UNDERDETERMINED SOURCE SEPARATION FOR COLORED SOURCES

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

[†]NTT Communication Science Laboratories,
NTT Corporation
2-4 Hikaridai Seika-cho Soraku-gun Kyoto
619-0237 Japan

[‡]Department of Multimedia Communication and Signal Processing, University Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen
Germany

ABSTRACT

This contribution focuses on the source separation stage as important part of underdetermined blind source separation (BSS). So far nearly all approaches for underdetermined BSS assume independently, identically distributed (i.i.d.) sources. They completely ignore the redundancy that is in the temporal structure of colored sources like speech signals. Instead, we propose multivariate models based on the multivariate Student's t or multivariate Gaussian distribution and investigate their potential for underdetermined BSS. We provide a simple yet effective filter based on the sources' autocorrelations for recovering the sources as basis for further advances in underdetermined BSS. The challenge is estimating the filter coefficients blindly. The experimental results support the idea that source separation for underdetermined BSS can be reduced to the separation of their autocorrelations.

1. INTRODUCTION

Blind source separation (BSS) describes techniques that aim at separating P signals if only Q mixed versions of the original signals are available. The need for BSS arises for example if the signals of simultaneous conversations are captured by several microphones. Most BSS approaches assume that there are at least as many microphones as source signals ($Q \geq P$), which is called (over-) determined BSS.

Instead, we consider underdetermined BSS, where we have less microphones than source signals ($Q < P$). Only few approaches have been proposed so far [2, 5, 17, 18], and the separation quality in terms of interference suppression and signal distortion is still not as good as with (over-) determined BSS. This is in particular true if wideband signals like speech signals are involved. The difficulty is that in contrast to (over-) 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.

So far, nearly all approaches to solve the latter problem assume independently, identically distributed (i.i.d.) sources [2, 5, 17, 18]. While this assumption may serve well as a first order approximation, it completely ignores the redundancy that is in the temporal structure of colored sources like speech signals. Therefore, we propose multivariate models, that take the temporal correlation explicitly into account and investigate their potential. We consider here only linear, instantaneous mixtures of speech signals in the time domain. This paper concentrates on the separation and assumes that the mixing matrix is known.

After formulating the problem analytically in Sec. 2, we provide statistical models for Bayesian inference in Sec. 3. In Sec. 4 we propose a closed-form and numerical approach for minimum mean square error (MMSE) source estimation based on the posterior distribution of the sources. Section 5 elaborates more on autocorrelation estimation, which turns out to be essential for high-quality underdetermined source separation. Section 6 presents experimental results which are discussed in Section 7.

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

2. PROBLEM FORMULATION

With $s_p(t) \in \mathbb{R}$ denoting the p-th source signal ($1 \leq p \leq P$) and $\tilde{\boldsymbol{A}} \in \mathbb{R}^{Q \times P}$ the mixing matrix, we obtain mixed signals $x_q(t) \in \mathbb{R}$ ($1 \leq q \leq Q$, $Q < P$) by

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_Q(t) \end{bmatrix} = \tilde{\boldsymbol{A}} \begin{bmatrix} s_1(t) \\ \vdots \\ s_P(t) \end{bmatrix} + \boldsymbol{n}(t). \tag{1}$$

$\boldsymbol{n}(t) \in \mathbb{R}^P$ denotes noise added to the sensors. Let

$$\boldsymbol{S}_p = [\ s_p(0) \quad \cdots \quad s_p(T-1)\ ]^{\mathcal{T}}, \qquad 1 \leq n \leq P \tag{2}$$

denote a frame of length T of the p-th original speech signal with autocorrelation

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

$(\cdot)^{\mathcal{T}}$ denotes the transpose. We summarize the P frames of the different source signals by the vector

