MAXIMUM A POSTERIORI ESTIMATOR FOR CONVOLUTIVE SOUND SOURCE SEPARATION WITH SUB-SOURCE BASED NTF MODEL AND THE LOCALIZATION PROBABILISTIC PRIOR ON THE MIXING MATRIX

Mieszko Fraś and Konrad Kowalczyk

AGH University of Science and Technology, Department of Electronics, Kraków, Poland
fras@agh.edu.pl, konrad.kowalczyk@agh.edu.pl

ABSTRACT

In this paper we present a method for the separation of sound source signals recorded using multiple microphones in a reverberant room. In particular, we propose a maximum a posteriori (MAP) estimator based on the multichannel nonnegative tensor factorization (NTF) model with the localization prior distribution on the mixing matrix, in which the latent data consists of the so-called sub-sources for an improved performance in a reverberant environment. For the proposed MAP estimator, we derive the sub-source based expectation maximization (EM) algorithm with the multiplicative update rules (MU) and the localization prior distribution (LP) on the mixing matrix (SSEM-MU-LP). We then perform several experiments for speech and instrumental sound sources recorded using two microphones, in determined and under-determined scenarios, and with different types of initialization of the model parameters. The results of these experiments clearly indicate a significant improvement of the proposed algorithm with the localization prior over the state-of-the-art NTF-based source separation algorithms, which can reach up to 50% in the signal-to-distortion ratio.

Index Terms— Sound source separation, nonnegative tensor factorization, probabilistic localization prior, expectation-maximization

1. INTRODUCTION

Blind sound source separation (BSS) from convolutive mixtures remains a challenging task, especially in under-determined scenarios where the number of sources is larger than the number of microphones. Over the years, various approaches have been presented in the literature including independent component analysis (ICA) [1], nonnegative matrix factorization (NMF) [2], nonnegative tensor factorization (NTF) [3], as well as deep learning methods [4]. In this work, we focus on the multichannel NMF/NTF based methods [5, 3], which proved to be highly effective in convolutive audio source separation [6, 7, 2]. Their advantage is that these adaptive algorithms perform source separation without the need to use any training data. Of particular interest to the presented work is the generalized expectation maximization algorithm with multiplicative updates (GEM-MU) [8], in which the multichannel NTF based model is incorporated into the EM algorithm. In [7], the sub-source based expectation maximization algorithm (SSEM) has been proposed to improve on source separation performance. In [9], the sub-source based expectation maximization algorithm with multiplicative update rules (SSEM-MU) [9] has been presented, however, to the best of our knowledge, the evaluation of this method has not been shown in any publication.

In this paper, we aim to improve the performance of the multichannel NTF based separation for convolutive mixtures by additionally incorporating the localization prior distribution. To this end, we adapt the approach presented in [10] to incorporate the localization prior on the mixing matrix into the posterior probability. We propose the MAP estimator with sub-source based multichannel NTF model, and present the update equations for the derived sub-source based expectation maximization (EM) algorithm with multiplicative update rules (MU) and the localization prior on the mixing matrix (SSEM-MU-LP). Subsequently, we perform an extensive evaluation of the proposed algorithm in experiments with speech and instrumental sound sources, in determined and under-determined scenarios, with random and perturbed oracle parameter initialization. The superior results of the proposed algorithm over the existing algorithms motivates the choice of incorporating the localization prior.

