Cogn Comput (2013) 5:493–503
DOI 10.1007/s12559-012-9185-9

Improved Convolutive and Under-Determined Blind Audio Source Separation with MRF Smoothing

Rafał Zdunek

Received: 22 January 2012 / Accepted: 17 August 2012 / Published online: 7 September 2012

Abstract Convolutive and under-determined blind audio source separation from noisy recordings is a challenging problem. Several computational strategies have been proposed to address this problem. This study is concerned with several modifications to the expectation-minimization-based algorithm, which iteratively estimates the mixing and source parameters. This strategy assumes that any entry in each source spectrogram is modeled using superimposed Gaussian components, which are mutually and individually independent across frequency and time bins. In our approach, we resolve this issue by considering a locally smooth temporal and frequency structure in the power source spectrograms. Local smoothness is enforced by incorporating a Gibbs prior in the complete data likelihood function, which models the interactions between neighboring spectrogram bins using a Markov random field. Simulations using audio files derived from stereo audio source separation evaluation campaign 2008 demonstrate high efficiency with the proposed improvement.

Keywords Blind source separation · Nonnegative matrix factorization · Expectation-maximization · Markov random field · Simultaneous auto-regression

Introduction

Blind source separation (BSS) aims to recover unknown source signals from observed mixtures with or without very limited information about their mixing process. BSS problems have been addressed in many previous studies, for example, [1–12], which were motivated by several real-world applications.

In a cocktail-party problem, microphones receive noisy mixtures of acoustic signals that propagate along multiple paths from their sources. In a real scenario, the number of audio sources may be greater than the number of microphones, audio sources may have different timbres and similar pitches, and audio signals may be only locally stationary.

A convolutive and under-determined mixing approach needs to be adopted to model this problem. There are several techniques for solving convolutive unmixing problems [13]. Some of these [14] operate in the time-domain by solving the alternative finite impulse response (FIR) inverse model using independent component analysis (ICA) methods [2]. Another method is to extract meaningful features from the time-frequency (TF) representations of mixtures. This approach seems to be more efficient than the ICA-based techniques especially when the number of microphones is lower than the number of sources. Acoustic signals are usually sparse in the TF domain, so the source signals can be separated efficiently even if they are partially overlapped and the problem is under-determined. These features can be extracted using several techniques, including TF masking [15, 16], frequency bin-wise clustering with permutation alignment (FBWC-PA) [17, 18], subspace projection [19], hidden Markov models (HMM) [20], interaural phase difference (IPD) [21], nonnegative matrix factorization (NMF) [22, 23], and nonnegative tensor factorization (NTF) [24].

Nonnegative matrix factorization [25] is a feature extraction method with many real-world applications [26]. A convolutive NMF-based unmixing model was proposed by Smaragdis [22]. Ozerov and Fevotte [23] developed the

R. Zdunek (✉)
Institute of Telecommunications, Teleinformatics,
and Acoustics, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
e-mail: rafal.zdunek@pwr.wroc.pl

EM-NMF algorithm, which is suitable for unsupervised convolutive and possibly under-determined unmixing of audio sources using only stereo observations. Their model of the sources was based on the generalized Wiener filtering model [27–29], which assumes that each source is locally stationary and that it can be expressed in terms of superimposed amplitude-modulated Gaussian components. Thus, a power spectrogram of each source can be factorized into lower-rank nonnegative matrices, which facilitates the use of NMF for estimating the frequency and temporal profiles of each latent source component. In the TF representation, the latent components are mutually and individually independent across frequency and time bins. However, this assumption is very weak for any adjacent bins because real audio signals have locally smooth frequency and temporal structures.

Motivated by several papers on smoothness [26, 28, 30, 31] in BSS models, we attempt to further improve the EM-NMF algorithm by enforcing local smoothness both in the frequency and temporal profiles of the NMF factors. Similar to [28, 30, 32], we introduce a priori knowledge to the NMF-based model using a Bayesian framework, although our approach is based on a Gibbs prior with a Markov random field (MRF) model to describe pairwise interactions among adjacent bins in spectrograms. As demonstrated in [33], the MRF model with Green's function, which is well known in many tomographic image reconstruction applications [34], can improve the EM-NMF algorithm. In this paper, we extend the results presented in [33] using other smoothing functions, particularly a more flexible simultaneous autoregressive (SAR) model that is more appropriate in term of hyperparameter estimation and computational complexity.

The rest of this paper is organized as follows. The next section reviews the underlying separation model. Section 3 is concerned with MRF smoothing. The optimization algorithm is described in Sect. 4. Audio source separation experiments are presented in Sect. 5. Finally, the conclusions are provided in Sect. 6.

Model

Let I microphones receive signals that can be modeled as a noisy convolutive mixture of J audio signals. The signal received by the i-th microphone ($i = 1, \ldots, I$) can be expressed as

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{l=0}^{L-1} \tilde{a}_{ijl} \tilde{s}_j(t-l) + \tilde{n}_i(t), \tag{1}$$

where $\tilde{a}_{ijl}$ represents the corresponding mixing filter coefficient, $\tilde{s}_j(t)$ is the j-th source signal ($j = 1, \ldots, J$), $\tilde{n}_i(t)$ is the additive noise, and L is the length of the mixing filter.

In the TF domain, the model (1) can be expressed as

$$x_{ift} = \sum_{j=1}^{J} a_{ijf} s_{jft} + n_{ift}, \quad \text{or equivalently,} \qquad \boldsymbol{X}_f = \boldsymbol{A}_f \boldsymbol{S}_f + \boldsymbol{N}_f, \tag{2}$$

where $\boldsymbol{X}_f = [x_{ift}]_f \in \mathbb{C}^{I \times T}, \boldsymbol{A}_f = [a_{ijf}]_f \in \mathbb{C}^{I \times J}, \boldsymbol{S}_f = [s_{jft}]_f \in \mathbb{C}^{J \times T}, \boldsymbol{N}_f = [n_{ift}]_f \in \mathbb{C}^{I \times T}$, and $f = 1, \ldots, F$ is the index of a frequency bin.

The noise n_{ift} is assumed to be stationary and spatially uncorrelated, i.e,

$$n_{ift} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_n), \tag{3}$$

where $\boldsymbol{\Sigma}_n = \operatorname{diag}\left([\sigma_i^2]\right)$ and $\mathcal{N}_c(0, \boldsymbol{\Sigma}_n)$ is a proper complex Gaussian distribution with a zero-mean and the covariance matrix $\boldsymbol{\Sigma}_n$.

Benaroya et al. [27] described an audio source $\tilde{s}(t)$ as a superimposed amplitude-modulated Gaussian process:

$$\tilde{s}(t) = \sum_{r=1}^{R} h_r(t) \tilde{w}_r(t), \tag{4}$$

where $h_r(t)$ is a slowly varying amplitude parameter in the r-th component ($r = 1, \ldots, R$), and $\tilde{w}_r(t)$ is a stationary zero-mean Gaussian process with the power spectral density $\sigma_r^2\,(f)$. The TF representation of (4) leads to

$$s(f,t) \sim \mathcal{N}_c\left(0, \sum_{r=1}^{R} h_r(t) \sigma_r^2(f)\right). \tag{5}$$

The power spectrogram of (5) is given by $|s_{ft}|^2 = \sum_{r=1}^{R} w_{fr} h_{rt}$, where $w_{fr} = \sigma_r^2(f)$. Thus, the spectrogram of the j-th source $\tilde{s}_j(t)$ can be factorized as follows:

$$|\boldsymbol{S}_j|^2 = \boldsymbol{W}_j \boldsymbol{H}_j, \tag{6}$$

where $\boldsymbol{S}_j \in \mathbb{C}^{F \times T}, \boldsymbol{W}_j \in \mathbb{R}_+^{F \times R_j}, \boldsymbol{H}_j \in \mathbb{R}_+^{R_j \times T}, R_j$ is the number of latent components in the j-th source, and $\mathbb{R}_+$ is the nonnegative orthant of the Euclidean space. The column vectors of $\boldsymbol{W}_j$ represent the frequency profiles of the j-th source, while the row vectors of $\boldsymbol{H}_j$ are the temporal profiles.

