TRINICON: A VERSATILE FRAMEWORK FOR MULTICHANNEL BLIND SIGNAL PROCESSING

Herbert Buchner, Robert Aichner, and Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{buchner,aichner,wk}@LNT.de

ABSTRACT

In this paper we present a framework for multichannel blind signal processing for convolutive mixtures, such as blind source separation (BSS) and multichannel blind deconvolution (MCBD). It is based on the use of multivariate pdfs and a compact matrix notation which considerably simplifies the representation and handling of the algorithms. By introducing these techniques into an information theoretic cost function, we can exploit the three fundamental signal properties nonwhiteness, nongaussianity, and nonstationarity. This results in a versatile tool that we call TRINICON ('Triple-N ICA for convolutive mixtures'). Both, links to popular algorithms and several novel algorithms follow from the general approach. In particular, we introduce a new concept of multichannel blind *partial* deconvolution (MCBPD) for speech which prevents a complete whitening of the output signals, i.e., the vocal tract is excluded from the equalization. This is especially interesting for automatic speech recognition applications. Moreover, we show results for BSS using multivariate spherically invariant random processes (SIRPs) to efficiently model speech, and show how the approach carries over to MCBPD. These concepts are also suitable for an efficient implementation in the frequency domain by using a rigorous broadband derivation avoiding the internal permutation problem and circularity effects.

1. INTRODUCTION

The task to perform blind signal processing on convolutive mixtures of unknown time series arises in several application domains, a prominent example being the so-called cocktail party problem, where we want to recover the speech signals of multiple speakers who are simultaneously talking in a room. The room may be very reverberant due to reflections on the walls, i.e., the original source signals $s_q(n)$, $q = 1, \ldots, Q$ are filtered by a linear multiple input and multiple output (MIMO) system before they are picked up by the sensors. In the following, we assume that the number Q of source signals $s_q(n)$ equals the number of sensor signals $x_p(n)$, $p = 1, \ldots, P$ (*Fig.* 1).

We distinguish two classes of signal processing problems in this scenario:

BSS for convolutive mixtures. In this approach, we want to determine a MIMO FIR demixing filter which separates the signals up to an arbitrary filtering and permutation by forcing the output signals to be mutually independent.

Multichannel Blind Deconvolution. Here, in addition to the separation, we want to recover the original signals up to an arbitrary (frequency-independent) scaling and permutation, i.e., we want to dereverberate the signals.

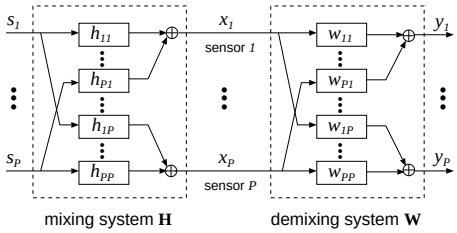


Fig. 1. Setup for blind MIMO signal processing.

For the blind estimation of the coefficients, it has to be considered that the speech excitation is colored and non-stationary. In addition, in MCBD, effectively, an inversion of (long and possibly non-minimum phase) room impulse responses is necessary. However, using the multiple-input/output inverse theorem (MINT) [1] any MIMO system $\mathbf{H}$ can exactly be inverted if $h_{qp}\ \forall\ p, q \in \{1, \ldots, P\}$ do not have common zeros in the z plane. Therefore, in principle, there is a general solution to the MCBD problem by using multiple microphones (According to our experience, 3 or 4 microphones are usually sufficient in real environments).

So far, the BSS problem has been mostly addressed for instantaneous mixtures or narrowband approaches in the frequency domain which adapt the coefficients independently in each DFT domain, e.g., [2]-[5]. In the case of MCBD, many approaches are either based on a whitening of the output signals (e.g., [6, 7]), which is problematic for speech and audio applications like automatic speech recognition, or are rather heuristically motivated, e.g., [8].

The aim of this paper is to present a unified treatment of BSS and MCBD algorithms applicable for speech signals in real acoustic environments. We propose the use of multivariate models in the cost function to capture the statistical description of the temporal structure of the source signals. Generalizing the approach in [9] allows the use of the cost function for both, BSS and MCBD, and also leads to an improved solution for speech dereverberation. Note that we consider here only time-domain algorithms. For the general frequency-domain broadband approach, see [10, 11].