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

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

$$[\ r_p(0) \quad \cdots \quad r_p(T-1)\ ] \tag{5}$$

We summarize the P autocorrelation matrices by

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

Extending the scalar elements $\tilde{A}_{qp}$ of $\tilde{\boldsymbol{A}}$ to diagonal matrices

$$\boldsymbol{A}_{qp} = \begin{bmatrix} \tilde{A}_{qp} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \tilde{A}_{qp} \end{bmatrix} \in \mathbb{R}^{T \times T}, \tag{7}$$

we define an extended mixing matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_{Q1} & \cdots & \boldsymbol{A}_{QP} \end{bmatrix} \in \mathbb{R}^{Q \cdot T \times P \cdot T} \tag{8}$$

Similar to the source signals we define

$$\boldsymbol{X}_q = [\ x_q(0) \quad \cdots \quad x_q(T-1)\ ]^{\mathcal{T}}, \qquad 1 \leq q \leq Q \tag{9}$$

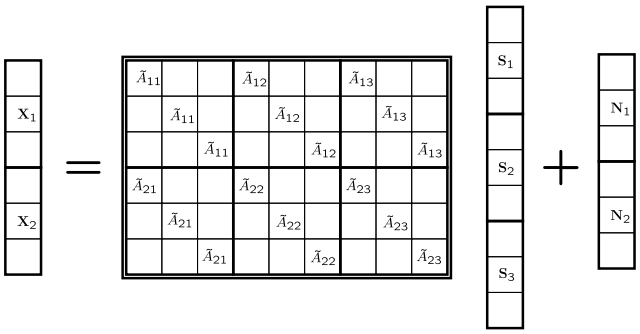

Figure 1: Mixing process for underdetermined BSS with $Q = 2$, $P = 3$

and summarize the $\boldsymbol{X}_q$ by

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_Q \end{bmatrix} \in \mathbb{R}^{Q \cdot T} \tag{10}$$

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

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{N} \tag{11}$$

as illustrated in Fig. 1. $\boldsymbol{N} \in \mathbb{R}^{Q \cdot T}$ is derived from $\boldsymbol{n}(t)$ in a similar way as $\boldsymbol{X}$ is derived 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 (over-) determined BSS it is sufficient to estimate the mixing matrix $\boldsymbol{A}$ or its inverse. Since the mixing matrix is square and therefore invertible (assuming that $\boldsymbol{A}$ is non-singular), the inverse can be used to separate the signals. In contrast to (over-) determined BSS, here the estimation of the mixing matrix $\boldsymbol{A}$ is not sufficient. Even if $\boldsymbol{A}$ is available, estimating the original signals from the mixtures poses a problem on its own, since $\boldsymbol{A}$ cannot be simply inverted. In the following we concentrate on estimating the original signals and assume that the mixing matrix is known or can otherwise be estimated [2, 14, 15, 16].

3. MODELS

The most general approach for estimating the unknown source signals is based on Bayesian inference. It yields the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ of the desired parameters $\boldsymbol{\theta}$ (here: source signals) by accounting for their prior distribution $p(\boldsymbol{\theta})$ and the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ of the observed data $\boldsymbol{X}$ (here: mixed signals). From Bayes' rule we obtain [12]

$$p(\boldsymbol{\theta}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \tag{12}$$

So far, nearly all approaches to underdetermined BSS assume independent and identical prior distributions. In contrast, we propose in this section two multivariate priors tailored to speech signals. We also define the likelihood derived from the commonly assumed Gaussian sensor noise model.

In both speech models we make the common assumption that the speech signals are mutually independent.

$$p(\boldsymbol{S}) = \prod_p p(\boldsymbol{S}_p) \tag{13}$$

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

In our first speech model, we model each frame $\boldsymbol{S}_p$ by a zero mean, multivariate Student's t distribution

$$p(\boldsymbol{S}_p) = \mathcal{ST}(\boldsymbol{S}_p|\alpha_p, \boldsymbol{\Sigma}_p) \tag{14}$$

with degrees of freedom α_p and scale $\boldsymbol{\Sigma}_p$. Modeling the speech signal by a multivariate Student's t distribution has several advantages:

