AUDIO SOURCE SEPARATION USING A REDUNDANT LIBRARY OF SOURCE SPECTRAL BASES FOR NON-NEGATIVE TENSOR FACTORIZATION

Mahmoud Fakhry[1,2]

[1]Doctoral School of ICT
University of Trento
via Sommarive 5, 38123 Trento, Italy
abdelraheem@fbk.eu

Piergiorgio Svaizer[2], *and Maurizio Omologo*[2]

[2]Center of Information Technology
Fondazione Bruno Kessler - irst
via Sommarive 18, 38123 Trento, Italy
{svaizer,omologo}@fbk.eu

ABSTRACT

This work proposes a solution to the problem of under-determined audio source separation using pre-trained redundant source-based prior information. In local Gaussian modeling of a mixing process, an observed mixture is modeled by a Gaussian distribution parameterized by source variances and spatial covariance matrices. The separation is performed by estimating the parameters, and applying Wiener filtering on the observed mixture. We propose, in a training phase, to build a redundant library of spectral basis matrices of all probable source power spectra, applying non-negative tensor factorization (NTF). In the testing phase, the matrices that match the observed mixture are detected using NTF. With the help of the detected matrices, a maximum likelihood algorithm is proposed in order to iteratively estimate the parameters of the model, exploiting the spatial redundancy of the observed mixture and using NTF. The proposed algorithm proves more flexibility and efficiency with respect to a baseline algorithm used as a reference.

Index Terms— Spectral bases, redundant library, non-negative tensor factorization, model parameters, audio source separation.

1. INTRODUCTION

Using side information has recently raised as a new trend to increase the performance of blind source separation [1, 2]. Several forms of the information have been exploited to solve the problem in [3, 4, 5, 6]. Audio source separation for indoor conversations can benefit from information about the parameters of mixing environments [7, 8], the spatial locations of speakers [9], and the spectral variances of source signals [10, 11]. The variance of source signals can be modeled using Gaussian mixture models (GMMs) [12] or non-negative matrix factorization (NMF) [11, 13]. In the NMF approaches, source spectrum is decomposed into the multiplication of two nonnegative matrices: a spectral basis matrix that contains parts of the spectrum, and a coefficient matrix that contains time-varying weights. These two matrices may be estimated during the separation process [13], or, to achieve a higher separation performance, the spectral basis matrix may be predefined as prior information [11, 13]. For speech enhancement purposes, a dictionry of spectral bases of multiple source signals is suggested to be used as prior information in [14, 15], and the selection of the optimal bases is done using some block sparsity constraints on top of the NMF objective. An extension to NMF has been considered by arranging multiple signal observations in a tensor form, under a parallel factorization analysis (PARAFAC) structure, where the observations form the slices of 3-D tensor [16, 17]. In non-negative tensor factorization (NTF), the tensor is decomposed into three matrices. The redundancy between the slices is described by two matrices as in NMF, while the diversity is represented by a third matrix. In the case of multichannel observations, the third matrix can be considered to describe source spatial diversity and its columns to convey spatial cues. In [18], the authors solve source separation of simple instantaneous mixtures by giving weights to the time-frequency points of observations according to their closeness to the spatial cues, assumed to be known as prior information.

In [11], we proposed a reverberant audio source separation algorithm, exploiting side information about the source variance. In a training step, the power spectra of each source signal files are concatenated and decomposed using NMF, and the side information is defined as the spectral basis matrices of the sources. Assuming that the spatial locations of speakers are labeled for each source, a spectral basis matrix is predefined and fixed for each label, and then it is used to perform the separation process. Using the information, we follow the local Gaussian model [19] to probabilistically parametrize the observed mixture by a set of parameters, source variances and spatial covariance matrices. To perform the separation using the information, the parameters are estimated and used to compute multichannel Wiener gains that are applied to extract the contribution of each source signal from the observed mixture. Although the algorithm performs well and achieves a good separation performance, it suffers from two main weak points that we try to mitigate in this work. The first point is the reliability of fixing spectral basis matrices for each source in advance, which means that the order of source spatial locations should be identified a priori. The second point is that we are loosing important spatial redundancy in the training and the estimation since we used NMF.