2. MATRIX NOTATION FOR CONVOLUTIVE MIXTURES

To derive an algorithm for block processing of convolutive mixtures, we first need to formulate the convolution of the FIR demix-

ing system of length L in the following matrix form [10]:

$$\mathbf{y}(m,j) = \mathbf{x}(m,j)\mathbf{W}(m), \tag{1}$$

where m denotes the block index, and $j = 0, \cdots, N-1$ is a time-shift index within a block of length N, and

$$\begin{aligned}
\mathbf{x}(m,j) &= [\mathbf{x}_1(m,j), \ldots, \mathbf{x}_P(m,j)], && (2)\\
\mathbf{y}(m,j) &= [\mathbf{y}_1(m,j), \ldots, \mathbf{y}_P(m,j)], && (3)\\
\mathbf{W}(m) &= \begin{bmatrix} \mathbf{W}_{11}(m) & \cdots & \mathbf{W}_{1P}(m) \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{P1}(m) & \cdots & \mathbf{W}_{PP}(m) \end{bmatrix}, && (4)\\
\mathbf{x}_p(m,j) &= [x_p(mL+j), \ldots, x_p(mL-2L+1+j)], && (5)\\
\mathbf{y}_q(m,j) &= [y_q(mL+j), \ldots, y_q(mL-D+1+j)] && (6)\\
&= \sum_{p=1}^{P} \mathbf{x}_p(m,j)\mathbf{W}_{pq}(m). && (7)
\end{aligned}$$

D in (6) denotes the number of lags taken into account to exploit the nonwhiteness of the source signals as shown below. $\mathbf{W}_{pq}(m)$ denotes a $2L \times D$ Sylvester matrix that contains all coefficients of the respective filter:

$$\mathbf{W}_{pq}(m) = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & w_{pq,L-1} \\ 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{8}$$

Note that this notation also allows a simple way to calculate the demixing filters exactly according to MINT for a known mixing system described by Sylvester matrices $\mathbf{H}_{qp}$ of the same structure as shown above and suitably chosen matrix dimensions [1]:

$$\mathbf{C} = \mathbf{H}\mathbf{W} \;\Rightarrow\; \mathbf{W} = \mathbf{H}^{-1}\mathbf{C}, \tag{9}$$

where $\mathbf{C}$ denotes the corresponding system matrix from the sources to the outputs, i.e., for exact dereverberation, we have $\mathbf{C} = \mathbf{I}$.

3. COST FUNCTION AND GENERAL OPTIMIZATION

3.1. Optimization Criterion

Different approaches exist to blindly estimate the demixing matrix $\mathbf{W}$ for the above mentioned tasks by utilizing the following source signal properties [2] which we all combine into an efficient and versatile algorithm:

(i) Nongaussianity is exploited by using higher-order statistics for independent component analysis (ICA). ICA approaches can be divided into several classes. Although they all lead to similar update rules, the minimization of the mutual information (MMI) among the output channels can be regarded as the most general approach for BSS [2]. To obtain an estimator not only allowing spatial separation but also temporal separation for MCBD, we use the Kullback-Leibler distance (KLD) [12] between a certain *desired* joint pdf (essentially representing a hypothesized stochastic source model) and the joint pdf of the actually estimated output signals. The desired pdf is factorized w.r.t. the different sources (for BSS) and possibly also w.r.t. certain temporal dependencies (for MCBD) as shown below. The KLD is guaranteed to be positive [12], which is a necessary condition for a useful cost function.

(ii) Nonwhiteness is exploited by simultaneous minimization of output cross-relations over multiple time-lags. We therefore consider multivariate pdfs, i.e., 'densities including 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.

Based on the KLD, we now define the following general cost function taking into account all three fundamental signal properties (i)-(iii):

$$\begin{aligned}
\mathcal{J}(m) &= -\sum_{i=0}^{\infty} \beta(i,m) \frac{1}{N} \sum_{j=0}^{N-1} \{\log(\hat{p}_{s,PD}(\mathbf{y}(i,j))) \\
&\quad - \log(\hat{p}_{y,PD}(\mathbf{y}(i,j)))\}, \qquad (10)
\end{aligned}$$

where $\hat{p}_{s,PD}(\cdot)$ and $\hat{p}_{y,PD}(\cdot)$ are the assumed or estimated PD-variate source model (i.e., desired) pdf and output pdf, respectively. Furthermore, D is the memory length, i.e., the number of time-lags to model the nonwhiteness of the P signals as above. β is a window function with finite support that is normalized according to $\sum_{i=0}^{m} \beta(i,m) = 1$ allowing for online, offline, and block-online algorithms [11].