2. SIGNAL MODEL

Let us consider a scenario in which J sources are recorded in a reverberant space using I microphones. Our aim is to blindly separate the unknown signals of J sources from a convolutive I-channel mixture. For $i = 1, ..., I$ and $j = 1, ..., J$, the vector of the microphone signals in the Short-Time Fourier Transform (STFT) domain $\mathbf{x}_{fn} = [X_{1fn}, X_{2fn}, ..., X_{Ifn}]^{\mathrm{T}} \in \mathbb{C}^I$ can be expressed as

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{jfn}, \tag{1}$$

where $\mathbf{y}_{jfn} = [Y_{j1fn}, Y_{j2fn}, \ldots, Y_{jIfn}]^{\mathrm{T}} \in \mathbb{C}^I$ denotes the vector with the signals of the j-th source in the STFT domain as captured by all I microphones (in the source separation literature such contribution of the source to the I-channel mixture is referred to as the spatial image of a source). The time and frequency indices of the STFT are given by $n = 1, \ldots, N$ and $f = 1, \ldots, F$, respectively. In order to model spatial and spectral cues, we assume the so-called local Gaussian model (LGM) [11] in which each j-th source spatial image is modeled as a zero-mean circular complex Gaussian vector

$$\mathbf{y}_{jfn} \sim N_c(0, \mathbf{R}_{jf} V_{jfn}), \tag{2}$$

with the time-invariant, complex-valued, full-rank spatial covariance matrix $\mathbf{R}_{jf} \in \mathbb{C}^{I \times I}$ and non-negative spectral variance $V_{jfn} \in \mathbb{R}_+$. In the following, we apply a joint NTF model which factorizes the

This research received financial support from the Foundation for Polish Science under grant number First TEAM/2017-3/23 (POIR.04.04.00-00-3FC4/17-00) which is co-financed by the European Union under the European Regional Development Fund.

matrix with spectral variances for all sources $\mathbf{V} \in \mathbb{R}_+^{J \times F \times N}$ into a sum of three nonnegative matrices: $\mathbf{Q} \in \mathbb{R}_+^{j \times k}$, $\mathbf{W} \in \mathbb{R}_+^{f \times k}$, $\mathbf{H} \in \mathbb{R}_+^{k \times n}$ [3, 8]. The spectral variance for the j-th source is then structured with the NTF model as

$$V_{jfn} = \sum_{k=1}^{K} Q_{jk} W_{fk} H_{kn}, \tag{3}$$

where $Q_{jk} = [\mathbf{Q}]_{jk}$, $W_{fk} = [\mathbf{W}]_{fk}$, $H_{kn} = [\mathbf{Q}]_{kn}$ are the elements of the respective matrices, K denotes the number of components, and $k = 1, \ldots, K$. Note that the columns of matrix $\mathbf{W}$ represent the frequency profiles, the rows of matrix $\mathbf{H}$ represent time activations, while matrix $\mathbf{Q}$ maps k-th component to the j-th source. Given (1) and (2), the following microphone signal model is obtained

$$\mathbf{x}_{fn} \sim \sum_{j=1}^{J} N_c(0, \mathbf{R}_{jf} V_{jfn}). \tag{4}$$

Next, following [7] we assume that the latent data consists of the so-called sub-sources which share the same spectral variance for a given source. For the j-th source, we introduce I sub-sources modeled as

$$S_{jifn} \sim N_c(0, V_{jfn}), \tag{5}$$

and collect them into a sub-source vector $\mathbf{s}_{fn} = [S_{11fn}, \ldots, S_{1Ifn}, S_{21fn}, \ldots, S_{2Ifn}, \ldots, S_{J1fn}, \ldots, S_{JIfn}]^{\mathrm{T}} \in \mathbb{R}_+^{JI}$. We assume that the components of $\mathbf{s}_{fn}$ are mutually independent between the sources. On the other hand, the spatial covariance matrix $\mathbf{R}_{jf}$ can be nonuniqualy represented as

$$\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H, \tag{6}$$

where $\mathbf{A}_{jf} \in \mathbb{C}^{I \times I}$ denotes the mixing matrix with steering vectors for all sub-sources of the j-th source and $\{\cdot\}^H$ denotes the conjugate transpose operator. In this work, we assume that the mixing matrix is full rank in order to model the entire room reverberation [10]. Finally, we can obtain the signal model

