$$\begin{aligned}\breve{\mathbf{w}}_{pq}^{\ell}(m) = \breve{\mathbf{w}}_{pq}^{\ell-1}(m) - \frac{\mu}{N}\sum_{i=0}^{\infty}\beta'(i,m)\Bigg[&\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\breve{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q(j)}{2\hat{\sigma}_{\tilde{y}_{q,i}}^2} \\ &-\left(\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\tilde{y}_q^4(j)}{3\hat{\sigma}_{\tilde{y}_{q,i}}^4}-1\right)\cdot\Bigg(\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\breve{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q^3(j)}{\hat{\sigma}_{\tilde{y}_{q,i}}^4} \\ &-\frac{\sum_{j=iN_L'}^{iN_L'+N-1}\breve{\mathbf{x}}_p^{(q)}(j)\tilde{y}_q(j)\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}_q^4(j)}{\hat{\sigma}_{\tilde{y}_{q,i}}^6}\Bigg)\Bigg] \\ &+\mu\sum_{i=0}^{\infty}\beta(i,m)\left[\mathcal{SC}\left\{\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\right)^{+}\right\}\right]_{pq},\end{aligned} \tag{16}$$

$p=1,\ldots,P,\ q=1,\ldots,P$. Analogously to the definition (15), the symbol $\breve{\mathbf{x}}_p^{(q)}$ denotes a column vector of filtered sensor signals x_p according to the prefiltering matrix $\mathbf{A}_q$ introduced above.

In other words, this update rule can be interpreted as a so-called *filtered-x*-type algorithm since both the input (i.e., microphone) signal vector and the output signals appear as filtered versions in the update. As a consequence, we immediately obtain

Fig. 2. While $\mathbf{W}$ ideally inverts the room acoustic mixing system $\mathbf{H}$, the (set of) linear prediction filter(s) $\mathbf{A}$ from the stochastic source model ideally inverts the (set of) speech production system(s) of the source(s). The coefficients $\mathbf{W}$ and $\mathbf{A}$ are estimated in an alternating fashion like the estimation of the other stochastic model parameters, as mentioned above. Note that (in accordance with the known filtered-x concept) the filtered input vector $\breve{\mathbf{x}}_p^{(q)}$ is obtained using the filter coefficients from the linear prediction (LP) analysis of the *output* signals y_p, i.e., the coefficients of the output LP analysis filters are copied to the input transformation filters.

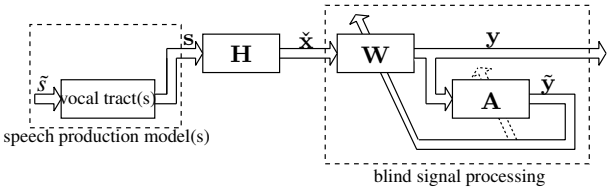

Figure 2: Inversion of the speech production models within the blind signal processing and filtered-x-type interpretation.

5 Experiments

The experiments have been conducted using $Q = 2$ speech signals (one male speaker and one female speaker) convolved with measured impulse responses of a real room with a reverberation time $T_{60} \approx 700$ms and a sampling frequency of 16kHz. A linear four-element microphone array ($P = 4$) with an inter-element spacing of 16cm was used. The speech signals arrived from $\pm 24^o$ relative to the normal plane of the array axis and the distance between the speakers and the center of the microphone array was 165cm.

As a performance measure for the evaluation of the dereverberation performance, the *signal-to-reverberation ratio* (SRR) measures the power ratio between the direct sound and the contribution by the reverberation. For speech signals, the first 50ms after the main peak of the impulse responses are also added to the contribution of the direct path (critical delay time n_{50}, which is known to contribute to the speech intelligibility). For a higher correlation to the quality perceived by auditory measurements, we study in our experiments the improvement of the so-called *segmental SRR*. In the case of multiple source signals, the performance of the additional source separation ability of MCBPD is evaluated by the improvement of the so-called (segmental) *signal-to-interference ratio* (SIR) at the output analogously to the (segmental) SRR.