Févotte et al. [28] transformed the model (5) to the following form:

$$s(f,t) = \sum_{r=1}^{R} c_r(f,t) \tag{7}$$

where $c_r(f,t) \sim \mathcal{N}_c\left(0, h_r(t)\sigma_r^2(f)\right) = \mathcal{N}_c\left(0, \boldsymbol{\Sigma}_{ft}^{(c)}\right)$. Thus,

$$\boldsymbol{\Sigma}_{ft}^{(c)} = \operatorname{diag}\left([w_{fr} h_{rt}]_r\right) = \operatorname{diag}\left(\left[|c_{rft}|^2\right]_r\right), \tag{8}$$

and $\left[\sum_{r=1}^{R} |c_{rft}|^2\right] = \boldsymbol{W}\boldsymbol{H}$, where $\boldsymbol{W} \in \mathbb{R}_+^{F \times R}, \boldsymbol{H} \in \mathbb{R}_+^{R \times T}$. Consequently, the model (2) can be expressed as

$$x_{ift} = \sum_{r=1}^{R} \bar{a}_{irf} c_{rft} + n_{ift}, \qquad \text{where} \quad s_{jft} = \sum_{r \in \Re_j} c_{rft}, \tag{9}$$

$R = |\Re|$ is the number of entries in the set $\Re = \bigcup_{j=1}^{J} \Re_{\mathrm{j}}$, and $\bar{\boldsymbol{A}}_f = [\bar{a}_{irf}] \in \mathbb{C}^{I \times R}$ is created from the columns of the matrix $\boldsymbol{A}_f$. For example, assuming $\forall j : \Re_{\mathrm{j}} = \{, \ldots, \Re\}$, we have $R = J\bar{R}$, and $\bar{\boldsymbol{A}}_f = [\boldsymbol{A}_f, \ldots, \boldsymbol{A}_f] \in \mathbb{C}^{I \times R}$ is the augmented mixing matrix [23] created from $\bar{R}$ matrices $\boldsymbol{A}_f$. From (8) and (9), we have $s_{jft} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{ft}^{(s)})$ where $\boldsymbol{\Sigma}_{ft}^{(s)} = \mathrm{diag}\left(\left[\sum_{r \in \Re_{\mathrm{j}}} w_{fr} h_{rt}\right]_j\right)$.

To estimate the parameters $\mathcal{A} = [a_{ijf}] \in \mathbb{C}^{I \times J \times F}$, $\mathcal{C} = [c_{rft}] \in \mathbb{C}^{R \times F \times T}, \boldsymbol{W} \in \mathbb{R}_+^{F \times R}, \boldsymbol{H} \in \mathbb{R}_+^{R \times T}$, and $\boldsymbol{\Sigma}_n \in \mathbb{R}_+^{I \times I}$, we formulate the following posterior:

$$P(\mathcal{C}, \boldsymbol{W}, \boldsymbol{H} | \mathcal{X}, \mathcal{A}, \boldsymbol{\Sigma}_n) = \frac{P(\mathcal{X} | \mathcal{C}, \mathcal{A}, \boldsymbol{\Sigma}_n) P(\mathcal{C} | \boldsymbol{W}, \boldsymbol{H}) P(\boldsymbol{W}) P(\boldsymbol{H})}{P(\mathcal{X} | \mathcal{A}, \boldsymbol{\Sigma}_n)}, \tag{10}$$

from which we obtain

$$\begin{aligned} \ln P(\mathcal{X}, \mathcal{C}, \boldsymbol{W}, \boldsymbol{H} | \mathcal{A}, \boldsymbol{\Sigma}_n) = \ln P(\mathcal{X} | \mathcal{C}, \mathcal{A}, \boldsymbol{\Sigma}_n) + \ln P(\mathcal{C} | \boldsymbol{W}, \boldsymbol{H}) \\ + \ln P(\boldsymbol{W}) + \ln P(\boldsymbol{H}). \end{aligned} \tag{11}$$

From (3) and (9), we have the joint conditional PDF for $\mathcal{X}$:

$$\begin{aligned} P(\mathcal{X} | \mathcal{C}, \mathcal{A}, \boldsymbol{\Sigma}_n) &= \prod_{i,f,t} \mathcal{N}_c\left(\sum_{r=1}^{R} \bar{a}_{irf} c_{rft}, \sigma_i^2\right) \\ &= \prod_{f,t} \mathcal{N}_c(\boldsymbol{A}_f \boldsymbol{s}_{ft}, \boldsymbol{\Sigma}_n). \end{aligned} \tag{12}$$

Based on (12), the log-likelihood term in (11) can be expressed as

$$\begin{aligned} \ln P(\mathcal{X} | \mathcal{C}, \mathcal{A}, \boldsymbol{\Sigma}_n) = &-\sum_{f,t} (\boldsymbol{x}_{ft} - \bar{\boldsymbol{A}}_f \boldsymbol{c}_{ft})^H \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{x}_{ft} - \bar{\boldsymbol{A}}_f \boldsymbol{c}_{ft}) \\ &- \sum_{f,t} \ln \det \boldsymbol{\Sigma}_n \\ = &-\sum_{f,t} (\boldsymbol{x}_{ft} - \boldsymbol{A}_f \boldsymbol{s}_{ft})^H \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{x}_{ft} - \boldsymbol{A}_f \boldsymbol{s}_{ft}) \\ &- \sum_{f,t} \ln \det \boldsymbol{\Sigma}_n + \text{const}, \end{aligned} \tag{13}$$

where $\boldsymbol{c}_{ft} = [c_{1ft}, \ldots, c_{Rft}]^T \in \mathbb{C}^R, \boldsymbol{s}_{ft} = [s_{1ft}, \ldots, s_{Jft}]^T \in \mathbb{C}^J$, and $\boldsymbol{x}_{ft} = [x_{1ft}, \ldots, x_{Ift}]^T \in \mathbb{C}^I$.

The joint conditional PDF for $\mathcal{C}$ comes from the model (5):

$$\begin{aligned} P(\mathcal{C} | \boldsymbol{W}, \boldsymbol{H}) &= \prod_{r=1}^{R} \prod_{f=1}^{F} \prod_{t=1}^{T} \mathcal{N}_c(0, w_{fr} h_{rt}) \\ &= \prod_{r=1}^{R} \prod_{f=1}^{F} \prod_{t=1}^{T} |\pi w_{fr} h_{rt}|^{-1} \exp\left\{-\frac{|c_{rft}|^2}{w_{fr} h_{rt}}\right\}. \end{aligned} \tag{14}$$

From (14), we have the log-likelihood functional for $\mathcal{C}$:

$$\ln P(\mathcal{C} | \boldsymbol{W}, \boldsymbol{H}) = -\sum_{r,f,t} \left(\ln(w_{fr} h_{rt}) + \frac{|c_{rft}|^2}{w_{fr} h_{rt}}\right) + \text{const}. \tag{15}$$

The negative log-likelihood in (15) is the Itakura-Saito (IS) divergence [35], which is particularly useful for measuring the goodness of fit between spectrograms. The IS divergence is the special case of the β-divergence when $\beta \to -1$ [26].

The priors $P(\boldsymbol{W})$ and $P(\boldsymbol{H})$ in (10) can be determined in many ways. Févotte et al. [28] proposed the determination of priors using Markov chains and the inverse Gamma distribution. In our approach, we propose to model the priors with the Gibbs distribution, which is particularly useful for enforcing local smoothness in images.

MRF Smoothing

Let us assume that prior information on the total smoothness of the estimated components $\boldsymbol{W}$ and $\boldsymbol{H}$ is modeled using the following Gibbs distributions:

$$\begin{aligned} P(\boldsymbol{W}) &= \frac{1}{Z_W} \exp\{-\alpha_W U(\boldsymbol{W})\}, \\ P(\boldsymbol{H}) &= \frac{1}{Z_H} \exp\{-\alpha_H U(\boldsymbol{H})\} \end{aligned} \tag{16}$$

where Z_W and Z_H are partition functions, α_W and α_H are regularization parameters, and $U(\boldsymbol{P})$ is a total energy function, which measures the total roughness in $\boldsymbol{P}$. The function $U(\boldsymbol{P})$ is often formulated with respect to the MRF model, which is commonly used in image reconstruction for modeling local smoothness.

The functions $U(\boldsymbol{W})$ and $U(\boldsymbol{H})$ can be determined for the matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ in the following way:

$$U(\boldsymbol{W}) = \sum_{f,r} \sum_{l \in S_f} v_{fl} \psi(w_{fr} - w_{lr}, \delta_W), \tag{17}$$

$$U(\boldsymbol{H}) = \sum_{t,r} \sum_{l \in S_t} v_{tl} \psi(h_{rt} - h_{rl}, \delta_H). \tag{18}$$

In the first-order interactions (nearest neighborhood), we have $S_f = \{f-1, f+1\}$ and the weighting factor $v_{fl} = 1$, and $S_t = \{t-1, t+1\}$ with $v_{tl} = 1$. In the second-order interactions, $S_f = \{f-2, f-1, f+1, f+2\}$ and $S_t = \{t-2, t-1, t+1, t+2\}$. The parameters δ_W and δ_H are scaling factors, while $\psi(\xi, \delta)$ is a potential function of ξ that can take different forms. The potential functions that can be applied to the EM-NMF algorithm are listed in Table 1.