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_f, \tag{7}$$

which follows from the model given by (4), in which $\mathbf{A}_f = [\mathbf{A}_{1f}, \mathbf{A}_{2f}, \ldots, \mathbf{A}_{Jf}]^{\mathrm{T}}$, and the noise term $\mathbf{b}_f = [B_{f1}, B_{f2}, \ldots, B_{fI}]^{\mathrm{T}} \in \mathbb{C}^I$ is added. The introduced noise term can be, e.g., used in the so-called simulated annealing [6]. The noise vector is modeled as

$$\mathbf{b}_f \sim N_c(0, \mathbf{\Sigma}_{b,f}), \tag{8}$$

with the noise covariance matrix given by $\mathbf{\Sigma}_{b,f} = \boldsymbol{\sigma}_{b,f}^2 \boldsymbol{I}_{I \times I}$.

3. MAXIMUM A POSTERIORI ESTIMATOR WITH THE LOCALIZATION GAUSSIAN PRIOR

In this section, we present the proposed maximum a posteriori (MAP) estimator for the model described in Sec. 2. with the localization prior distribution on the mixing matrix which consists of the steering vectors. In order to estimate the parameters of the probabilistic model $\mathbf{\Theta} = \{\mathbf{A}, \mathbf{Q}, \mathbf{W}, \mathbf{H}, \mathbf{\Sigma}_b\}$ with a prior distribution on $\mathbf{A}$, we formulate the following posterior where $\mathbf{X}$ is the observed microphone mixture and the latent data consists of the sub-sources $\mathbf{S}$. The negative log-posterior of the complete data $\{\mathbf{X}, \mathbf{S}\}$ can then be written as

$$-\log P(\mathbf{X}, \mathbf{S}|\mathbf{\Theta}) = \log P(\mathbf{X}|\mathbf{S}, \mathbf{\Theta}) + \log P(\mathbf{S}|\mathbf{\Theta}) + \log P(\mathbf{A}). \tag{9}$$

In this work, we assume that the mixing matrix for the j-th source can be modeled as a complex Gaussian distribution [10]

$$\mathbf{A}_{jf} \sim N_c(\boldsymbol{U}_{jf}, \mathbf{\Sigma}_{jf}), \tag{10}$$

with mean $\boldsymbol{U}_{jf} \in \mathbb{C}^{I \times I}$ and covariance matrix $\mathbf{\Sigma}_{jf} \in \mathbb{C}^{I \times I}$. Such a mixing matrix $\mathbf{A}_{jf}$ can be selected so that the signal that propagates over the direct path is concentrated only in the mean of the prior for the first sub-source (i.e., for $i = 1$), which can be written as

$$\boldsymbol{U}_{jf} = \left[\mathbf{d}_{jf}, \mathbf{O}_{I \times (I-1)}\right], \tag{11}$$

where $\mathbf{O}_{I \times (I-1)}$ is the zero matrix and the steering vector for the direct path signal is given by $\mathbf{d}_{jf} = \left[1, e^{-\jmath\omega\,\tau_2}, ..., e^{-\jmath\omega\,\tau_I}\right]^{\mathrm{T}}$, where $\omega = 2\pi f$, $\jmath = \sqrt{-1}$, and τ_i denotes the delay between the direct signal that propagated from the source to the i-th microphone relative to the delay for the first microphone. In addition, we model the late room reverberation using the so-called spatial coherence matrix $\mathbf{\Omega}_f \in \mathbb{C}^{I \times I}$ [12], whose elements are given by $[\mathbf{\Omega}_f]_{ii`} = \mathrm{sinc}(\frac{\omega||\mathbf{p}_i - \mathbf{p}_{i'}||_2}{\nu})$, where $\mathbf{p}_i$ and $\mathbf{p}_{\mathrm{i}`}$ are the respective microphone positions and ν denotes the wave velocity. Since the diffuse late reverberation cancels out on average, it is taken into account only in the covariance of the prior, which can be expressed as

