UNDERDETERMINED SOURCE SEPARATION BY ICA AND HOMOMORPHIC SIGNAL PROCESSING

[1†]*Stefan Winter,* [2]*Walter Kellermann,* [1]*Hiroshi Sawada and* [1]*Shoji Makino*

[1]`wifan@cslab.kecl.ntt.co.jp`
[1]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0237 Japan
[2]Department of Multimedia Communication and Signal Processing, University Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen

ABSTRACT

Nearly all approaches for underdetermined blind source separation (BSS) assume independent and identically distributed (i.i.d.) sources. They completely ignore the temporal structure of colored sources such as speech signals. Instead, we propose a multivariate model based on a multivariate Gaussian distribution that is then used to determine an unmixing matrix for underdetermined BSS. Based on parameterization by cepstral coefficients we present a novel ICA-based cost function for estimating the speech-related parameters of the unmixing matrix. Experimental results support the proposed approach.

1. INTRODUCTION

Blind source separation (BSS) describes techniques that aim at separating P signals if no information is available other than Q mixed versions of the original signals. Most BSS approaches assume that there are at least as many microphones as source signals ($Q \geq P$), which is called (over-) determined BSS. Since overdetermined BSS can be reduced to determined BSS [1], we refer to both as determined BSS.

Instead, we consider underdetermined BSS, where we have fewer microphones than source signals ($Q < P$). In this case, (see e.g. [2, 3, 4], the separation quality in terms of interference suppression and signal distortion is still not as good as with determined BSS. This is particularly true if wideband signals such as speech signals are involved. The difficulty is that in contrast to determined BSS the solution of underdetermined BSS goes beyond system identification. Even if the mixing system is fully identified, additional effort is required to separate the mixtures.

Nearly all approaches designed to solve the latter problem assume independent and identically distributed (i.i.d.) sources [2, 3, 4].While this assumption may serve well as a first order approximation, it completely ignores the temporal structure of colored sources such as speech signals. Here we propose a multivariate model that takes the temporal correlation explicitly into account. We consider convolutive mixtures of speech signals in the time domain. This paper concentrates on the separation and assumes that the mixing matrix is known.

The contribution of this paper is based on the following ideas: With the multivariate source model we can derive an unmixing matrix for underdetermined BSS, which results in high quality source separation if the parameters of this unmixing matrix are known. We rely on two principles when estimating the source-related parameters of the unmixing matrix. First we reduce the number of parameters by utilizing real cepstral coefficients based on principles of homomorphic signal processing [5]. Then we apply independent component analysis (ICA) to estimate the reduced set of parameters.

After formulating the problem analytically in Sec. 2, we provide statistical models for Bayesian inference in Sec. 3 together with the resulting unmixing matrix. Section 4 elaborates the concise parameterization of the unmixing matrix. In Sec. 5 we derive a novel algorithm for estimating the speech parameters based on ICA. Section 6 presents experimental results which are discussed in Section 7.

2. PROBLEM FORMULATION

With $s_p(t) \in \mathbb{R}$ denoting the p-th source signal ($1 \leq p \leq P$) and $h_{qp}(\tau)$ ($0 \leq \tau \leq L-1$) the impulse response of length L from source p to sensor q, we obtain mixed signals $x_q(t) \in \mathbb{R}$ ($1 \leq q \leq Q, Q < P$) by

$$x_q(t) = \sum_p \sum_\tau s_p(t) h(\tau - t) + n(t). \qquad (1)$$

$n(t)$ denotes noise added to the sensors. Let

$$\boldsymbol{S}_p = \begin{bmatrix} s_p(0) & \cdots & s_p(T-1) \end{bmatrix}^{\mathcal{T}}, \quad 1 \leq p \leq P \qquad (2)$$

denote a frame of length T of the p-th original speech signal. $(\cdot)^{\mathcal{T}}$ denotes the transpose. We assume that the