According to Lange [41], a robust potential function in the Gibbs prior should have the following properties: nonnegative, even, 0 at $\xi = 0$, strictly increasing for $\xi > 0$,

Table 1 Potential functions

Author(s) (name)	Functions: $\psi(\xi, \delta)$	Reference
(Gaussian)	$(\xi/\delta)^2$	
Besag (Laplacian)	$\lvert\xi/\delta\rvert$	[36]
Bouman and Sauer (GGMRF)	$\lvert\xi/\delta\rvert^p$	[37]
Geman and McClure	$\frac{16}{3\sqrt{3}}\frac{(\xi/\delta)^2}{(1+(\xi/\delta)^2)}$	[38]
Geman and Reynolds	$\frac{\lvert\xi/\delta\rvert}{1+\lvert\xi/\delta\rvert}$	[39]
Green	$\delta\ln[\cosh(\xi/\delta)]$	[34]
Hebert and Leahy	$\delta\ln[1+(\xi/\delta)^2]$	[40]

unbounded, and convex with bounded first-derivative. Of the functions listed in Table 1, Green's function satisfies all of these properties, and consequently, it was selected for use in the tests in [33]. Unfortunately, the application of Green's function to both matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ demands the determination of two hyperparameters δ_W and δ_H, and two penalty parameters α_W and α_H. Moreover, data-driven hyperparameter estimation usually involves an approximation of the partition functions $\mathbf{Z}_W$ and $\mathbf{Z}_H$, which is not easy in this task.

The Gaussian function $\psi(\xi,\delta)=(\xi/\delta)^2$, as shown in Table 1, does not have a bounded first-derivative, but its scaling parameter δ may be merged with a penalty parameter α. Consequently, only two parameters need to be determined. The MRF model with a Gaussian potential function is actually the SAR model [42–44], which is used widely in many scientific fields [44, 45] to represent the interactions among spatial data with Gaussian noise. Let $\boldsymbol{w}_r \in \mathbb{R}_+^F$ be the r-th column of the matrix $\boldsymbol{W}$, and $\underline{\mathbf{h}}_r \in \mathbb{R}_+^{1\times T}$ be the r-th row of the matrix $\boldsymbol{H}$. Random variables in the vectors $\boldsymbol{w}_r$ and $\underline{\mathbf{h}}_r$ can be modeled using the following stochastic equations:

$$\boldsymbol{w}_r=\boldsymbol{S}^{(W)}\boldsymbol{w}_r+\epsilon, \quad \underline{\mathbf{h}}_r=\underline{\mathbf{h}}_r\boldsymbol{S}^{(H)}+\epsilon, \tag{19}$$

where $\boldsymbol{S}^{(W)}\in\mathbb{R}^{F\times F}$ and $\boldsymbol{S}^{(H)}\in\mathbb{R}^{T\times T}$ are symmetric matrices of spatial dependencies between the random variables, $\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ is an i.i.d. Gaussian noise, and $\mathbf{I}$ is an identity matrix with a corresponding size.

According to [45, 46], the spatial dependence matrices can be expressed as $\boldsymbol{S}^{(W)}=\gamma\mathbf{Z}^{(W)}$ and $\boldsymbol{S}^{(H)}=\gamma\mathbf{Z}^{(H)}$, where γ is a constant that ensures that the matrices $\boldsymbol{C}^{(W)}=\mathbf{I}_F-\boldsymbol{S}^{(W)}$ and $\boldsymbol{C}^{(H)}=\mathbf{I}_T-\boldsymbol{S}^{(H)}$ are positive-definite, while $\mathbf{Z}^{(W)}=[z_{mf}^{(W)}]$ and $\mathbf{Z}^{(H)}=[z_{tn}^{(H)}]$ are binary symmetric band matrices indicating the neighboring entries in $\boldsymbol{w}_r$ and $\underline{\mathbf{h}}_r$, respectively. In the first-order interactions, we have $z_{1,2}^{(W)}=z_{F,F\text{-}1}^{(W)}=z_{m,m\text{-}1}^{(W)}=z_{m,m+1}^{(W)}=1$ for $m\in\{2,\ldots,F-1\}$; $z_{2,1}^{(H)}=z_{T-1,T}^{(H)}=z_{n-1,n}^{(H)}=z_{n+1,n}^{(H)}=1$ for $n\in\{2,\ldots,T-1\}$, and $z_{mf}=z_{tn}=0$ otherwise. In the P-order interactions, each entry w_{fr} and h_{rt} has the corresponding sets of neighbors: $\{w_{f\text{-}v,r}\}$, $\{w_{f+v,r}\}$, $\{h_{r,t\text{-}v}\}$, $\{h_{r,t+v}\}$ with $v=1,\ldots,P$. As a consequence, $\boldsymbol{Z}^{(W)}$ and $\boldsymbol{Z}^{(H)}$ are symmetric band matrices with P sub-diagonals and P super-diagonals, the entries of which are equal to ones, but zeros otherwise. The matrices $\boldsymbol{C}^{(W)}$ and $\boldsymbol{C}^{(H)}$ are positive-definite, if $\gamma<(2P)^{-1}$ for P-order interactions [45, 46]. We selected $\gamma=(2P)^{-1}-\tilde{\epsilon}$, where $\tilde{\epsilon}$ is a small constant, for example, $\tilde{\epsilon}=10^{-16}$.

In the SAR model, Gibbs priors (16) may be expressed as their joint multivariate Gaussian priors:

$$P(\boldsymbol{W})=\prod_{r=1}^{R}Z_W^{-1}\exp\left\{-\frac{\alpha_W}{2}||\boldsymbol{C}^{(W)}\boldsymbol{w}_r||_2^2\right\}, \tag{20}$$

$$P(\boldsymbol{H})=\prod_{r=1}^{R}Z_H^{-1}\exp\left\{-\frac{\alpha_H}{2}||\underline{\mathbf{h}}_r\boldsymbol{C}^{(H)}||_2^2\right\} \tag{21}$$

where $Z_W=\left(\frac{2\pi}{\alpha_W}\right)^{F/2}\left(\prod_{f=1}^{F}\lambda_f^2(\boldsymbol{C}^{(W)})\right)^{-1/2}$, and $\lambda_f(\boldsymbol{C}^{(W)})$ is the f-th eigenvalue of the matrix $\boldsymbol{C}^{(W)}$. Similarly, $Z_H=\left(\frac{2\pi}{\alpha_H}\right)^{T/2}\left(\prod_{t=1}^{T}\lambda_t^2(\boldsymbol{C}^{(H)})\right)^{-1/2}$. If $P=1:\lambda_f(\boldsymbol{C}^{(W)})\cong 1-\cos\left(\frac{\pi f}{F}\right)$ and $\lambda_t(\boldsymbol{C}^{(H)})\cong 1-\cos\left(\frac{\pi t}{T}\right)$, which simplifies the hyperparameter estimation.

Algorithm

The EM algorithm [47] is applied to maximize $\ln P(\mathcal{X},\mathcal{C},\boldsymbol{W},\boldsymbol{H}|\mathcal{A},\boldsymbol{\Sigma}_n)$ in (11). To calculate the E-step, the log-likelihood functional (13) is transformed to the following form

$$\begin{aligned}\ln P(\mathcal{X}|\mathcal{C},\mathcal{A},\boldsymbol{\Sigma}_n)=&-T\sum_f\mathrm{tr}\left\{\boldsymbol{\Sigma}_n^{-1}\mathbf{R}_f^{(xx)}\right\}\\&+T\sum_f\mathrm{tr}\left\{\boldsymbol{A}_f^H\boldsymbol{\Sigma}_n^{-1}\mathbf{R}_f^{(xs)}\right\}\\&+T\sum_f\mathrm{tr}\left\{\boldsymbol{\Sigma}_n^{-1}\boldsymbol{A}_f(\mathbf{R}_f^{(xs)})^H\right\}\\&-T\sum_f\mathrm{tr}\left\{\boldsymbol{A}_f^H\boldsymbol{\Sigma}_n^{-1}\boldsymbol{A}_f\mathbf{R}_f^{(ss)}\right\}\\&-\sum_{f,t}\ln\det\boldsymbol{\Sigma}_n,\end{aligned} \tag{22}$$

where the correlation matrices are given by $\mathbf{R}_f^{(xx)}=\frac{1}{T}\sum_t\boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H$, $\mathbf{R}_f^{(ss)}=\frac{1}{T}\sum_t\boldsymbol{s}_{ft}\boldsymbol{s}_{ft}^H$, and the cross-correlation $\mathbf{R}_f^{(xs)}=\frac{1}{T}\sum_t\boldsymbol{x}_{ft}\boldsymbol{s}_{ft}^H$.