$$\mathbf{\Sigma}_{jf} = \mathbf{\Omega}_f. \tag{12}$$

Next, by introducing the conditional expectation operator $\mathbb{E}_{X|S,\Theta^l}[\cdot]$ to the negative log-posterior (9), the following cost function $Q_p(\Theta, \Theta^l)$ to be minimized is obtained (up to constant values)

$$\begin{aligned} Q_p(\Theta, \Theta^l)_{fn} = \sum_{f,n} \mathrm{Tr}\Big\{ \mathbf{\Sigma}_{b,fn}^{-1} \Big(\widehat{\mathbf{R}}_{xx,fn} - \mathbf{A}_f \widehat{\mathbf{R}}_{xs,fn}^H - \\ \widehat{\mathbf{R}}_{xs,fn} \mathbf{A}_f^H + \mathbf{A}_f \widehat{\mathbf{R}}_{ss,fn} \mathbf{A}_f^H \Big) \Big\} + \sum_{f,n} \log |\mathbf{\Sigma}_{b,f}| + \\ I \sum_{j,f,n} \mathrm{d}_{IS}\left(\widehat{\xi}_{jfn} | V_{jfn} \right) + \gamma \log N_c(\mathbf{A}_f | \boldsymbol{U}_f, \mathbf{\Sigma}_f), \end{aligned} \tag{13}$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace operator, γ is the trade-off hyperparameter for the localization prior, $\mathrm{d}_{IS}(\widehat{\xi}_{jfn}|V_{jfn})$ denotes the Itakura-Saito divergence [13], and the so-called sufficient statistics are given by $\widehat{\mathbf{R}}_{xx,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{X}_{fn}\mathbf{X}_{fn}^H]$, $\widehat{\mathbf{R}}_{xs,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{X}_{fn}\mathbf{S}_{fn}^H]$, $\widehat{\mathbf{R}}_{ss,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{S}_{fn}\mathbf{S}_{fn}^H]$ and $\widehat{\xi}_{jfn} = \mathbb{E}_{X|S,\Theta^l}[\frac{1}{I}\sum_i |S_{ji,fn}|^2]$.

The proposed expectation maximization algorithm which estimates the model parameters $\mathbf{\Theta}$ is derived in Sec. 4. The final step of the presented source separation is to retrieve the spatial source images from the microphone signals in the STFT domain using the well-known multichannel Wiener filter, which is given by

$$\widehat{\mathbf{Y}}_{jfn} = \mathbf{R}_{jf} V_{jfn} \left[\sum_{j=1}^{J} \mathbf{R}_{jf} V_{jfn} \right]^{-1} \mathbf{X}_{fn}. \tag{14}$$

4. THE PROPOSED SSEM-MU-LP ALGORITHM

In this section, we present the update equations for the proposed sub-source based expectation maximization (EM) algorithm with multiplicative update rules and the localization prior distribution on the mixing matrix (hereafter denoted as SSEM-MU-LP or the proposed algorithm). The derived update equations for minimizing the cost function (13) using the proposed EM algorithm are provided below.

Note that since our algorithm builds on the SSEM-MU algorithm with sub-source modeling, the presented update rules for the E-step are analogous to the ones presented in [9], while they are different for the M-step. In the E-step, we compute the conditional expectation of sufficient statistics $\widehat{\mathbf{R}}_{xx,fn}, \widehat{\mathbf{R}}_{xs,fn}, \widehat{\mathbf{R}}_{ss,fn}$ and $\widehat{\xi}_{jfn}$ using the following update equations [9]

$$\widehat{\mathbf{R}}_{xx,fn} = \mathbf{A}_f \mathbf{R}_{ss,fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{b,f}, \tag{15}$$

$$\widehat{\mathbf{R}}_{xs,fn} = \mathbf{R}_{xx,fn} \mathbf{G}_{s,fn}^H, \tag{16}$$