3.2. General Coefficient Update

It can be shown (after a somewhat tedious but straightforward derivation) that by taking the *natural gradient* [2] of $\mathcal{J}(m)$ with respect to the demixing filter matrix $\mathbf{W}(m)$ [10],

$$\Delta\mathbf{W} \propto \mathbf{W}\mathbf{W}^H \frac{\partial \mathcal{J}}{\partial \mathbf{W}^*}, \tag{11}$$

we obtain the following generic TRINICON-based update rule:

$$\begin{aligned}
\mathbf{W}(m) &= \mathbf{W}(m-1) - \mu\Delta\mathbf{W}(m), && (12)\\
\Delta\mathbf{W}(m) &= \frac{2}{N} \sum_{i=0}^{\infty} \beta(i,m) \sum_{j=0}^{N-1} \mathbf{W}(i)\mathbf{y}^H(i,j) \\
&\quad \cdot \{\mathbf{\Phi}_{s,PD}(\mathbf{y}(i,j)) - \mathbf{\Phi}_{y,PD}(\mathbf{y}(i,j))\}, && (13)
\end{aligned}$$

with the *desired* score function

$$\mathbf{\Phi}_{s,PD}(\mathbf{y}(i,j)) = -\frac{\partial \log \hat{p}_{s,PD}(\mathbf{y}(i,j))}{\partial \mathbf{y}(i,j)} \tag{14}$$

resulting from the hypothesized source model (a factorization of $\hat{p}_{s,PD}(\cdot)$ among the sources yields BSS, while a complete factorization leads to the traditional MCBD approach), and the actual score function

$$\mathbf{\Phi}_{y,PD}(\mathbf{y}(i,j)) = -\frac{\partial \log \hat{p}_{y,PD}(\mathbf{y}(i,j))}{\partial \mathbf{y}(i,j)}. \tag{15}$$

4. SPECIAL CASES AND ILLUSTRATION

There are many interesting known and novel practical approximations within the framework. To begin with, we first consider algorithms based on second-order statistics (SOS) as they are particularly illustrative.

4.1. Realizations based on Second-Order Statistics

Here, the source models are simplified to seqences of multivariate Gaussian functions described by $PD \times PD$ correlation matrices $\mathbf{R}_{\cdot\cdot}$ within the length-N signal blocks. This leads to the coefficient update

$$\begin{aligned}\Delta\mathbf{W}(m) &= 2\sum_{i=0}^{\infty}\beta(i,m)\mathbf{W}(i)\mathbf{R}_{\mathbf{yy}}\left\{\hat{\mathbf{R}}_{\mathbf{ss}}^{-1}-\hat{\mathbf{R}}_{\mathbf{yy}}^{-1}\right\}\\ &= 2\sum_{i=0}^{\infty}\beta(i,m)\mathbf{W}(i)\left\{\hat{\mathbf{R}}_{\mathbf{yy}}-\hat{\mathbf{R}}_{\mathbf{ss}}\right\}\hat{\mathbf{R}}_{\mathbf{ss}}^{-1}\end{aligned} \quad (16)$$

4.1.1. Generic SOS-based BSS

The BSS variant of the generic SOS natural gradient update (16) follows immediately by setting

$$\hat{\mathbf{R}}_{\mathbf{ss}}(i) = \mathrm{bdiag}_D\,\hat{\mathbf{R}}_{\mathbf{yy}}(i). \quad (17)$$

The update (16) together with (17) was originally obtained independently in [10] as a generalization of the cost function of [13]:

$$\mathcal{J}_{\mathrm{SOS}}(m) = \sum_{i=0}^{\infty}\beta(i,m)\left\{\log\det\hat{\mathbf{R}}_{\mathbf{ss}}(i) - \log\det\hat{\mathbf{R}}_{\mathbf{yy}}(i)\right\}. \quad (18)$$

In *Fig.* 2 the mechanism of (16) based on the model (17) is illustrated. By minimizing $\mathcal{J}_{\mathrm{SOS}}(m)$, all cross-correlations for D time-lags are reduced and will ideally vanish, while the auto-correlations are untouched to preserve the structure of the individual signals. This algorithms leads to very robust practical solutions

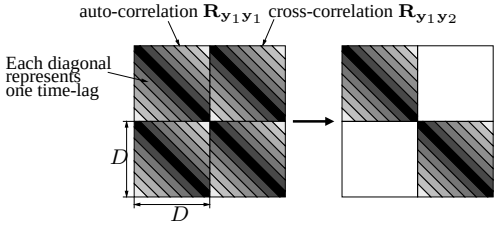


Fig. 2. Illustration of SOS-based BSS.

even for a large number of filter taps due to an inherent normalization by the auto-correlation matrices, reflected by the inverse in (16) of $\mathrm{bdiag}_D\,\hat{\mathbf{R}}_{\mathbf{yy}}$. Note that there are also efficient approximations of this algorithm [14],[15] with a reduced computational complexity allowing already realtime operation on a regular PC platform. Moreover, a close link has been established [10],[11] to some popular frequency-domain algorithms [3],[4].