Ozerov et al. [23] observed that the set $\{\mathbf{R}_f^{(xx)}, \mathbf{R}_f^{(xs)}, \mathbf{R}_f^{(ss)}, |c_{rft}|^2\}$ provides sufficient statistics for the exponential family [47], so the sources $\boldsymbol{s}_{ft}$ and the latent components $\boldsymbol{c}_{ft}$ can be estimated by computing the conditional expectations of the natural statistics. According to [23], we have the following posterior estimates:

$$\hat{\boldsymbol{s}}_{ft} = \boldsymbol{\Sigma}_{ft}^{(s)}\boldsymbol{A}_f^H(\boldsymbol{A}_f\boldsymbol{\Sigma}_{ft}^{(s)}\boldsymbol{A}_f^H + \boldsymbol{\Sigma}_n)^{-1}\boldsymbol{x}_{ft}, \tag{23}$$

$$\hat{\boldsymbol{\Sigma}}_{ft}^{(s)} = \boldsymbol{\Sigma}_{ft}^{(s)} - \boldsymbol{\Sigma}_{ft}^{(s)}\boldsymbol{A}_f^H(\boldsymbol{A}_f\boldsymbol{\Sigma}_{ft}^{(s)}\boldsymbol{A}_f^H + \boldsymbol{\Sigma}_n)^{-1}\boldsymbol{A}_f\boldsymbol{\Sigma}_{ft}^{(s)}. \tag{24}$$

Similarly, for the latent components, we have

$$\hat{\boldsymbol{c}}_{ft} = \boldsymbol{\Sigma}_{ft}^{(c)}\bar{\boldsymbol{A}}_f^H(\bar{\boldsymbol{A}}_f\boldsymbol{\Sigma}_{ft}^{(c)}\bar{\boldsymbol{A}}_f^H + \boldsymbol{\Sigma}_n)^{-1}\boldsymbol{x}_{ft}, \tag{25}$$

$$\hat{\boldsymbol{\Sigma}}_{ft}^{(c)} = \boldsymbol{\Sigma}_{ft}^{(c)} - \boldsymbol{\Sigma}_{ft}^{(c)}\bar{\boldsymbol{A}}_f^H(\bar{\boldsymbol{A}}_f\boldsymbol{\Sigma}_{ft}^{(c)}\bar{\boldsymbol{A}}_f^H + \boldsymbol{\Sigma}_n)^{-1}\bar{\boldsymbol{A}}_f\boldsymbol{\Sigma}_{ft}^{(c)}. \tag{26}$$

The conditional expectations for the sufficient statistics are as follows:

$$\hat{\mathbf{R}}_f^{(xx)} = \mathbf{R}_f^{(xx)}, \quad \hat{\mathbf{R}}_f^{(xs)} = \frac{1}{T}\sum_t \boldsymbol{x}_{ft}\mathcal{E}(\boldsymbol{s}_{ft}^H) = \frac{1}{T}\sum_t \boldsymbol{x}_{ft}\hat{\boldsymbol{s}}_{ft}^H, \tag{27}$$

$$\hat{\mathbf{R}}_f^{(ss)} = \frac{1}{T}\sum_t \mathcal{E}(\boldsymbol{s}_{ft})\mathcal{E}(\boldsymbol{s}_{ft}^H) + \hat{\boldsymbol{\Sigma}}_{ft}^{(s)} = \frac{1}{T}\sum_t \hat{\boldsymbol{s}}_{ft}\hat{\boldsymbol{s}}_{ft}^H + \hat{\boldsymbol{\Sigma}}_{ft}^{(s)}, \tag{28}$$

$$|c_{rft}|^2 \leftarrow \mathcal{E}(c_{rft})\mathcal{E}(c_{rft}^H) + (\hat{\boldsymbol{\Sigma}}_{ft}^{(c)})_{rr} = |\hat{c}_{rft}|^2 + (\hat{\boldsymbol{\Sigma}}_{ft}^{(c)})_{rr}. \tag{29}$$

Detailed derivations of the formulae (23)–(26) are presented in the ‘‘Appendix’’.

From the M-step, we have $\frac{\partial}{\partial \boldsymbol{A}_f}\ln P(\mathcal{X}, \mathcal{C}, \boldsymbol{W}, \boldsymbol{H}|\mathcal{A}, \boldsymbol{\Sigma}_n) = 2T(-\boldsymbol{\Sigma}_n^{-1}\mathbf{R}_f^{(xs)} + \boldsymbol{\Sigma}_n^{-1}\boldsymbol{A}_f\mathbf{R}_f^{(ss)}) = 0$, which gives $\boldsymbol{A}_f = \hat{\mathbf{R}}_f^{(xs)}(\hat{\mathbf{R}}_f^{(ss)})^{-1}$. From

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_n^{-1}}\ln P(\mathcal{X}, \mathcal{C}, \boldsymbol{W}, \boldsymbol{H}|\mathcal{A}, \boldsymbol{\Sigma}_n) = 0,$$

we have

$$\boldsymbol{\Sigma}_n = \mathrm{diag}\left\{\mathbf{R}_f^{(xx)} - \boldsymbol{A}_f(\hat{\mathbf{R}}_f^{(xs)})^H - \hat{\mathbf{R}}_f^{(xs)}\boldsymbol{A}_f^H + \boldsymbol{A}_f\hat{\mathbf{R}}_f^{(ss)}\boldsymbol{A}_f^H\right\}.$$

From $\frac{\partial}{\partial w_{fr}}\ln P(\mathcal{X}, \mathcal{C}, \boldsymbol{W}, \boldsymbol{H}|\mathcal{A}, \boldsymbol{\Sigma}_n) = 0$, we have

$$w_{fr} = \frac{1}{T}\sum_{t=1}^{T}\frac{|c_{rft}|^2}{h_{rt}} - \alpha_W\nabla_{w_{fr}}U(\boldsymbol{W}). \tag{30}$$

Similarly, from $\frac{\partial}{\partial h_{rt}}\ln P(\mathcal{X}, \mathcal{C}, \boldsymbol{W}, \boldsymbol{H}|\mathcal{A}, \boldsymbol{\Sigma}_n) = 0$, we get

$$h_{rt} = \frac{1}{F}\sum_{f=1}^{F}\frac{|c_{rft}|^2}{w_{fr}} - \alpha_H\nabla_{h_{rt}}U(\boldsymbol{W}). \tag{31}$$

The terms $\nabla_{w_{fr}}U(\boldsymbol{W})$ and $\nabla_{h_{rt}}U(\boldsymbol{W})$ in (30) and (31) take the following forms with respect to the potential functions:

- *Gaussian* (SAR model):

$$\nabla_{w_{fr}}U(\boldsymbol{W}) = \left[(\boldsymbol{C}^{(W)})^T\boldsymbol{C}^{(W)}\boldsymbol{W}\right]_{fr}, \tag{32}$$

$$\nabla_{h_{rt}}U(\boldsymbol{H}) = \left[\boldsymbol{H}\boldsymbol{C}^{(H)}(\boldsymbol{C}^{(H)})^T\right]_{rt}, \tag{33}$$

- *GR function* (proposed by Green [34]):

$$\nabla_{w_{fr}}U(\boldsymbol{W}) = \sum_{l\in S_f} v_{fl}\tanh\left(\frac{w_{fr} - w_{lr}}{\delta_W}\right), \tag{34}$$

$$\nabla_{h_{rt}}U(\boldsymbol{H}) = \sum_{l\in S_t} v_{tl}\tanh\left(\frac{h_{rt} - h_{rl}}{\delta_H}\right). \tag{35}$$