$$\widehat{\mathbf{R}}_{ss,fn} = \mathbf{G}_{s,fn} \mathbf{R}_{xx,fn} \mathbf{G}_{s,fn}^H + (\mathbf{I}_{JI} - \mathbf{G}_{s,fn}\mathbf{A}_f)\mathbf{R}_{ss,fn}, \tag{17}$$

$$\widehat{\xi}_{jfn} = \frac{1}{I} \sum_{i=(j-1)I+1}^{jI} \widehat{\mathbf{R}}_{ss,fn}(i,i), \tag{18}$$

where $\mathbf{G}_{s,fn}$ and $\mathbf{R}_{ss,fn}$ are calculated using the current estimates of parameters, and they are given by

$$\mathbf{G}_{s,fn} = \mathbf{R}_{ss,fn} \mathbf{A}_f^H \widehat{\mathbf{R}}_{xx,fn}^{-1}, \tag{19}$$

$$\mathbf{R}_{ss,fn} = \text{diag}([\overbrace{V_{1fn}, \ldots, V_{1fn}}^{I \text{ times}}, \ldots, \overbrace{V_{Jfn}, \ldots, V_{Jfn}}^{I \text{ times}}]). \tag{20}$$

In the M step, following [10], we minimize the cost function (13) over $\mathbf{A}_f$, which yields the following closed-form solution

$$\begin{aligned} \mathbf{A}_f = & \left[\gamma \mathbf{\Sigma}_f^{-1} + \frac{1}{\sigma_{b,f}^2 N} \sum_n \left(\widehat{\mathbf{R}}_{ss,fn} \otimes \mathbf{I}_I \right)^T \right]^{-1} \\ & \left[\gamma \mathbf{\Sigma}_f^{-1} \boldsymbol{U}_f + \frac{1}{\sigma_{b,f}^2 N} \sum_n \left(\widehat{\mathbf{R}}_{xs,fn}^H \right) \right], \end{aligned} \tag{21}$$

where $\otimes$ denotes the Kronecker product operator, and

$$\boldsymbol{U}_f = [\boldsymbol{U}_{1f}, \boldsymbol{U}_{2f}, \ldots, \boldsymbol{U}_{Jf}]^{\mathrm{T}}, \tag{22}$$

$$\mathbf{\Sigma}_f = \begin{bmatrix} \mathbf{\Sigma}_{jf} & & 0 \\ & \ddots & \\ 0 & & \mathbf{\Sigma}_{Jf} \end{bmatrix}. \tag{23}$$

However, since $\boldsymbol{U}_f$ and $\mathbf{\Sigma}_f$ are defined in our work as power-normalized matrices, while the powers of $\widehat{\mathbf{R}}_{ss}$ and $\widehat{\mathbf{R}}_{xs}$ depend on the powers of the microphones signals, we modify (22) to achieve the correct normalization of the latter matrices. The proposed closed-form solution is given by

$$\begin{aligned} \mathbf{A}_f = & \left[\gamma \mathbf{\Sigma}_f^{-1} + \frac{1}{\sigma_{b,f}^2 N} \sum_n \left(\overline{\mathbf{R}}_{ss,fn} \otimes \mathbf{I}_I \right)^T \right]^{-1} \\ & \left[\gamma \mathbf{\Sigma}_f^{-1} \boldsymbol{U}_f + \frac{1}{\sigma_{b,f}^2 N} \sum_n \left(\overline{\mathbf{R}}_{xs,fn}^H \right) \right], \end{aligned} \tag{24}$$

where

$$\overline{\mathbf{R}}_{ss,fn} = \frac{\widehat{\mathbf{R}}_{ss,fn}}{\text{Tr}\{\widehat{\mathbf{R}}_{ss,fn}\}}, \tag{25}$$

$$\overline{\mathbf{R}}_{xs,jfn} = \frac{\widehat{\mathbf{R}}_{xs,jfn}}{\text{Tr}\{\widehat{\mathbf{R}}_{xs,jfn}\}}, \tag{26}$$

and $\overline{\mathbf{R}}_{xs,fn} = \left[\overline{\mathbf{R}}_{xs,1fn}, \overline{\mathbf{R}}_{xs,2fn}, \ldots, \overline{\mathbf{R}}_{xs,Jfn} \right]^{\mathrm{T}}$.