[†] The author is on leave from the Department of Multimedia Communication and Signal Processing, University Erlangen-Nuremberg

speech signal is stationary within a frame of length T with autocorrelation

$$r_p(\tau) = E\{s_p(t)s_p(t+\tau)\}. \qquad (3)$$

We summarize the P frames of the different source signals by the vector

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_P \end{bmatrix} \in \mathbb{R}^{P \cdot T}. \qquad (4)$$

For each source $\boldsymbol{S}_p$ the autocorrelation matrix $\boldsymbol{R}_p \in \mathbb{R}^{T \times T}$ is given by a symmetric Toeplitz matrix. Its first row is defined by

$$\boldsymbol{r}_p = \begin{bmatrix} r_p(0) & \cdots & r_p(T-1) \end{bmatrix}. \qquad (5)$$

We summarize the P autocorrelation matrices by

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}_P \end{bmatrix} \in \mathbb{R}^{P \cdot T \times P \cdot T}. \qquad (6)$$

We define the mixing matrix

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \cdots & \boldsymbol{H}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_{Q1} & \cdots & \boldsymbol{H}_{QP} \end{bmatrix} \in \mathbb{R}^{Q \cdot (T-L+1) \times P \cdot T}, \qquad (7)$$

where $\boldsymbol{H}_{qp}$ denotes convolution matrices with Toeplitz structure

$$\boldsymbol{H}_{qp} = \begin{bmatrix} h_{qp}(L-1) & \ldots & h_{qp}(0) & & \boldsymbol{0} \\ & \ddots & & & \\ \boldsymbol{0} & & h_{qp}(L-1) & \ldots & h_{qp}(0) \end{bmatrix}$$

$$\in \mathbb{R}^{T-L+1 \times T}. \qquad (8)$$

Similar to the source signals we define

$$\boldsymbol{X}_q = \begin{bmatrix} x_q(0) & \cdots & x_q(T-L) \end{bmatrix}^{\mathcal{T}}, \quad 1 \le q \le Q \qquad (9)$$

and summarize $\boldsymbol{X}_q$ by

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_Q \end{bmatrix} \in \mathbb{R}^{Q \cdot (T-L+1)}. \qquad (10)$$

This results in a compact description of the mixing process for one frame given by

$$\boldsymbol{X} = \boldsymbol{H}\boldsymbol{S} + \boldsymbol{N} \qquad (11)$$

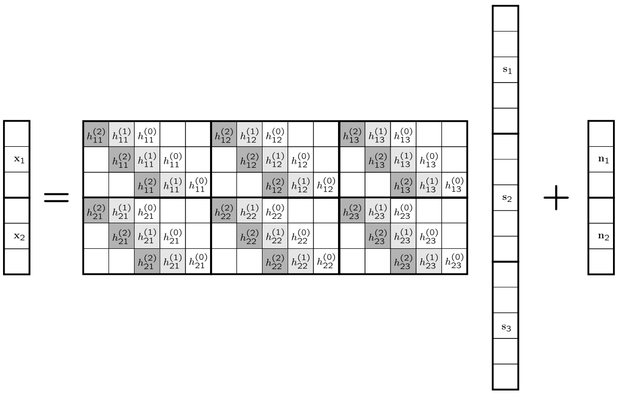
x1
x2
=
$h_{11}^{(2)}$ $h_{11}^{(1)}$ $h_{11}^{(0)}$ $h_{12}^{(2)}$ $h_{12}^{(1)}$ $h_{12}^{(0)}$ $h_{13}^{(2)}$ $h_{13}^{(1)}$ $h_{13}^{(0)}$
$h_{21}^{(2)}$ $h_{21}^{(1)}$ $h_{21}^{(0)}$ $h_{22}^{(2)}$ $h_{22}^{(1)}$ $h_{22}^{(0)}$ $h_{23}^{(2)}$ $h_{23}^{(1)}$ $h_{23}^{(0)}$
s1
s2
s3
+
n1
n2