- *HL function* (proposed by Hebert and Leahy [40]):

$$\nabla_{w_{fr}}U(\boldsymbol{W}) = \sum_{l\in S_f} v_{fl}\frac{2\delta_W(w_{fr} - w_{lr})}{\delta_W^2 + (w_{fr} - w_{lr})^2}, \tag{36}$$

$$\nabla_{h_{rt}}U(\boldsymbol{H}) = \sum_{l\in S_t} v_{tl}\frac{2\delta_H(h_{rt} - h_{rl})}{\delta_H^2 + (h_{rt} - h_{rl})^2}. \tag{37}$$

Experiments

Experiments were conducted using selected sound recordings taken from the stereo audio source separation evaluation campaign (SiSEC)[1] in 2007. This campaign aimed to evaluate the performance of source separation algorithms using stereo under-determined mixtures. We selected the benchmarks given in Table 2, which included speech recordings (three male voices—`male3`, and three female voices—`female3`), three nonpercussive music sources—`nodrums`, and three music sources that included drums—`wdrums`. The mixed signals were recordings that lasted 10 s, which were sampled at 16 kHz (the standard settings of recordings from the “Under-determined speech and music mixtures“ datasets in the SiSEC2008). For each benchmark, the number of true sources was three ($J = 3$) but it only had two microphones ($I = 2$), that is, stereo recordings. Thus, for each case, we faced an under-determined BSS problem. All instantaneous mixtures were obtained using the same mixing matrix with positive coefficients. Synthetic convolutive mixtures were obtained for a meeting room with a 250 ms reverberation time using omnidirectional microphones with 1 m spacing.

The spectrograms were obtained by a short-time fourier transform (STFT) using half-overlapping sine windows. To create the spectrograms and recover the time-domain signals from STFT coefficients, we used the corresponding `stft_multi` and `istft_multi` Matlab functions from the SiSEC2008 webpage[2] [48]. For instantaneous and convolutive mixtures, the window lengths were set to 1,024 and 2,048 samples, respectively.

[1] http://sisec.wiki.irisa.fr.

[2] http://sisec2008.wiki.irisa.fr.

Table 2 Benchmarks

Instantaneous	Convolutive
male3_inst_mix	*male3_synthconv_250ms_1m_mix*
female3_inst_mix	*female3_synthconv_250ms_1m_mix*
nodrums_inst_mix	*nodrums_synthconv_250ms_1m_mix*
wdrums_inst_mix	*wdrums_synthconv_250ms_1m_mix*

The EM-NMF algorithm was taken from Ozerov's homepage[3], while the MRF-EM-NMF algorithm was coded and extensively tested by Ochal [49].

The proposed algorithm is based on an alternating optimization scheme, which is intrinsically non-convex, and hence, its initialization plays an important role. An incorrect initialization may result in slow convergence and early stagnation at an unfavorable local minimum of the objective function. As done in many NMF algorithms, the factors $\boldsymbol{W}$ and $\boldsymbol{H}$ are initialized with uniformly distributed random numbers, whereas the entries in the matrix $\boldsymbol{A}$ are drawn from a zero-mean complex Gaussian distribution. After $\boldsymbol{W}$ and $\boldsymbol{H}$ have been initialized, the covariance matrices $\boldsymbol{\Sigma}_{ft}^{(s)}$ and $\boldsymbol{\Sigma}_{ft}^{(c)}$ given by (8) can be computed. A noise covariance matrix $\boldsymbol{\Sigma}_n$ is needed to update the E-step. Ozerov and Fevotte [23] tested several techniques for determining this matrix. The E-step in MRF-EM-NMF is identical to that in EM-NMF [23], and hence, all of these techniques can be used in this experiment. The initial matrix $\boldsymbol{\Sigma}_n$ was determined based on the empirical variance of the observed power spectrograms.

The MRF-EM-NMF and EM-NMF algorithms were initialized using the same random values (given as $\bar{R}$) and run for 1,500 iterations.

The choice of the parameters $\{\alpha_W, \alpha_H, \gamma_W, \gamma_H\}$ used in the Gibbs distributions also affected the performance. The regularization parameters can be fixed or changed with iterations. Motivated by iterative thresholding strategies [26], we used the following strategies:

- Linear thresholding:

$$\alpha(k) = \alpha \frac{k}{k_{\max}},$$

- Nonlinear thresholding:

$$\alpha(k) = \frac{\alpha}{2}\left(1 + \tanh\left(\frac{k - \nu k_{\max}}{\tau k_{\max}}\right)\right),$$

- Fixed thresholding:

$$\alpha(k) = \begin{cases} \alpha & \text{if} \quad k > k_1, \\ 0 & \text{otherwise} \end{cases}$$

where k is the current iteration, $k_{\max}$ is the maximum number of iterations, $\tau \in (0, 1)$ is the shape parameter, $\nu \in (0, 1)$ is the shift parameter, k_1 is the threshold, and α can be equal to α_W or α_H. All of the above thresholding strategies aim to relax smoothing during the early iterations when the descent directions in the updates are sufficiently steep and to emphasize smoothing if noisy perturbations become significantly detrimental to the overall smoothness. These strategies are motivated by standard regularization rules that apply to ill-posed problems. We tested all of the thresholding strategies using instantaneous and convolutive mixtures, and we obtained the best performance with fixed thresholding using $k_1 = k_{\max}/2$.

The parameters δ_W and δ_H in the MRF models can be estimated using standard marginalization procedures or by maximizing the Type II ML estimate for (10). However, these techniques have a huge computational cost for the nonlinear potential functions in the MRF models. For practical reasons, they are not very useful for the GR or HR functions.

In this study, we tested all of the benchmarks in Table 2 and the following potential functions: the first- and second-order Gaussian, GR, and HR. For the Gaussian functions, we tested all combinations of the regularization parameters α_W and α_H from the discrete set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. For GR and HL, the regularization parameters could take only two values, $\{0.001, 0.01\}$, although the parameters δ_W and δ_H were tested with the following values: $\{0.1, 1, 10\}$. The optimal values of the smoothing parameters are summarized in Table 3.

The separation results were evaluated in terms of the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [50]. Figure 1 shows the SDRs and SIRs averaged for the sources, which were estimated using the EM-NMF and MRF-EM-NMF with various smoothing functions based on instantaneous and convolutive mixing models. For each sample in Table 2 and each smoothing function, the smoothing parameters were tuned optimally for a given fixed initializer. This unsupervised learning approach evaluated the efficiency of the smoothing functions with respect to a given recording scenario. However, the smoothing parameters need to be determined with a supervised learning framework in practice. To test this option, each recording in Table 2 was divided into two 5 s excerpts during the training and testing stages. For each training excerpt, the smoothing parameters and initializer were selected to maximize the SDR performance. Testing was performed on the other excerpt with the same initializer. The results obtained during the testing stage with the instantaneous mixtures are shown in Fig. 2.

For comparison, Table 4 shows the average SDR results produced and the running time taken when using several

[3] http://www.irisa.fr/metiss/ozerov/.

Table 3 Parameters of the MRF-EM-NMF algorithm for each test case shown in Fig. 1

Benchmark	Smoothing	Instantaneous mixture					Convolutive mixture				
		$\bar{R}$	α_W	α_H	δ_W	δ_H	$\bar{R}$	α_W	α_H	δ_W	δ_H
Male	GR	12	0.01	0.01	1	1	4	0.01	0.01	0.1	10
Male	HL	12	0.001	0.001	1	10	4	0.001	0.01	1	1
Male	1-Gaussian	12	0.001	0.01	–	–	4	0.05	0.05	–	–
Male	2-Gaussian	12	0.001	0.01	–	–	4	0.05	0.01	–	–
Female	GR	12	0.01	0.01	10	10	4	0.01	0.01	1	1
Female	HL	12	0.001	0.001	1	10	4	0.001	0.001	0.1	10
Female	1-Gaussian	12	0.001	0.001	–	–	4	0.1	0.001	–	–
Female	2-Gaussian	12	0.001	0.001	–	–	4	0.05	0.005	–	–
Nodrums	GR	4	0.01	0.01	10	1	4	0.01	0.01	10	0.1
Nodrums	HL	4	0.01	0.001	1	10	4	0.01	0.01	0.1	0.1
Nodrums	1-Gaussian	4	0.001	0.01	–	–	4	0.001	0.05	–	–
Nodrums	2-Gaussian	4	0.01	0.001	–	–	4	0.005	0.01	–	–
Wdrums	GR	4	0.01	0.01	1	10	4	0.01	0.01	1	1
Wdrums	HL	4	0.01	0.001	1	10	4	0.001	0.01	1	0.1
Wdrums	1-Gaussian	4	0.001	0.001	–	–	4	0.001	0.1	–	–
Wdrums	2-Gaussian	4	0.001	0.001	–	–	4	0.005	0.1	–	–