In addition, in the M-step, parameters Q_{jk}, W_{fk}, and H_{kn} are updated using the so-called multiplicative update rules, which for the NTF model are given by

$$Q_{jk} \longleftarrow Q_{jk} \frac{\sum_{fn} W_{fk} H_{kn} \hat{\xi}_{jfn} V_{jfn}^{-2}}{\sum_{fn} W_{fk} H_{kn} V_{jfn}^{-1}}, \tag{27}$$

$$W_{fk} \longleftarrow W_{fk} \frac{\sum_{jn} H_{kn} Q_{jk} \hat{\xi}_{jfn} V_{jfn}^{-2}}{\sum_{jn} H_{kn} Q_{jk} V_{jfn}^{-1}}, \tag{28}$$

$$H_{kn} \longleftarrow H_{kn} \frac{\sum_{jf} W_{fk} Q_{jk} \hat{\xi}_{jfn} V_{jfn}^{-2}}{\sum_{jf} W_{fk} Q_{jk} V_{jfn}^{-1}}. \tag{29}$$

Note that at the end of each iteration, the estimated matrices $\mathbf{A}_f$, $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ should be normalized, as described, e.g., in [8], to avoid scale and phase ambiguity. In the proposed algorithm, $\mathbf{\Sigma}_b$ is updated using the so-called simulated annealing (for more details please refer to [6]). Regarding computational complexity of the proposed algorithm, note that the localization prior adds very little to the overall complexity since both first terms in square brackets of (21) do not require recalculation over the iterations. In fact, the localization information significantly increases algorithm's convergence speed, and as a consequence much less iterations are required.

5. EXPERIMENTAL EVALUATION

5.1. Experimental setup and evaluation criteria

This section presents the description of performed experiments, the compared methods, and the criteria used in evaluation. The performance of the compared source separation methods is evaluated in a set of numerical experiments performed in a room of size $10 \times 10 \times 4$ m with a reverberation time of 250 ms. In all experiments, two microphones ($I = 2$) with the microphone spacing of 0.05 m are located around the room center and the sources are located randomly at a distance of 2 m from the microphones under an additional constraint that the distance between the sources is larger or equal to 1 m. The microphone signals are obtained as a sum of convolutions of non-reverberant source signals and the respective room impulse responses between the sources and the microphones simulated using the image-source method [14]. Two types of source signals are selected, namely, the speech signals and the instrumental signals. All non-reverberant recordings of the source signals are taken from the EBU SQAM database [15] with 6 speech excerpts of 3 male and 3 female speakers and 6 instrumental recordings of flute, violoncello, piano, accordion, trumpet, and marimba. The final microphone signals sampled at 16 kHz are trimmed to 8 s, and a 2048-point STFT with 50% overlap is used to process the signals in the STFT domain.

The proposed SSEM-MU-LP algorithm is compared with three state-of-the-art source separation algorithms, which all contain various elements of the proposed processing. These three methods are: the generalized expectation maximization algorithm with multiplicative update rules (GEM-MU) [8], the sub-source based expectation maximization algorithm (SSEM) [7] and the sub-source based expectation maximization algorithm with multiplicative update rules (SSEM-MU) [9]. Four standard evaluation measures are used to investigate the separation performance, namely the signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR), respectively [16]. The experiments are performed for two types of scenarios, namely, the determined scenario (in which $J = I = 2$) and the under-determined scenario (in which $J = 3$ and

Table 1. SDR, SIR, ISR, and SAR results obtained with parameters estimated in the 100-th iteration for the Proposed, SSEM-MU, SSEM, and GEM-MU algorithms for determined and under-determined scenarios with speech and instrumental sources, initialized randomly and with the perturbed oracle values.