Figure 1: Mixing process for underdetermined BSS with $Q = 2$, $P = 3$, $T = 5$ and $L = 3$

as illustrated in Fig. 1. $\boldsymbol{N} \in \mathbb{R}^{Q \cdot (T-L+1)}$ is derived from $n(t)$ in a similar way to that used to derive $\boldsymbol{X}$ from $x(t)$. The final goal in BSS is the estimation of signals $\boldsymbol{Y}$ that resemble the original signals $\boldsymbol{S}$ as closely as possible. Since only the mixed signals are available, this is at best possible up to arbitrary permutation and scaling. With determined BSS it is sufficient to estimate the mixing matrix $\boldsymbol{H}$ or its inverse. Since the mixing matrix is square and therefore invertible (assuming that $\boldsymbol{H}$ is non-singular), the inverse can be used to separate the signals. In contrast to determined BSS, in the underdetermined case estimation of the mixing matrix $\boldsymbol{H}$ is insufficient. Even if $\boldsymbol{H}$ is available, estimating the original signals from the mixtures poses its own problem, since $\boldsymbol{H}$ cannot be simply inverted. In the following we concentrate on estimating the original signals and assume that the mixing matrix is known or else can be estimated [2, 6, 7].

3. MODELS

We make the common assumption that the speech signals are mutually independent

$$p(\boldsymbol{S}) = \prod_p p(\boldsymbol{S}_p). \qquad (12)$$

We further assume that the speech signal is approximately stationary within a frame of appropriately chosen length T.

Nearly all approaches to underdetermined BSS assume independent and identical prior distributions. In contrast, in this section we propose a multivariate prior based on the analytically tractable, multivariate Gaussian distribution

$$p(\boldsymbol{S}_p) = \mathcal{N}(\boldsymbol{S}_p | \boldsymbol{\mu}_p, \boldsymbol{R}_p) \qquad (13)$$

with zero mean vector $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{R}_p$.

We also define the likelihood derived from the commonly assumed additive white Gaussian sensor noise model with identical variance σ^2 across all sensors. Based on the mixing process (11), this assumption results in the Gaussian likelihood

$$p(\boldsymbol{X}|\boldsymbol{H},\boldsymbol{S},\sigma^2\boldsymbol{I}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{H}\boldsymbol{S},\sigma^2\boldsymbol{I}). \tag{14}$$

From the Bayes' rule [8] we then obtain the posterior of the source signal given the mixed signals as

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{X}) &\propto p(\boldsymbol{S})\cdot p(\boldsymbol{X}|\boldsymbol{H},\boldsymbol{S},\sigma^2\boldsymbol{I}) \\ &\propto \mathcal{N}(\boldsymbol{S}|\boldsymbol{\mu}_{\boldsymbol{S}},\boldsymbol{\Sigma}_{\boldsymbol{S}}) \end{aligned} \tag{15}$$

with the correlation matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{S}} = \left(\boldsymbol{R}^{-1} + \frac{1}{\sigma^2}\boldsymbol{H}^{\mathcal{T}}\boldsymbol{H}\right)^{-1} \tag{16}$$

and the mean

$$\boldsymbol{\mu}_{\boldsymbol{S}} = \left(\sigma^2\boldsymbol{R}^{-1} + \boldsymbol{H}^{\mathcal{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathcal{T}}\boldsymbol{X}. \tag{17}$$

Since the posterior distribution is again Gaussian, the MMSE estimate is given by its mean, resulting in the estimated source signals

$$\boldsymbol{Y} = \underbrace{\left(\sigma^2\boldsymbol{R}^{-1} + \boldsymbol{H}^{\mathcal{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathcal{T}}}_{\boldsymbol{W}}\boldsymbol{X}. \tag{18}$$

As shown in Sec. 6, the unmixing matrix $\boldsymbol{W}$ is sufficient for high quality source separation. The problem is now to estimate the parameters of the unmixing matrix.