- It is a good approximation of real speech signals. This does not just follow from experimental results but is also based on results with spherically invariant random processes (SIRPs) [3] and the univariate Student's t distribution [5].
- The multivariate Student's t distribution can be expressed as a multivariate Gaussian mixture model (MGM) with an infinite number of mixtures [1] as follows

$$\mathcal{ST}(\boldsymbol{S}_p|\alpha_p, \boldsymbol{\Sigma}_p) = \int \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0}, \boldsymbol{R}_p) \cdot \mathcal{IW}(\boldsymbol{R}_p|\alpha_p, \boldsymbol{\Sigma}_p) d\boldsymbol{R}_p. \tag{15}$$

  $\mathcal{N}$ and $\mathcal{IW}$ denote a multivariate Gaussian and inverse Wishart distribution, respectively. According to this relation, the multivariate Student's t distribution is the marginal distribution obtained from the joint distribution of the speech signal and its autocorrelation. This leads to a hierarchical Bayesian model with a multivariate Gaussian distribution for the speech signal $\boldsymbol{S}_p$

$$p(\boldsymbol{S}_p|\boldsymbol{R}_p) = \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0}, \boldsymbol{R}_p) \tag{16}$$

  and an inverse Wishart distribution for the autocorrelation matrix $\boldsymbol{R}_p$

$$p(\boldsymbol{R}_p|\alpha_p, \boldsymbol{\Sigma}_p) = \mathcal{IW}(\boldsymbol{R}_p|\alpha_p, \boldsymbol{\Sigma}_p). \tag{17}$$

  The expected value of $\boldsymbol{R}_p$ is [8]

$$E\{\boldsymbol{R}_p\} = \frac{1}{\alpha_p - T - 1} \boldsymbol{\Sigma}_p^{-1}. \tag{18}$$

  The hierarchical model will be used in Sec. 4 to obtain conditional posterior PDFs.
- The inverse Wishart distribution is the conjugate prior of a multivariate Gaussian likelihood with respect to its correlation matrix [7]. This means that the corresponding posterior distribution is again an inverse Wishart distribution, which is also exploited in Sec. 4.

The degrees of freedom α_p and the scale $\boldsymbol{\Sigma}_p$ in (17) can either be considered deterministic or random. We summarize all parameters in the parameter vector $\boldsymbol{\theta}$ and obtain the joint prior distribution as

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = p(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{A}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \tag{19}$$

$$= p(\boldsymbol{S}|\boldsymbol{R}) p(\boldsymbol{R}|\boldsymbol{\alpha}, \boldsymbol{\Sigma}) p(\boldsymbol{A}) p(\sigma^2) p(\boldsymbol{\alpha}) p(\boldsymbol{\Sigma}) \tag{20}$$

with $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_P \end{bmatrix}^T$. The relation between the different variables is illustrated in Fig. 2.

As a special case of the first speech model we propose an analytically feasible, multivariate Gaussian prior as second speech model

$$p(\boldsymbol{S}_p) = \lim_{\alpha_p \to \infty} \mathcal{ST}(\boldsymbol{S}_p|\alpha_p, \boldsymbol{\Sigma}_p) = \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0}, \boldsymbol{R}_p) \tag{21}$$

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

For deriving the likelihood we assume additive white Gaussian noise with identical variance σ^2 across all sensors. Based on the mixing process (11), this assumption results in the Gaussian likelihood

$$p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{S}, \sigma^2) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{A}\boldsymbol{S}, \sigma^2). \tag{22}$$

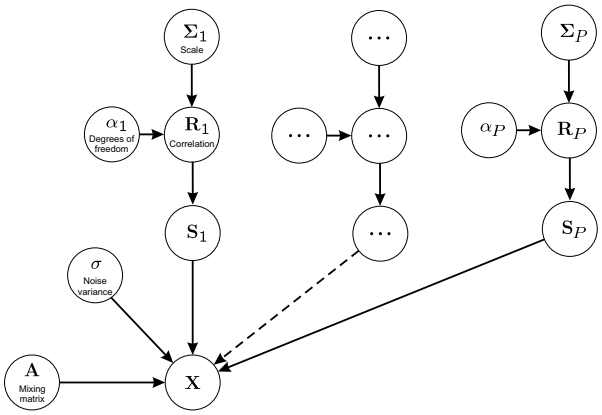

Figure 2: Graphical model

4. BAYESIAN INFERENCE

The general goal in Bayesian inference is to estimate the posterior PDF of unknown parameters based on the available data. Except for few special cases like the Gaussian speech model (21) with Gaussian likelihood (22), the posterior cannot be derived in closed-form. Instead, it requires numerical methods based for example on Monte Carlo techniques [4]. Based on our multivariate models in the previous section, we derive in the following MMSE estimates of the sources in closed-form and by numerical approximation. Further details are provided in the appendix.