The notations "1-Gaussian" and "2-Gaussian" represent the first- and second-order Gaussian functions, respectively

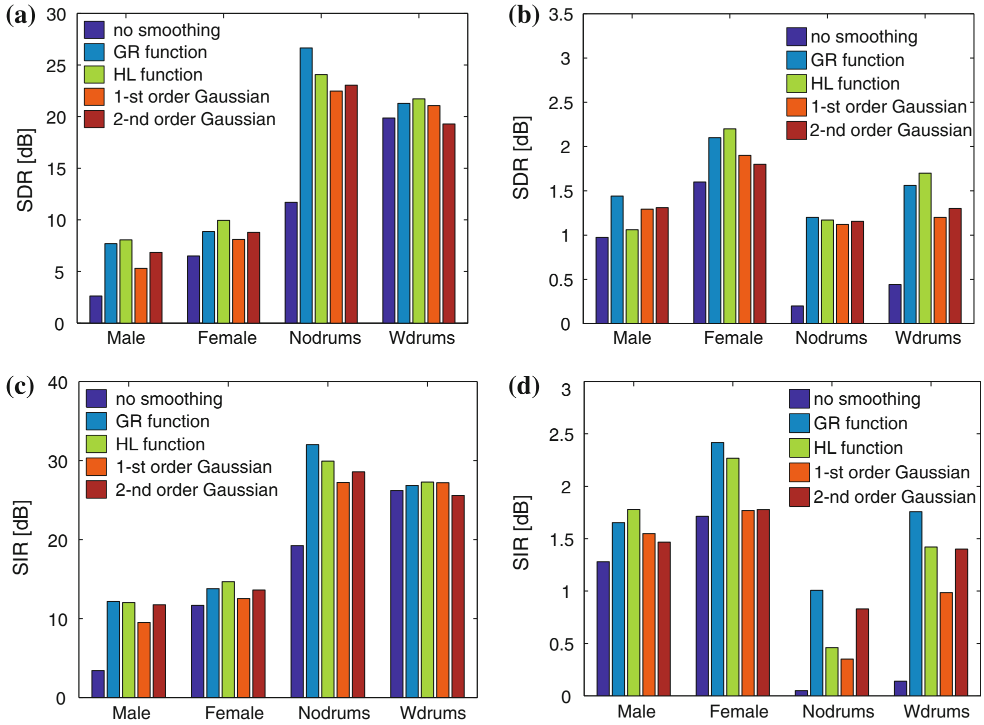

Fig. 1 Source separation results obtained with the MRF-EM-NMF (first- and second-order Gaussian, GR, and HL functions) and EM-NMF (no smoothing) algorithms after 1,500 iterations: **a** mean SDR (dB) for instantaneous mixture, **b** mean SDR (dB) for convolutive mixture, **c** mean SIR (dB) for instantaneous mixture, **d** mean SIR (dB) for convolutive mixture. The smoothing parameters were tuned separately for each mixture in Table 2

state-of-the-art algorithms, which were applied to the mixtures in Table 2. The generalized Gaussian prior (GGP) algorithm [51] and the statistically sparse decomposition principle (SSDP) algorithms [52] were applied to the instantaneous mixtures. The convolutive mixtures were unmixed with the IPD [21], two versions of the FBWC-PA

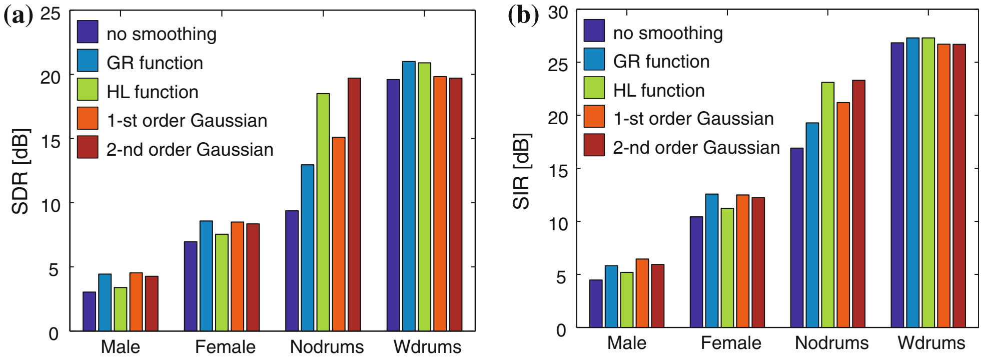

Fig. 2 Source separation results obtained in the testing stage with the MRF-EM-NMF (first- and second-order Gaussian, GR, and HL functions) and EM-NMF (no smoothing) algorithm after 1,500 iterations: **a** mean SDR (dB), **b** mean SIR (dB). The smoothing parameters were determined during the training stage. 5 s excerpts were used in the training and testing stages

Table 4 Mean SDR (dB) and running time (s) for sources estimated from the mixtures shown in Table 2

Benchmark	Mixture	Male	Female	Nodrums	Wdrums	Time
MRF-EM-NMF (HR)	inst	8.06	9.95	24.07	21.72	2487
MRF-EM-NMF (GR) [33]	inst	7.69	8.86	26.65	21.28	2498
EM-NMF [23]	inst	2.62	6.5	11.7	19.87	2456
GGP [51]	inst	8.4	8.57	13.9	10.3	5
SABM+SSDP [52]	inst	4.25	3.82	5.83	9.43	2
MRF-EM-NMF (HR)	conv	1.06	2.2	1.17	1.7	2760
MRF-EM-NMF (GR) [33]	conv	1.4	2.1	1.2	1.56	2762
EM-NMF [23]	conv	0.95	1.6	0.2	0.44	2720
IPD [21]	conv	1.53	1.43	2.2	−2.7	1200
FBWC-PA [17]	conv	−0.1	4.43	0.77	−2.53	40
Generalized FBWC-PA [18]	conv	5.95	7.45	1.2	−0.69	8
ConvNMF [22]	conv	−0.7	−0.47	3.85	8.13	347

[17, 18] algorithm, and the Convolutive NMF [22]. Note that the last method in this list was based on supervised learning, whereas the others were unsupervised learning algorithms. In this case, the first 8 s excerpts of the 10 s source recordings were used for learning, while the remainder was used for testing.

The averaged elapsed time measured using Matlab 2008a for 1,500 iterations with $\bar{R} = 12$, executed on a 64-bit Intel Quad Core CPU 3 GHz with 8 GB RAM was almost the same for the MRF-EM-NMF and EM-NMF algorithms (see Table 4).

The simulations demonstrate that MRF smoothing improved the source separation results in almost all test cases. The results confirmed that instantaneous mixtures were considerably easier to separate than convolutive ones. The MRF-EM-NMF algorithm delivered the best mean SDR performance of all the algorithms tested with instantaneous mixtures. The highest SDR values were produced with instantaneously mixed non-percussive music sources. This was justified by the smooth frequency and temporal structures of non-percussive music spectrograms. If the source spectrograms were not very smooth (as with the percussive audio recordings), MRF smoothing gave only a slight improvement (see Figs. 1, 2) in the first-order MRF interactions, and even a slight deterioration in the higher-order MRF interactions. According to Fig. 1, the HL function delivered the most promising SDR results, which were stable with a wide range of parameters. In each case with the instantaneous mixtures, the best results were produced with the same hyperparameter values, $\delta_W = 1$ and $\delta_H = 10$, and almost the same penalty parameter values, α_W and α_H. The SAR model also improved the results compared with the standard EM-NMF algorithm. Moreover, the SAR model was tuned using only two penalty parameters, and the partition function of the associated Gibbs prior could be derived using a closed-form expression, which might be very useful for data-driven hyperparameter estimation.

The source separation results produced with the MRF-EM-NMF algorithm for convolutive and under-determined mixtures were better than those obtained with the EM-NMF algorithm. Unfortunately, the SDR values showed that these results were still a long way from being perfect, even after 1,500 iterations, and thus, further research is needed in this field. It is likely that some additional prior information could be imposed, especially on a mixing operator, which might increase the efficiency considerably.

It should be noted that the SDR performance with both mixtures could still be improved by refining the associated parameters, especially in the MRF models, and by using more efficient initializers.