4.1.2. MCBD based on SOS

Traditionally, ICA-based MCBD algorithms assume i.i.d. source models, e.g., [6, 7]. In the SOS case, this corresponds to a complete whitening of the output signals by not only applying a joint de-cross-correlation, but also a joint de-auto-correlation, i.e., $\hat{\mathbf{R}}_{\mathbf{ss}} = \mathrm{diag}\,\hat{\mathbf{R}}_{\mathbf{yy}}$ over multiple time-instants, as illustrated in *Fig.* 4 (b).

4.1.3. MCBPD based on SOS

Especially for distant-talking automatic speech recognition (ASR), there is a very strong need for speech dereverberation without introducing artifacts to the signals. In ASR, certain features for the actual recognition process are extracted from short signal blocks. These blocks are generally much shorter than the reverberation time in real environments. The resulting temporal 'smearing' often significantly degrades the performance of large-vocabulary ASR. On the other hand, it is known that ASR is relatively insensitive to very short reverberation components due to the block-based feature extraction. Equations (13)-(15) inherently contain a statistical source model (signal properties (i)-(iii) in Sect. 3.1), expressed by the multivariate densities, and thus provide all necessary requirements for the MCBPD approach which allows to distinguish between the actual speech production system, i.e., the vocal tract, and the reverberant room (*Fig.* 3). Ideally, only the influence of the room acoustics should be minimized. In the SOS case, the auto-

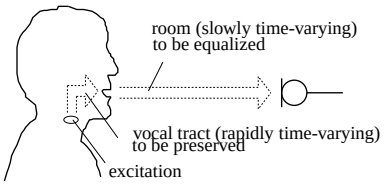
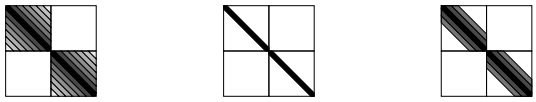


Fig. 3. Illustration of the MCBPD principle.

correlation structure of the speech signals can be taken into account, as shown in *Fig.* 4 (c). While the room acoustics influences all off-diagonals, the effect of the vocal tract is concentrated in the first few lags around the main diagonal. These first off-diagonals of $\hat{\mathbf{R}}_{\mathbf{yy}}$ are now taken over into $\hat{\mathbf{R}}_{\mathbf{ss}}$. Alternatively, the structure in *Fig.* 4 (c) may be approximated by small sub-matrices making its handling somewhat more efficient. Note that there is a close link to linear prediction techniques which gives guidelines for the number of lags to be preserved. Experiments with a state-of-the-art large-vocabulary speech recognizer (by Dragon Systems) confirm the effectiveness of this approach by a significant improvement of the word error rate for the consideration of 20-30 off-diagonals in $\hat{\mathbf{R}}_{\mathbf{ss}}$ at a sampling rate of 16kHz.



(a) BSS (b) MCBD (c) MCBPD

Fig. 4. Desired correlation matrices for BSS, MCBD, and MCBPD.

4.2. Realizations based on Higher-Order Statistics

The general HOS approach (13)-(15) provides the possibility to take into account all information that we possibly have on the statistical properties of the desired source signals. This provides an increased flexibility and improved performance of BSS. Moreover, the more accurate modeling of the desired source signals gives an improved MCBPD.

To apply the general approach in a real-world scenario, an appropriate multivariate score functions (14) and (15) have to be determined. Fortunately, there is an efficient solution to this prob-

lem by assuming so-called spherically invariant random processes (SIRPs). The general form of correlated SIRPs of D-th order is given with a properly chosen function $f_D(\cdot)$ by [16]

$$\hat{p}_D(\mathbf{y}_p(i,j)) = \frac{1}{\sqrt{\pi^D \det(\mathbf{R}_{pp}(i))}} f_D\left(\mathbf{y}_p(i,j)\mathbf{R}_{pp}^{-1}(i)\mathbf{y}_p^H(i,j)\right) \tag{19}$$

for the p-th channel. These models are representative for a wide class of stochastic processes. Speech signals in particular can very accurately be represented by SIRPs [16]. A great advantage arising from the SIRP model is that multivariate pdfs can be derived analytically from the corresponding univariate pdf together with the (lagged) correlation matrices. The function $f_D(\cdot)$ can thus be calculated from the well-known univariate models for speech, e.g., the Laplacian density. Using the chain rule, the corresponding score function (14) can be derived from (19), as shown in [9, 11] in more detail.