4. PARAMETERIZATION

The fewer the parameters that need to be estimated in relation to the available data, the better the estimate will be. In order to reduce the number of parameters in the unmixing matrix, we describe the autocorrelation by cepstral coefficients. We exploit the fact that speech signals can generally be concisely described by a few cepstral coefficients [5]. In addition, the autocorrelation is related to the real cepstrum by a one-to-one mapping based on the Wiener-Khintchine theorem [9].

We summarize the D_p real cepstral coefficients used for the p-th source by the vector $\widehat{\boldsymbol{S}}_p$. By zero-padding $\widehat{\boldsymbol{S}}_p$ to length T, the autocorrelation sequence $\boldsymbol{r}_p$ can be shown to be given by

$$\boldsymbol{r}_p = \mathsf{IDFT}\left\{\exp\left(2\cdot\mathsf{DFT}\left\{\widehat{\boldsymbol{S}}_{\mathsf{p}}\right\}\right)\right\}, \tag{19}$$

where $\mathsf{IDFT}\left\{\cdot\right\}$ and $\mathsf{DFT}\left\{\cdot\right\}$ denote the (inverse) Discrete Fourier Transform and $\exp(\cdot)$ is applied elementwise.

To estimate optimal real cepstral coefficients based on the cost function presented in Sec. 5, we need the derivative of the autocorrelation sequence $\boldsymbol{r}_p$ with respect to the cepstral coefficient $\widehat{S}_p^d$, $1 \leq d \leq D$, which denotes the d-th element of $\widehat{\boldsymbol{S}}_p$. The derivative yields

$$\frac{\partial \boldsymbol{r}_p}{\partial \widehat{S}_p^d} = 2\cdot\mathsf{IDFT}\left\{\exp\left(2\cdot\mathsf{DFT}\left\{\widehat{\boldsymbol{S}}_{\mathsf{p}}\right\} - \jmath\frac{2\pi\mathsf{d}}{\mathsf{T}}\mathsf{k}\right)\right\}. \tag{20}$$

Then, the derivative of the correlation matrix $\boldsymbol{R}_p$ is given by a symmetric Toeplitz matrix with $\frac{\partial \boldsymbol{r}_p}{\partial \widehat{S}_p^d}$ as its first column.

5. PARAMETER ESTIMATION

To estimate the cepstral coefficients of each source in a given frame, we apply the principles of ICA. Based on minimum mutual information (MMI) [10], which is the most general approach to ICA [11], we define the cost function

$$\mathcal{J}_{\text{ICA}} = -E\left\{\log|\det\boldsymbol{R}| + \log|\det\boldsymbol{C}| + \right. \tag{21}$$

$$\left. \sum_p \log\left(p_{\boldsymbol{Y}_p}(\boldsymbol{Y}_p)\right)\right\} \tag{22}$$

with $E\{\cdot\}$ denoting the expectation and $\boldsymbol{C}$ being defined as

$$\boldsymbol{C} := \left(\sigma^2\boldsymbol{I} + \boldsymbol{H}\boldsymbol{R}\boldsymbol{H}^{\mathcal{T}}\right)^{-1}. \tag{23}$$

For a gradient-based optimization of the cost function we determined the derivative of the cost function by using the multivariate source model (13). The result is given in (24), where tr $(\cdot)$ denotes the matrix trace.