Conclusions

This study demonstrated that imposing MRF smoothing on the power spectrograms of audio sources estimated from under-determined unmixing problems may improve the quality of estimated audio sounds considerably. This was justified because any type of meaningful prior information improves the performance, especially with under-determined problems. This study addressed the application of MRF smoothing in the EM-NMF algorithm, but this type of smoothing could be applied to many other related BSS algorithms based on feature extraction from power spectrograms. Thus, the theoretical results presented in this paper may have broad practical applications. Clearly, further studies are needed to improve this technique for convolutive mixtures and to integrate regularization parameter estimation techniques in the main algorithm.

Acknowledgments This work was supported by habilitation grant N N515 603139 (2010–2012) from the Ministry of Science and Higher Education, Poland. The author would like to thank the reviewers for their valuable comments.

Appendix

The conditional expectations of the natural statistics can be derived from the a posteriori distributions $P(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft})$ and $P(\boldsymbol{c}_{ft}|\boldsymbol{x}_{ft})$. Thus,

$$\begin{aligned} P(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft}) &= \frac{P(\boldsymbol{x}_{ft}, \boldsymbol{s}_{ft})}{P(\boldsymbol{x}_{ft})} \\ &= \frac{(\pi^{I+J} \det \boldsymbol{\Sigma}_{ft})^{-1} \exp\left\{ -\begin{bmatrix} \bar{\boldsymbol{x}}_{ft} \\ \bar{\boldsymbol{s}}_{ft} \end{bmatrix}^H (\boldsymbol{\Sigma}_{ft})^{-1} \begin{bmatrix} \bar{\boldsymbol{x}}_{ft} \\ \bar{\boldsymbol{s}}_{ft} \end{bmatrix} \right\}}{(\pi^I \det \boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \exp\left\{ -(\bar{\boldsymbol{x}}_{ft})^H (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \bar{\boldsymbol{x}}_{ft} \right\}} \\ &= (\pi^J \det \Gamma_{ft})^{-1} \exp\{-\Psi_{ft}\}, \end{aligned} \tag{38}$$

where $\bar{\boldsymbol{x}}_{ft} = \boldsymbol{x}_{ft} - \mathcal{E}(\boldsymbol{x}_{ft}), \bar{\boldsymbol{s}}_{ft} = \boldsymbol{s}_{ft} - \mathcal{E}(\boldsymbol{s}_{ft}), \Gamma_{ft} = \boldsymbol{\Sigma}_{ft}^{(s)} - \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \boldsymbol{\Sigma}_{ft}^{(xs)}$,

$$\boldsymbol{\Sigma}_{ft} = \begin{bmatrix} \boldsymbol{\Sigma}_{ft}^{(x)} & \boldsymbol{\Sigma}_{ft}^{(xs)} \\ \boldsymbol{\Sigma}_{ft}^{(sx)} & \boldsymbol{\Sigma}_{ft}^{(s)} \end{bmatrix},$$

$$\Psi_{ft} = \begin{bmatrix} \bar{\boldsymbol{x}}_{ft} \\ \bar{\boldsymbol{s}}_{ft} \end{bmatrix}^H (\boldsymbol{\Sigma}_{ft})^{-1} \begin{bmatrix} \bar{\boldsymbol{x}}_{ft} \\ \bar{\boldsymbol{s}}_{ft} \end{bmatrix} - (\bar{\boldsymbol{x}}_{ft})^H (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \bar{\boldsymbol{x}}_{ft}.$$

We can transform $\boldsymbol{\Sigma}_{ft}^{-1}$ in (38) into the following form:

$$\boldsymbol{\Sigma}_{ft}^{-1} = \begin{bmatrix} \left(\boldsymbol{\Sigma}_{ft}^{(x)} - \boldsymbol{\Sigma}_{ft}^{(xs)} (\boldsymbol{\Sigma}_{ft}^{(s)})^{-1} \boldsymbol{\Sigma}_{ft}^{(sx)} \right)^{-1} & -(\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \boldsymbol{\Sigma}_{ft}^{(xs)} \Gamma_{ft}^{-1} \\ -\Gamma_{ft}^{-1} \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} & \Gamma_{ft}^{-1} \end{bmatrix}.$$

Using the Woodbury matrix identity, we have

$$\begin{aligned} &(\boldsymbol{\Sigma}_{ft}^{(x)} - \boldsymbol{\Sigma}_{ft}^{(xs)} (\boldsymbol{\Sigma}_{ft}^{(s)})^{-1} \boldsymbol{\Sigma}_{ft}^{(sx)})^{-1} \\ &\quad = (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} + (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \boldsymbol{\Sigma}_{ft}^{(xs)} \Gamma_{ft}^{-1} \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1}, \end{aligned}$$

and finally,

$$\begin{aligned} \Psi_{ft} &= \left(\bar{\boldsymbol{s}}_{ft} - \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \bar{\boldsymbol{x}}_{ft} \right)^H \Gamma_{ft}^{-1} \left(\bar{\boldsymbol{s}}_{ft} - \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \bar{\boldsymbol{x}}_{ft} \right) \\ &= \left(\boldsymbol{s}_{ft} - \hat{\boldsymbol{s}}_{ft} \right)^H (\hat{\boldsymbol{\Sigma}}_{ft}^{(s)})^{-1} \left(\boldsymbol{s}_{ft} - \hat{\boldsymbol{s}}_{ft} \right), \end{aligned} \tag{39}$$

where

$$\hat{\boldsymbol{s}}_{ft} = \mathcal{E}(\boldsymbol{s}_{ft}) + \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \left(\boldsymbol{x}_{ft} - \mathcal{E}(\boldsymbol{x}_{ft}) \right), \tag{40}$$

$$\hat{\boldsymbol{\Sigma}}_{ft}^{(s)} = \Gamma_{ft} = \boldsymbol{\Sigma}_{ft}^{(s)} - \boldsymbol{\Sigma}_{ft}^{(sx)} (\boldsymbol{\Sigma}_{ft}^{(x)})^{-1} \boldsymbol{\Sigma}_{ft}^{(xs)}. \tag{41}$$

Thus, $P(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft}) = \mathcal{N}_c(\boldsymbol{s}_{ft}; \hat{\boldsymbol{s}}_{ft}, \hat{\boldsymbol{\Sigma}}_{ft}^{(s)})$. From (5), it follows that $\mathcal{E}(\boldsymbol{s}_{ft}) = 0$, so $\mathcal{E}(\boldsymbol{x}_{ft}) = 0$. Since the zero-mean noise $\mathbf{n}_{ft}$ from (3) is not correlated with $\boldsymbol{s}_{ft}$, we have

$$\begin{aligned} \boldsymbol{\Sigma}_{ft}^{(xs)} &= \mathcal{E}\left((\boldsymbol{x}_{ft} - \mathcal{E}(\boldsymbol{x}_{ft}))(\boldsymbol{s}_{ft} - \mathcal{E}(\boldsymbol{s}_{ft}))^H \right) \\ &= \mathcal{E}\left((\boldsymbol{A}_f \boldsymbol{s}_{ft} + \mathbf{n}_{ft}) \boldsymbol{s}_{ft}^H \right) = \boldsymbol{A}_f \mathcal{E}(\boldsymbol{s}_{ft} \boldsymbol{s}_{ft}^H) + \mathcal{E}(\mathbf{n}_{ft} \boldsymbol{s}_{ft}^H) = \boldsymbol{A}_f \boldsymbol{\Sigma}_{ft}^{(s)}. \end{aligned} \tag{42}$$

Inserting (42) and $\boldsymbol{\Sigma}_{ft}^{(sx)} = (\boldsymbol{\Sigma}_{ft}^{(xs)})^H = \boldsymbol{\Sigma}_{ft}^{(s)} \boldsymbol{A}_f^H$ into (40) and (41), we obtain the update rules (23) and (24), respectively.

Analyzing $P(\boldsymbol{c}_{ft}|\boldsymbol{x}_{ft})$, one can obtain $P(\boldsymbol{c}_{ft}|\boldsymbol{x}_{ft}) = \mathcal{N}_c(\boldsymbol{c}_{ft}; \hat{\boldsymbol{c}}_{ft}, \hat{\boldsymbol{\Sigma}}_{ft}^{(c)})$, which yields the update rules (25) and (26).