The calculation of (15) becomes particularly simple in most practical realizations by transforming the output pdf $\hat{p}_{y,PD}(\cdot)$ into the corresponding multivariate input signal pdf using $\mathbf{W}$ which is considered as a mapping matrix of a linear transformation. The derivative of the input signal pdf vanishes as it is independent of the demixing system. Note that in the extreme case of a full deconvolution and i.i.d assumption, this approach boils down to the traditional MCBD approach in [7] which is an improved version of [6].

5. SIMULATION RESULTS

We conducted our experiments on BSS for convolutive mixtures using speech signals from the TIMIT database convolved with impulse responses of a real room with reverberation time $T_{60} \approx 150$ ms. Note that this class of algorithms can cope with a long filter length (or reverberation time) due to the inherent normalization property discussed in Sect. 4.1.1. The sampling rate was $f_s = 16$ kHz. We used a two-element microphone array with an inter-element spacing of 16 cm. For the filter adaptation (offline) we used both, the generic SOS algorithm in the time-domain, and the generic HOS algorithm with SIRP model from the Laplacian pdf. We chose the following parameters: $L = 512$, $N = 1024$, $D = 512$ (note that N has to be chosen greater than D to get improved estimates in the HOS case). To evaluate the performance, as shown in *Fig.* 5 we used the signal-to-interference ratio (SIR), defined as the ratio of the signal power of the target signal to the signal power from the jammer signal. For *Fig.* 5, the stepsizes have been maximized up to the stability margin.

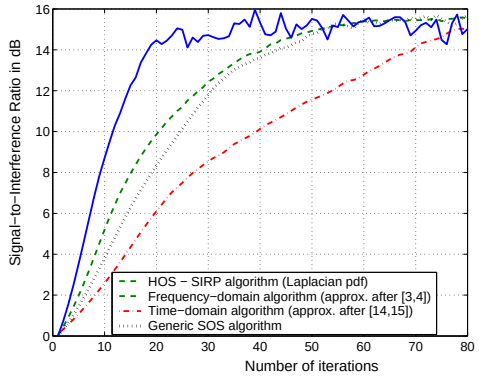


Fig. 5. BSS simulation results for two sensors and $L = 512$.

6. CONCLUSIONS

We presented a versatile framework for blind signal processing exploiting all three fundamental statistical source properties in a rigorous way. There are various interesting links to both, known and novel time-domain and frequency-domain algorithms. An efficient realization for blind source separation and dereverberation incorporating a SIRP-based model has been described.

7. REFERENCES

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

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

[3] H.-C. Wu and J. C. Principe,"Simultaneous diagonalization in the frequency domain (SDIF) for source separation," in *Proc. ICA*, pp. 245-250, 1999.

[4] C.L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of non-stationary signals," in Proc. *NNSP*, 2001.

[5] S. Amari, T.-P. Chen, and A. Cichocki, "Nonholonomic orthogonal learning algorithms for blind source separation," *Neural Computation*, vol. 12, no. 6, pp. 1463-1484, 2000.

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

[7] S. Choi et al., "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *Proc. Int. Symp ICA*, pp. 371-376, 1999.

[8] B.W. Gillespie and L. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," *Proc. IEEE ICASSP*, 2003.

[9] H. Buchner, R. Aichner, and W. Kellermann, "Blind Source Separation for Convolutive Mixtures Exploiting Nongaussianity, Nonwhiteness, and Nonstationarity," *Proc. Int. Workshop on Acoustic Echo and Noise Control*, 2003.

[10] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," *Proc. Int. Symp ICA*, 2003.

[11] 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, to appear 2004.

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

[13] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved non-stationary signals," *Neurocomputing*, vol. 22, pp. 157-171, 1998.

[14] T. Nishikawa, H. Saruwatari, and K. Shikano, "Comparison of time-domain ICA, frequency-domain ICA and multistage ICA for blind source separation," in *Proc. European Signal Processing Conference*, vol. 2, pp. 15-18, Sep. 2002.

[15] R. Aichner et al., "Time-domain blind source separation of non-stationary convolved signals with utilization of geometric beamforming," in Proc. *NNSP*, 2002.

[16] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing* vol. 12, pp. 119-141, 1987.