4.1 Multivariate Student's t prior

In order to estimate the sources based on the multivariate Student's t prior (14), we employ a Gibbs sampler, which is a Monte Carlo technique. In contrast to our multivariate approach, the Gibbs sampler was applied in [5] to an i.i.d. speech model based on a univariate Student's t distribution.

4.1.1 Basic idea

The Gibbs sampler approximates the posterior PDF by sampling iteratively from appropriately chosen conditional PDFs. In fact, the conditional PDFs form a first-order Markov chain, whose stationary PDF is the desired posterior PDF [4, 8].

Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$ denote the unknown parameters with prior PDF $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $\boldsymbol{X}$ the available data with likelihood $p_{\boldsymbol{X}|\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{\theta})$. We further define $\boldsymbol{\theta}_{-i}$ as summarizing all parameters without parameter $\boldsymbol{\theta}_i$. For the Gibbs sampler we need the conditional PDFs $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{X})$, which can be shown to be [5]

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{X}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) p_{\boldsymbol{X}|\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{\theta}) \tag{23}$$

Once they are derived, we can sample iteratively from each conditional PDF using the previous samples as summarized in Algorithm 1. As soon as a sufficient number of samples is drawn, they approximate the posterior PDF. They can further be used to calculate point estimates $\boldsymbol{Y}$ (MAP, MMSE, ...) of the original signals $\boldsymbol{S}$. In order to avoid transient effects a certain number of initial samples should be discarded [9, 13].

Algorithm 1: General Gibbs sampler

```
Initialize θ (e.g. randomly) ;
for k=1..K do
    for i do
        θ_i^(k+1) ~ p(θ_i | θ_-i^(k,k+1), x);
    end
end
```

4.1.2 Implementation

We now derive the conditional PDFs for our specific problem based on the multivariate Student's t speech model in Sec. 3.

For the conditional PDF of the speech signal we obtain

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{\theta}_{-\boldsymbol{S}}, \boldsymbol{X}) &\propto p(\boldsymbol{S}|\boldsymbol{R}) \cdot p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{S}, \sigma^2) && (24)\\ &\propto \mathcal{N}(\boldsymbol{S}|\boldsymbol{0}, \boldsymbol{R}) \cdot \mathcal{N}(\boldsymbol{X}|\boldsymbol{A}\boldsymbol{S}, \sigma^2) && (25)\\ &\propto \mathcal{N}(\boldsymbol{S}|\boldsymbol{\mu}_{\boldsymbol{S}}, \boldsymbol{\Sigma}_{\boldsymbol{S}}) && (26) \end{aligned}$$

with

$$\boldsymbol{\Sigma}_{\boldsymbol{S}} = \left(\boldsymbol{R}^{-1} + \frac{1}{\sigma^2} \boldsymbol{A}^{\mathcal{T}} \boldsymbol{A} \right)^{-1} \tag{27}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{S}} = \left(\sigma^2 \boldsymbol{R}^{-1} + \boldsymbol{A}^{\mathcal{T}} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^{\mathcal{T}} \boldsymbol{X}. \tag{28}$$