Scenario			Determined								Under-determined							
Sources			Speech				Instrumental				Speech				Instrumental			
Parameter			SDR	SIR	ISR	SAR	SDR	SIR	ISR	SAR	SDR	SIR	ISR	SAR	SDR	SIR	ISR	SAR
Initialization	Random	Proposed	**8.6**	**12.0**	**13.0**	**13.6**	**13.1**	**17.3**	**17.6**	**20.3**	**5.3**	**8.1**	**9.5**	**8.2**	**9.2**	**13.9**	**13.5**	**13.3**
		SSEM-MU[9]	1.6	2.2	6.6	4.3	1.8	1.9	6.1	6.9	0.7	-0.8	4.2	2.8	0.8	-0.2	4.2	4.0
		SSEM [7]	1.7	1.5	6.3	4.7	1.9	1.6	6.0	6.9	0.6	-1.4	4.0	2.4	0.4	-1.2	3.9	3.4
		GEM-MU [8]	-0.3	1.9	6.0	0.8	0.0	1.8	5.5	2.6	-0.7	-0.4	4.2	0.1	-0.7	0.0	4.1	1.1
	Perturbed Oracle	Proposed	**9.5**	14.0	**14.7**	**12.9**	**15.6**	**20.6**	**20.8**	**20.6**	**5.7**	**9.1**	**10.1**	**8.4**	**9.4**	**14.9**	**14.5**	**13.3**
		SSEM-MU[9]	7.6	**14.1**	14.4	10.6	7.2	16.6	16.8	8.7	3.0	7.2	9.5	4.4	5.0	10.7	12.0	7.0
		SSEM [7]	6.7	10.9	12.1	9.7	7.4	12.7	13.6	10.3	2.8	4.8	7.2	4.5	3.4	5.8	7.8	6.8
		GEM-MU [8]	4.4	11.9	12.5	6.1	5.2	13.9	14.5	6.7	1.0	5.5	8.0	2.1	5.5	12.4	13.4	7.0

$I = 2$). In each scenario, 16 speech and 16 instrumental recordings with random source positioning in a room are used. Each result (in dB) presented in this paper is obtained by averaging over 160 results, obtained by performing 10 experiments with random initialization for each of the 16 recordings.

Finally, since the compared state-of-the-art algorithms strongly depend on the initialization of parameters, we perform the experiments for two types of initialization. At first, the parameters of the respective algorithms ($\mathbf{A}, \mathbf{Q}, \mathbf{W}, \mathbf{H}$) are initialized with random values drawn from the uniform distribution, whereby the random values in matrices $\mathbf{W}$ and $\mathbf{H}$ are additionally scaled with the power of the observed mixture averaged over time and frequency, respectively. Secondly, in order to provide 'good' initialization, the parameters are initialized with perturbed oracle values following a procedure presented in detail in [6]. In particular, the matrices $\mathbf{Q}$, $\mathbf{W}$, and $\mathbf{H}$ are initialized with the results of the IS-NMF algorithm [13] for a single source recording, while $\mathbf{A}$ is initialized using the original mixing system; all matrices with the parameters are then perturbed with high-level additive noise. Irrespective of the initialization procedure, in case some of the already existing techniques did not converge (SDR $<$-10 dB), such results were omitted from the calculation of the final averaged values.

5.2. Results and discussion

Figure 1 presents the SDR results obtained using the proposed SSEM-MU-LP algorithm and three existing algorithms (SSEM-MU, SSEM, and GEM-MU) for a number of performed experiments. Several general observations can be made. Firstly, the results of source separation for instrumental signals are in general better than those for the speech signals irrespective of the investigated algorithm. This can be attributed to the larger differences in time-frequency representations for different instruments. Secondly, the results obtained for the determined scenarios are much better than those for under-determined scenarios with the gain of circa 1/3 to 1/2 of the final SDR score. Comparing four studied algorithms, one can clearly observe that the proposed algorithm yields significant improvement in SDR compared to the state-of-the-art techniques for all performed experiments. In addition, the proposed algorithm is shown to be significantly more robust against random initialization in comparison with all three existing techniques. In contrary, other algorithms struggle to achieve good performance unless their parameters are sufficiently well initialized.