References

1. Cichocki A, Amari SI. Adaptive blind signal and image processing (new revised and improved edition). New York: Wiley; 2003.
2. Hyvrinen A, Karhunen J, Oja E. Independent component analysis. New York: Wiley; 2001.
3. Comon P, Jutten C. Handbook of blind source separation: independent component analysis and applications. 1st ed. Burlington, MA: Academic Press, Elsevier; 2010, ISBN: 0123747260, 9780 123747266.
4. Naik GR, Kumar DK. Dimensional reduction using blind source separation for identifying sources. Int J Innov Comput Inf Control (IJICIC). 2011;7(2):989–1000.
5. Popescu TD. A new approach for dam monitoring and surveillance using blind source separation. Int J Innov Comput Inf Control (IJICIC). 2011;7(6):3811–3824.
6. Zhang Z, Miyake T, Imamura T, Enomoto T, Toda H. Blind source separation by combining independent component analysis with the complex discrete wavelet transform. Int J Innov Comput Inf Control (IJICIC). 2010;6(9):4157–4172.
7. Khosravy M, Asharif MR, Yamashita K: A PDF-matched short-term linear predictability approach to blind source separation. Int J Innov Comput Inf Control (IJICIC). 2009;5(11(A)):3677–3690.
8. Yang Z, Zhou G, Ding S, Xie S. Nonnegative blind source separation by iterative volume maximization with fully nonnegativity constraints. ICIC Express Lett. 2010;4(6(B)):2329–2334.
9. Pao TL, Liao WY, Chen YT, Wu TN. Mandarin audio-visual speech recognition with effects to the noise and emotion. Int J Innov Comput Inf Control (IJICIC). 2010;6(2):711–724.
10. Lin SD, Huang CC, Lin JH. A hybrid audio watermarking technique in cepstrum domain. ICIC Express Lett. 2010;4(5(A)):1597–1602.
11. Zin TT, Hama H, Tin P, Toriu T. HOG embedded markov chain model for pedestrian detection. ICIC Express Lett. 2010;4(6(B)):2463–2468.
12. Virtanen T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans Audio Speech Lang Process. 2007;15(3):1066–1074.
13. Pedersen MS, Larsen J, Kjems U, Parra LC. Convolutive blind source separation methods. In: Benesty J, Huang Y, Sondhi M, editors. Springer handbook of speech processing. Berlin: Springer; 2008. p. 1065−94, ISBN: 978-3-540-49125-5.
14. Parra L, Spence C. Convolutive blind separation of non-stationary sources. IEEE Trans Speech Audio Process. 2000;8(3) 320–327.
15. Yilmaz O, Rickard S. Blind separation of speech mixtures via time-frequency masking. IEEE Trans Signal Process. 2004;52(7):1830–1847.
16. Reju VG, Koh SN, Soon IY. Underdetermined convolutive blind source separation via time-frequency masking. IEEE Trans Audio Speech Lang Process. 2010;18(1):101–116.
17. Sawada H, Araki S, Makino S. Measuring dependence of binwise separated signals for permutation alignment in frequency-domain bss. In: ISCAS; 2007. p. 3247–3250.
18. Sawada H, Araki S, Makino S. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans Audio Speech Lang Process. 2011;19(3):516–527.
19. Aïssa-El-Bey A, Abed-Meraim K, Grenier Y. Blind separation of underdetermined convolutive mixtures using their time-frequency representation. IEEE Trans Audio Speech Lang Process. 2007;15(5):1540–1550.
20. Weiss RJ, Ellis DPW. Speech separation using speaker-adapted eigenvoice speech models. Comput Speech Lang. 2010; 24(1): 16–29.
21. Mandel MI, Ellis DPW, Jebara T. An EM algorithm for localizing multiple sound sources in reverberant environments. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in neural information processing systems 19. Cambridge: MIT Press; p. 953–960.
22. Smaragdis P. Convolutive speech bases and their application to supervised speech separation. IEEE Trans Audio Speech Lang Process. 2007;15(1):1–12.
23. Ozerov A, Févotte C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans Audio Speech Lang Process. 2010;18(3):550–563.
24. Ozerov A, Févotte C, Blouet R, Durrieu JL (2011) Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In: ICASSP; p. 257–260.
25. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–791.
26. Cichocki A, Zdunek R, Phan AH, Amari SI. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. Chichester, UK: Wiley and Sons; 2009.
27. Benaroya L, Gribonval R, Bimbot F. Non-negative sparse representation for Wiener based source separation with a single sensor. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP'03), Hong Kong; 2003. p. 613–616.
28. Févotte C, Bertin N, Durrieu JL. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Computation. 2009;21(3):793–830.
29. Duong NQK, Vincent E, Gribonval R. Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans Audio Speech Lang Process. 2010; 18(7);1830–1840.
30. Zdunek R, Cichocki A. Blind image separation using nonnegative matrix factorization with Gibbs smoothing. In: Ishikawa M, Doya K, Miyamoto H, Yamakawa T editors. Neural information processing, vol 4985 of Lecture notes in computer science. Berlin: Springer; 2008. p. 519–528 ICONIP 2007.
31. Zdunek R, Cichocki A. Improved M-FOCUSS algorithm with overlapping blocks for locally smooth sparse signals. IEEE Trans Signal Process. 2008;56(10):4752–4761.
32. Ozerov A, Vincent E, Bimbot F. A general flexible framework for the handling of prior information in audio source separation. IEEE Trans Audio Speech Lang Process. 2012;20(4):1118–1133.
33. Zdunek R. Convolutive nonnegative matrix factorization with Markov random field smoothing for blind unmixing of multichannel speech recordings. In: Travieso-Gonzalez CM, Alonso-Hernandez JB, editors. Advances in nonlinear speech processing, vol 7015 of Lecture notes in artificial intelligence (LNAI). Springer Berlin/Heidelberg; 2011. p. 25–32 NOLISP 2011.
34. Green PJ. Bayesian reconstruction from emission tomography data using a modified EM algorithm. IEEE Trans Med Imaging. 1990;9:84–93.
35. Itakura F, Saito S. An analysis-synthesis telephony based on the maximum likelihood method, vol c-5-5. In: Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan. New York: Elsevier; 1968. p. 17–20.
36. Besag J. Toward Bayesian image analysis. J Appl Stat. 1989;16: 395–407.

37. Bouman CA, Sauer K. A generalized Gaussian image model for edge-preserving MAP estimation. IEEE Trans Image Process. 1993;2:296–310.
38. Geman S, McClure D (1987) Statistical methods for tomographic image reconstruction. Bull Int Stat Inst. 1987;LII-4: 5–21.
39. Geman S, Reynolds G. Constrained parameters and the recovery of discontinuities. IEEE Trans Pattern Anal Mach Intell. 1992; 14:367–383.
40. Hebert T, Leahy R. A generalized EM algorithm for 3-D Bayesian reconstruction from poisson data using Gibbs priors. IEEE Trans Med Imaging. 1989;8:194–202.
41. Lange K. Convergence of EM image reconstruction algorithms with Gibbs smoothing. IEEE Trans Med Imaging. 1990;9(4): 439–446.
42. Whittle P. On stationary processes in the plane. Biometrika. 1954;41(3):434–449.
43. Besag J. Spatial interactions and the statistical analysis of lattice systems. J R Stat Soc Ser B. 1974;36:192–236.
44. Ripley BD. Spatial statistics. New York: Wiley; 1981.
45. Molina R, Katsaggelos A, Mateos J. Bayesian and regularization methods for hyperparameter estimation in image restoration. IEEE Trans Image Process. 1999;8(2):231–246.
46. Galatsanos N, Mesarovic V, Molina R, Katsaggelos A. Hierarchical Bayesian image restoration for partially-known blurs. IEEE Trans Image Process. 2000;9(10):1784–1797.
47. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc. 1977;39(1): 1–38.
48. Vincent E, Araki S, Theis FJ, Nolte G, Bofill P, Sawada H, Ozerov A, Gowreesunker BV, Lutter D, Duong QKN. The signal separation evaluation campaign (2007–2010): achievements and remaining challenges. Signal Process. 2012;92:1928–1936.
49. Ochal P. Application of convolutive nonnegative matrix factorization for separation of muscial instrument sounds from multichannel polyphonic recordings. M.Sc. thesis (supervised by Dr. R. Zdunek), Wroclaw University of Technology, Poland (2010) (in Polish).
50. Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation. IEEE Trans Audio Speech Lang Process 2006;14(4):1462–1469.
51. Vincent E. Complex nonconvex lp norm minimization for underdetermined source separation. In: Proceedings of the 7th international conference on Independent component analysis and signal separation. ICA'07. Berlin: Springer; 2007. p. 430–437.
52. Xiao M, Xie S, Fu Y. A statistically sparse decomposition principle for underdetermined blind source separation. In: Proceedings of 2005 international symposium on intelligent signal processing and communication systems (ISPACS 2005); 2005. p. 165–168.