6. EXPERIMENTAL RESULTS

We performed experiments with three speech signals (two male, one female) of 1.62 seconds (8 kHz sampling rate). We generated two convolutive mixtures with artificial impulse responses of length $L = 11$ and assumed that the mixing matrix was available. The signals were processed in frames of $T = 128$ samples and shifted by 64 samples. No noise was added to the mixtures. Therefore, the variance σ^2 served as a regularization parameter and was set at $\sigma^2 = 10^{-4}$. The sources had the same variance of 0.0037. We compared the following approaches:

E1 This constitutes the reference of the unprocessed mixed signals

E2 The autocorrelation was directly estimated from the original signals to provide an upper limit for the separation performance of $\boldsymbol{W}$ (18).

$$\frac{\partial \mathcal{J}_{\text{ICA}}}{\partial \widehat{S}_p^d} = E\left\{ \text{tr}\left(\left(\boldsymbol{H}^{\mathcal{T}} \boldsymbol{C} \boldsymbol{H} - \frac{1}{2} \boldsymbol{R}^{-1} - \left(\boldsymbol{I} - 2\left(\boldsymbol{W}\boldsymbol{H}\right)^{\mathcal{T}} \right) \boldsymbol{I} \boldsymbol{H}^{\mathcal{T}} \boldsymbol{C} \boldsymbol{X} \boldsymbol{X}^{\mathcal{T}} \boldsymbol{C}^{\mathcal{T}} \boldsymbol{H} \right) \frac{\partial \boldsymbol{R}}{\partial \widehat{S}_p^d} \right) \right\} \tag{24}$$

Method	SDR	SIR	SAR
E1	−20.80	−6.01	−10.47
E2	18.39	33.13	18.60
E3	6.64	8.88	11.60

Table 1: Experimental results

E3 The autocorrelation was parameterized by $D = 30$ cepstral coefficients. These, in turn, were estimated by the ICA-based approach as described in Sec. 5. To obtain values for the real cepstral coefficients that led to invertible correlation matrices we provided lower and upper bounds. They were trained based on clean speech signals.

To evaluate of the separation results we used the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) as defined in [12]. The averaged results for the different approaches are shown in Table 1.

7. CONCLUSION

The results in Table 1 together with a subjective evaluation suggest that knowing the autocorrelation together with the mixing matrix is sufficient to perform high quality underdetermined source separation in noiseless environments (E1 compared to E2). In other words, the problem of underdetermined source separation can be reduced to the estimation of the underlying autocorrelations once the mixing matrix is available. We proposed a novel ICA-based cost function for estimating the autocorrelations based on a parameterization by cepstral coefficients. An algorithm based on the gradient of the cost function provides a good estimate of the source signals in underdetermined BSS (E3).

In the future we plan to estimate both the speech parameters of the unmixing matrix, and the mixing parameters based on ICA. The estimation of both kinds of parameters would result in a fully blind algorithm for underdetermined source separation.

8. REFERENCES

[1] S. Winter, H. Sawada, and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006, Article ID 71632, 11 pages.

[2] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1693–1700, 2005.

[3] C. Févotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," Tech. Rep., Cambridge University Engineering Dept., January 2005.

[4] S. Winter, H. Sawada, and S. Makino, "On real and complex valued L1-norm minimization for overcomplete blind source separation," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2005, pp. 86–89.

[5] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice Hall, 1978.

[6] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Proc. ICA 2004*, Sep 2004, pp. 652–660.

[7] L. Vielva, I. Santamaria, C. Pantaleon, J. Ibanez, and D. Erdogmus, "Estimation of the mixing matrix for underdetermined blind source separation using spectral estimation techniques," in *Proc. EUSIPCO 2002*, Sep 2002, vol. 1, pp. 557–560.

[8] A. Papoulis and S.U. Pillai, *Probability, random variables, and stochastic processes*, McGraw-Hill, 4th edition, 2002.

[9] J.G. Proakis and D.G. Manolakis, *Digital signal processing*, Prentice Hall, 3rd edition, 1996.

[10] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 7 edition, Aug 2004.

[11] H. Buchner, R. Aichner, and W. Kellermann, *Audio signal processing for next-generation multimedia communication systems*, chapter Blind Source Separation for convolutive mixtures: a unified treatment, Kluwer, 2004.

[12] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide – Revision 2.0," Tech. Rep. 1706, IRISA, April 2005.