The averaged results of all performed experiments in terms of four evaluation measures are presented in Table 1. These results confirm the conclusions drawn from Fig. 1. As can be observed, with random initialization, some of the existing methods do not converge, and they achieve poor source separation. Furthermore, three existing techniques produce much lower SAR than the proposed method, while SIR and ISR can be considered closer to each other, which in turn causes SDR scores to be significantly better for the proposed method than for all other studied methods. Finally, including more advanced models such as using subs-source and adding multiplicative update rules yields improved separation performance.

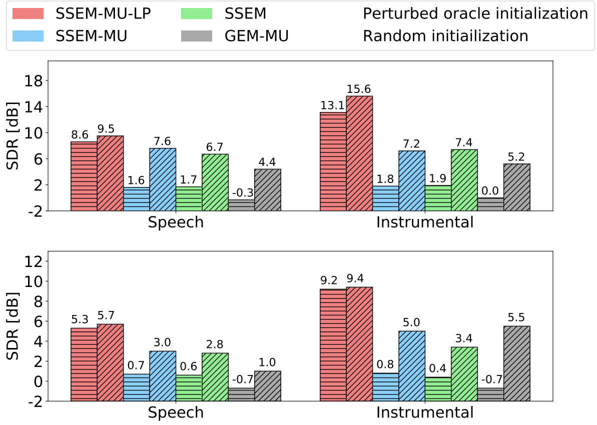

Fig. 1. Signal-to-distortion ratio (SDR) averaged over 10 random experiments in a determined (upper plot) and under-determined (bottom plot) scenarios for source signals estimated with perturbed initialization which are found in the 100-th iteration.

6. CONCLUSIONS

In this paper, we have presented a novel sound source separation method based on the maximum a posteriori estimator with NTF model and the localization prior. The derived sub-source based expectation maximization algorithm with multiplicative update rules and the localization prior has been shown in a number of experiments to outperform state-of-the-art source separation algorithms that do not account for the a priori information. The key advantage of the proposed algorithm is that it achieves very good separation results without any training or knowledge about the room characteristics.

7. ACKNOWLEDGEMENTS

Mr Jan Wilczek and Mr Mateusz Guzik are thanked for their help with early developments of the studied separation algorithms.

8. REFERENCES

[1] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, “Blind source separation combining independent component analysis and beamforming,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 569270, 2003.

[2] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[4] A. A. Nugraha, A. Liutkus, and E. Vincent, “Multichannel audio source separation with deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[5] L. Benaroya, L. M Donagh, F. Bimbot, and R. Gribonval, “Non negative sparse representation for wiener based source separation with a single sensor,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03).* IEEE, 2003, vol. 6, pp. VI–613.

[6] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[7] A. Ozerov, E. Vincent, and F. Bimbot, “A general flexible framework for the handling of prior information in audio source separation,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[8] A. Ozerov, C. Févotte, R. Blouet, and J. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2011, pp. 257–260.

[9] A. Ozerov, C. Févotte, and E. Vincent, “An introduction to multichannel nmf for audio source separation,” in *Audio Source Separation*, pp. 73–94. Springer, 2018.

[10] N. QK. Duong, E. Vincent, and R. Gribonval, “Spatial location priors for gaussian model based reverberant audio source separation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 149, 2013.

[11] N. QK. Duong, E. Vincent, and R. Gribonval, “Underdetermined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[12] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[13] C. Févotte, N. Bertin, and J. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[14] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[15] “Sound quality assessment material recordings for subjective tests,” https://tech.ebu.ch/publications/sqamcd, 2008, Accessed: 2020-06-11.

[16] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.