For the conditional PDF of the autocorrelation we obtain

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{\theta}_{-\boldsymbol{R}}, \boldsymbol{X}) &\propto p(\boldsymbol{S}|\boldsymbol{R}) \cdot p(\boldsymbol{R}|\boldsymbol{\alpha}, \boldsymbol{\Sigma}) && (29)\\ &\propto \prod_p \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0}, \boldsymbol{R}_p) \cdot \mathcal{IW}(\boldsymbol{R}_p|\alpha_p, \boldsymbol{\Sigma}_p)\\ &\propto \prod_p \mathcal{IW}\left(\boldsymbol{R}_p|\alpha_p + 1, \left(\boldsymbol{S}_p \boldsymbol{S}_p^{\mathcal{T}} + \boldsymbol{\Sigma}_p^{-1} \right)^{-1} \right) \end{aligned}$$

To successfully implement this Gibbs sampler, appropriate choices for the degrees of freedom α_p and the scale $\boldsymbol{\Sigma}_p$ are necessary. It is common in Bayesian inference to manually tune the degrees of freedom. Usually there is also sufficient prior data available to train the scale (e.g. speaker adaptation for speech recognition). Here we face the problem, that we do not have sufficient prior data. We consider the following options:

- We could consider the scale as random variable as well and choose an uninformative prior (e.g. Jeffrey's prior, uniform prior, ...). So far experiments with uninformative priors did not yet lead to reasonable results. Therefore, we don't consider this option here.
- Knowing that the choice of the scale $\boldsymbol{\Sigma}_p$ directly influences the autocorrelation by (18), we can set it to a specific value, determined for example by some preprocessing (e.g. ℓ_1-norm minimization [17]) as detailed in the next section.

4.2 Multivariate Gaussian prior

Plugging the Gaussian prior (21) and the likelihood (22) in (12) we obtain for the Gaussian posterior similar to (24)

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{X}) &\propto p(\boldsymbol{S}|\boldsymbol{R}) \cdot p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{S}, \sigma^2) && (30)\\ &\propto \mathcal{N}(\boldsymbol{S}|\boldsymbol{\mu}_{\boldsymbol{S}}, \boldsymbol{\Sigma}_{\boldsymbol{S}}). \end{aligned}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{S}}$ and $\boldsymbol{\mu}_{\boldsymbol{S}}$ given by (27) and (28), respectively.

Since the MMSE estimate of a normally distributed random variable is its mean, the resulting estimate of the sources is

$$\boldsymbol{Y} = \underbrace{\left(\sigma^2 \boldsymbol{R}^{-1} + \boldsymbol{A}^{\mathcal{T}} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^{\mathcal{T}}}_{=\boldsymbol{W}} \boldsymbol{X}. \tag{31}$$

The unmixing filter $\boldsymbol{W}$ is a generalized inverse modified by the autocorrelation matrix. The influence of the autocorrelation matrix $\boldsymbol{R}$ is determined by the noise variance σ^2. Without modification by the autocorrelation matrix (31) could not be determined, since $\boldsymbol{A}^{\mathcal{T}} \boldsymbol{A}$ is a singular matrix.

5. AUTOCORRELATION

For the proposed speech models the autocorrelation plays a crucial role. With the multivariate Gaussian model, the estimate is used directly. With the multivariate Student's t model, the estimate can be used to set the scale $\mathbf{\Sigma}_p$, if it is treated as a deterministic value. It then determines the expected value of the autocorrelation. Therefore, for high quality separation in terms of interference suppression and signal distortion [6] we need a good estimate of the autocorrelation. In this section we will discuss possible options for estimating the autocorrelation. In general, we can distinguish between non-parametric and parametric approaches.

5.1 Non-parametric estimation

Non-parametric estimation has the advantage that no assumption about the structure is necessary. For our purpose with only a limited number T of samples of the signal in question available, we choose a biased estimation technique [10]. It guarantees that the estimated autocorrelation is positive semi-definite. For a signal frame $s(t)$, $0 \leq t \leq T-1$ the biased autocorrelation estimate is given by

$$r(\tau) = E\{s(t)s(t+\tau)\} \tag{32}$$

$$\approx \frac{1}{T}\sum_{t=0}^{T-\tau-1} s(t)s(t+\tau) \tag{33}$$

5.2 Parametric estimation

Non-parametric estimation of the autocorrelation suffers from an insufficient amount of data compared to the number of parameters (if we define all time lags of the autocorrelation as parameters). This results in high variance and low resolution [11]. In general, it is desirable to reduce the number of parameters by an appropriate parametric model that exploits the inherent structure.