Our simulations are based on the coefficient update (16) using the correlation method. We chose $L = 3000$, the block length $N = N'_L = 320$ corresponding to a stationarity interval of 20ms, and $n_A = 32$. Figure 3 shows both the SIR improvement (i.e.,

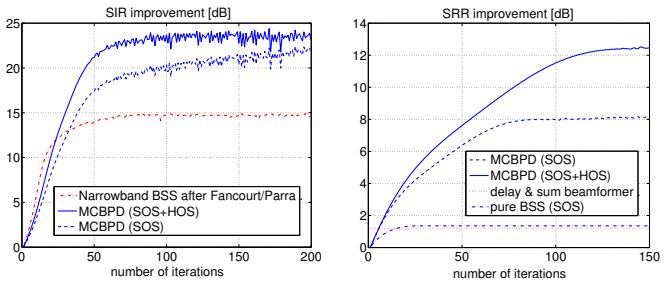

Figure 3: SIR and SRR performance of MIMO-based MCBPD.

source separation at the ouputs) and the SRR improvement for offline (batch) adaptation, i.e., $\beta(i,m) = \beta(i)$ in (6) (and thus $\beta'(i,m) = \beta'(i)$ in (16)) corresponds to a rectangular window function over the entire available signal length, and the outer sum in (6) and (16) turns into a summation of the contributions from all blocks with equal weights. The SIR and SRR curves were averaged between the contributions from the two sources. We see that the optimization based purely on second-order statistics (SOS, only the first term in the brackets in (16) was used) exhibits a rapid initial convergence. Further considerations [3] show that the approach purely based on the kurtosis (only the second term in the brackets in (16)) finally achieves a higher level of SRR improvement at the cost of a slower initial convergence. By exploiting all the available statistical signal properties (SOS+HOS, both terms in the brackets in (16) were used), the TRINICON framework combines the advantages of the former two approaches. These synergies can be seen in both the separation and the dereverberation performances. As reference, we also included the SIR convergence curve of the popular narrowband BSS algorithm after Fancourt and Parra [8] which is based on SOS. The reference curve for a pure separation algorithm [7] based on SOS (as a special case of (16) with $n_A = L-1$, $N = L$, and using only the first term in the brackets) in the SRR plot, and the comparison with a conventional delay-and-sum beamformer confirms the high efficiency of the MCBPD extension presented in this paper.

6 Conclusions

Based on the TRINICON framework for broadband adaptive MIMO filtering we developed in this paper a class of novel algorithms for the problem of blind dereverberation. Due to the design of the stochastic source model specifically for speech signals, we effectively exploit the nonwhiteness, the nonstationarity, and the nongaussianity leading to a high separation and dereverberation performance without whitening artifacts.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley & Sons, Inc., New York, 2001.

[2] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON-based blind system identification with application to multiple-source localization and separation," in S. Makino, T.-W. Lee, and S. Sawada (eds.), *Blind Speech Separation*, Springer, Berlin, pp. 101-147, Sept. 2007.

[3] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in P.A. Naylor and N.D. Gaubitch (eds.), *Speech Dereverberation*, Springer, London, to appear in 2010.

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol 36, no. 2, pp. 145-152, Feb. 1988.

[5] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Montreal, Canada, vol. 3, pp. 889-892, May 2004.

[6] S. Amari et al.,"Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. IEEE Int. Workshop Signal Processing Advances in Wireless Communications*, pp. 101-107, 1997.

[7] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in Y. Huang and J. Benesty (eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer Academic Publishers, Boston, pp. 255-293, Feb. 2004.

[8] C.L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of nonstationary signals," in *Proc. Int. Workshop Neural Networks Signal Processing (NNSP)*, 2001, pp. 303-312.

[9] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley & Sons, New York, 1991.

[10] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, no. 6, pp.1260–1277, 2006.

[11] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[12] J.D. Markel and A.H.Gray, *Linear Prediction of Speech*, Springer, Berlin, 3rd edition, 1976.

[13] R.A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol 16, pp. 21-35, 1978.

[14] R.H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*, Ph.D. dissertation, Univ. of Southern California, Los Angeles, CA, May 1996.

[15] B.W. Gillespie, H.S. Malvar, and D.A.F. Florêncio,"Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE ICASSP*, Salt Lake City, UT, USA, May 2001.