Based on the harmonicity of speech we propose approximating the biased autocorrelation by M harmonic, decaying cosine functions.

$$r_p(\tau) \approx \left(\sum_{m=1}^{M} \gamma_{m,p}\cos(p\cdot f_{0,p}\cdot\tau)\right)\exp(-\beta_p\tau) \tag{34}$$

Here, $\gamma_{m,p}$ denotes the gain of each cosine, $f_{0,p}$ is the fundamental frequency and β_p is the decay factor. This reduces the number of parameters to $M+2$ for each source.

5.2.1 *Least square estimation*

If we have access to the autocorrelation that we want to approximate, we can use a least square approach to estimate the parameters. We apply this method to the following scenarios.

- To explore the possibilities of the parametric approach, we can use the autocorrelations of the original signals to estimate the parameters.
- Other parameter estimation methods might need to be initialized. Then we can use initial estimates of the sources (e.g. ℓ_1-norm minimization [17]) and use their autocorrelations.
- Linear mixtures of mutually independent signals have the property, that their autocorrelations are a linear mixture of the autocorrelation of the original signals.

$$r_{xq}(\tau) = E\{x_q(t)x_q(t-\tau)\} \tag{35}$$

$$= E\left\{\left(\sum_p \tilde{A}_{qp}s_p(t)\right)\cdot\left(\sum_p \tilde{A}_{qp}s_p(t-\tau)\right)\right\}$$

$$= E\left\{\sum_p \tilde{A}_{qp}^2 s_p(t)s_p(t-\tau)\right\} \tag{36}$$

$$= \sum_p \tilde{A}_{qp}^2 E\{s_p(t)s_p(t-\tau)\} \tag{37}$$

$$= \sum_p \tilde{A}_{qp}^2 r_{s_p}(\tau) \tag{38}$$

If we assume, that the fundamental frequency $f_{0,p}$ is different for each speaker, we can estimate the parameters from the mixtures for each source in parallel.

In each scenario a non-parametric estimation is the basis for the parameter estimation. As a general least square cost function for all scenarios we propose

$$\mathcal{J}(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{f}_0) = \sum q\sum_\tau (rq(\tau) - \tag{39}$$

$$\sum_p \tilde{A}_{q,p}^2\left(\left(\sum_{m=1}^{M}\gamma_{mp,p}\cos(m\cdot f_{0,p}\cdot\tau)\right)\exp(-\beta_p\tau)\right)\Bigg)^2$$

6. EXPERIMENTAL RESULTS

We performed experiments with three speech signals (two male, one female) of 1.62 seconds (8kHz sampling rate). We generated two instantaneous mixtures with the mixing matrix

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 1.3 & -0.9 & 0.8 \end{bmatrix} \tag{40}$$

and assumed that the mixing matrix is available. The signals were processed in frames of $T = 256$ samples and shifted by 64 samples. No noise was added to the mixtures, but the noise variance was set to $\sigma^2 = 10^{-4}$. This led to a good influence of the autocorrelation matrix $\boldsymbol{R}$. For approximations of the autocorrelation we set $M = 2$. The sources had equal variance of 0.0037. We compared the following approaches:

L1	Separating the mixtures by ℓ_1-norm minimization in the frequency domain, which is based on a univariate speech model [17].
L1G	Using the results from L1 to estimate the autocorrelation. This estimate was used to further enhance the separated signals by the Gibbs sampler (24)-(29) ($K = 1000$ iterations).
OG	Estimating the signals with the Gibbs sampler (24)-(29) ($K = 10$ iterations), whereby the autocorrelation was estimated from the original signals (33).
OI	With the autocorrelation estimated from the original signals (33), the signals were obtained with the modified generalized inverse (31).
OAI	Estimating the signals with the modified generalized inverse (31), whereby each autocorrelation was approximated by two harmonic, exponentially weighted cosines based on the original signals (34).
OSI	Estimating the signals with the modified generalized inverse (31), whereby the autocorrelation matrix was a diagonal matrix filled with the first lag of the autocorrelation of the original signals (33).
MAI	Estimating the signals with the modified generalized inverse (31), whereby the autocorrelation was approximated directly from the mixtures (34).

For the evaluation 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 [6]. The averaged results of the different approaches are shown in Table 1.

7. CONCLUSION

The experimental results in Table 1 together with subjective evaluation suggest that knowing the autocorrelation together with the mixing matrix is sufficient to perform high quality underdetermined source separation in noiseless environments (OG and OI compared to L1G, OSI and MAI). 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. The

Method	SDR	SIR	SAR
L1	9.28	14.84	11.79
L1G	5.29	17.65	7.25
OG	12.3	35.61	12.56
OI	18.21	34.33	18.51
OAI	10.42	20.75	11.40
OSI	11.11	18.08	12.89
MAI	3.45	13.63	5.83

Table 1: Experimental results

proposed parameterization of the autocorrelation did not yet lead to satisfying results (OAI, MAI). Therefore, either better autocorrelation models or improved non-parametric estimation methods are necessary.

The experimental results obtained by the MMSE approach OI resemble the results of the Gibbs sampler OG. This suggests that the exact prior distribution of the sources plays only a minor role as long as the autocorrelation is taken into account. Therefore, the comparatively easy and fast solution given by (31) might provide a basis for algorithms that yield better autocorrelation estimates and eventually underdetermined source separation.

Possible options for improving the autocorrelation estimate include linear prediction based models and independent component analysis (ICA) techniques. Linear prediction has already proven in speech coding, that it allows very compact yet high quality descriptions of speech signals. ICA exploits the statistical independence between the sources and might be helpful for estimating the involved parameters.

A. CONDITIONAL PDFS FOR GIBBS SAMPLER

- Conditional PDF of speech signal

$$
\begin{aligned}
p(\boldsymbol{S}|\boldsymbol{\theta}_{-\boldsymbol{S}},\boldsymbol{X}) &\propto p(\boldsymbol{S}|\boldsymbol{R})\cdot p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{S},\sigma^2) \qquad (41)\\
&\propto \mathcal{N}(\boldsymbol{S}|\boldsymbol{R})\cdot\mathcal{N}(\boldsymbol{X}|\boldsymbol{A}\boldsymbol{S},\sigma^2)\\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{S}^{\mathcal{T}}\boldsymbol{R}^{-1}\boldsymbol{S}\right)\cdot\\
&\quad \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})^{\mathcal{T}}\frac{\boldsymbol{I}}{\sigma^2}(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\underbrace{\boldsymbol{S}^{\mathcal{T}}\left(\boldsymbol{R}^{-1}+\frac{1}{\sigma^2}\boldsymbol{A}^{\mathcal{T}}\boldsymbol{A}\right)\boldsymbol{S}}_{:=\boldsymbol{S}^{\mathcal{T}}\boldsymbol{\Sigma}_{\boldsymbol{S}}^{-1}\boldsymbol{S}}-\right.\right.\\
&\quad \left.\left.-\boldsymbol{S}^{\mathcal{T}}\underbrace{\frac{1}{\sigma^2}\boldsymbol{A}^{\mathcal{T}}\boldsymbol{X}}_{:=\boldsymbol{\Sigma}_{\boldsymbol{S}}^{-1}\boldsymbol{\mu}_{\boldsymbol{S}}}-\frac{1}{\sigma^2}\boldsymbol{X}^{\mathcal{T}}\boldsymbol{A}\boldsymbol{S}+\frac{1}{\sigma^2}\boldsymbol{X}^{\mathcal{T}}\boldsymbol{X}\right)\right)\\
&= \mathcal{N}(\boldsymbol{S}|\boldsymbol{\mu}_{\boldsymbol{S}},\boldsymbol{\Sigma}_{\boldsymbol{S}})
\end{aligned}
$$

with $\boldsymbol{\Sigma}_{\boldsymbol{S}}$ and $\boldsymbol{\mu}_{\boldsymbol{S}}$ given by (27) and (28), respectively.

- Conditional PDF of autocorrelation

$$
\begin{aligned}
p(\boldsymbol{R}|\boldsymbol{\theta}_{-\boldsymbol{R}},\boldsymbol{X}) &\propto p(\boldsymbol{S}|\boldsymbol{R})\cdot p(\boldsymbol{R}|\boldsymbol{\alpha},\boldsymbol{\Sigma}) \qquad (42)\\
&\propto \prod_p \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0},\boldsymbol{R}_p)\cdot\prod_p \mathcal{IW}(\boldsymbol{R}_p|\alpha_p,\boldsymbol{\Sigma}_p)\\
&\propto \prod_p \mathcal{N}(\boldsymbol{S}_p|\boldsymbol{0},\boldsymbol{R}_p)\cdot\mathcal{IW}(\boldsymbol{R}_p|\alpha_p,\boldsymbol{\Sigma}_p)\\
&\propto \prod_p |\boldsymbol{R}_p|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{S}_p^{\mathcal{T}}\boldsymbol{R}_p^{-1}\boldsymbol{S}_p\right)\cdot\\
&\quad |\boldsymbol{R}_p|^{-\frac{1}{2}(\alpha_p+T+1)}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{R}_p^{-1}\boldsymbol{\Sigma}_p^{-1}\right)\right)\\
&\propto \prod_p |\boldsymbol{R}_p|^{-\frac{1}{2}(\alpha_p+T+2)}\cdot\\
&\quad \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{R}_p^{-1}\boldsymbol{S}_p\boldsymbol{S}_p^{\mathcal{T}}\right)-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{R}_p^{-1}\boldsymbol{\Sigma}_p^{-1}\right)\right)\\
&= \prod_p \mathcal{IW}\left(\boldsymbol{R}_p|\alpha_p+1,\left(\boldsymbol{S}_p\boldsymbol{S}_p^{\mathcal{T}}+\boldsymbol{\Sigma}_p^{-1}\right)^{-1}\right)
\end{aligned}
$$

References

[1] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. Wiley series in probability and statistics. Wiley, 2000.

[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*, E88-A(7):1693–1700, 2005.

[3] H. Brehm and W. Stammler. Description and generation of spherically invariant speech-model signals. *Signal Processing*, 12:119–141, 1987.

[4] A. Doucet and X. Wang. Monte Carlo methods for signal processing: a review in the statistical signal processing context. *IEEE Signal Processing Magazine*, 22(6):152–170, Nov 2005.

[5] C. Févotte and S. J. Godsill. A bayesian approach for blind separation of sparse sources. Technical report, Cambridge University Engineering Dept., January 2005.

[6] C. Févotte, R. Gribonval, and E. Vincent. BSS_EVAL toolbox user guide – Revision 2.0. Technical Report 1706, IRISA, April 2005.

[7] C. Fraley and A.E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Technical Report 486, University of Washington, Statistics, August 2005.

[8] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall, 1995.

[9] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 7 edition, Aug 2004.

[10] D.G. Manolakis, V.K. Ingle, and S.M. Kogon. *Statistical and adaptive signal processing*. McGraw-Hill, 2000.

[11] J.M. Mendel. Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305, March 1991.

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

[13] D.B. Rowe. *Multivariate Bayesian statistics: models for source separation and signal unmixing*. Chapman & Hall/CRC, 2003.

[14] F.J. Theis. *Mathematics in independent component analysis*. PhD thesis, University of Regensburg, 2002.

[15] 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*, volume 1, pages 557–560, Sep 2002.

[16] S. Winter, H. Sawada, S. Araki, and S. Makino. Overcomplete BSS for convolutive mixtures based on hierarchical clustering. In *Proc. ICA 2004*, pages 652–660, Sept. 2004.

[17] 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 Processingto Audio and Acoustics (WASPAA)*, pages 86–89, New Paltz, NY, USA, 2005.

[18] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition. *Neural Computations*, 13(4):863 – 882, 2001.