In order to mitigate the first weak point and to increase flexibility of the algorithm, we propose to build a redundant (over-complete) library of spectral basis matrices of all available sources. Then the matrices matching the observed mixture are detected and exploited to separate the mixture. To mitigate the second weak point and to exploit spatial redundancy of the observed mixture, we replace NMF for the training and the estimation with non-negative tensor factorization (NTF). The rest of the paper is organized as follows. In Section 2, we present the formulation of the problem. The proposed algorithm is explained in Section 3 and the experimental evaluation are reported in Section 4. Finally Section 5 concludes the paper.

2. PROBLEM FORMULATION

We assume that N sources are observed by an array of M-microphones. In the time-frequency domain, let the signal generated by the *n-th* source and the signal at the *m-th* microphone be denoted by $S_n(f, l)$ and $X_m(f, l)$, respectively, where (f, l) indicates the in-

The research leading to these results has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement n. 288121 - DIRHA (see http://dirha.fbk.eu/).

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

dexes of a time-frequency point out of L total number of frames and F total number of frequency bins. The vector of the observed mixture $\mathbf{X}(f,l) = [X_1(f,l) \cdots X_M(f,l)]^T$ can be modeled, as follows

$$\mathbf{X}(f,l) = \sum_{n=1}^{N} \mathbf{c}_n(f,l), \tag{1}$$

where $\mathbf{c}_n(f,l)$ is a $M \times 1$ vector that defines the spatial image of the *n-th* source at the microphones as

$$\mathbf{c}_n(f,l) = \mathbf{h}_n(f) S_n(f,l), \tag{2}$$

and $\mathbf{h}_n(f)$ is a $M \times 1$ vector that corresponds to the transfer function between the *n-th* source and the microphones at the *f-th* frequency, and defines the contribution of the source in the mixture.

In the local Gaussian modeling of a mixing process [19], the coefficients of the source spatial image $\mathbf{c}_n(f,l)$ are assumed to be independent from each other, and from one source to the others. The coefficients are probabilistically modeled by a zero-mean Gaussian random vector, i.e. $\mathbf{c}_n(f,l) \sim \mathcal{N}(0, \mathbf{\Sigma}_{\mathbf{c}_n}(f,l))$, with a covariance matrix $\mathbf{\Sigma}_{\mathbf{c}_n}(f,l)$ computed as

$$\mathbf{\Sigma}_{\mathbf{c}_n}(f,l) = v_n(f,l) \mathbf{R}_n(f), \tag{3}$$

where $v_n(f,l)$ is a scalar variance encoding the power spectrum of the source signal $S_n(f,l)$, and $\mathbf{R}_n(f)$ is a $M \times M$ spatial covariance matrix encoding the source spatial information $\mathbf{h}_n(f)$. Source separation can be performed by estimating the model parameters $\theta = \{v_1(f,l), ..., v_N(f,l), \mathbf{R}_1(f), ..., \mathbf{R}_N(f)\}$. Moreover, the spatial images of all sources are derived in the minimum mean square error (MMSE) sense using the multichannel Wiener filtering of the mixture such as

$$\tilde{\mathbf{c}}_n(f,l) = \mathbf{\Sigma}_{\mathbf{c}_n}(f,l) (\sum_{n=1}^{N} \mathbf{\Sigma}_{\mathbf{c}_n}(f,l))^{-1} \mathbf{X}(f,l). \tag{4}$$

3. THE PROPOSED ALGORITHM

We propose a new framework of source separation using source-based pre-trained information. In the first step, a redundant library of spectral basis matrices is built by factorizing the power spectra of all available source signals using NTF. Secondly, unlike as in [14, 15], we detect multiple spectral basis matrices best matching the actual source signals in the observed mixture $\mathbf{X}(f,l)$. Finally, the detected matrices are used to iteratively estimate the set of parameters θ in the ML sense, and the Wiener filtering process in (4) is applied on the observed mixture to extract the contribution of each source signal.

3.1. Building of the redundant library

Let us define $\mathbf{V}^I(f,l)$ as a tensor of I signal examples, each *i-th* slice of the tensor is the power spectrum matrix of each signal example, where $\mathbf{V}^i(f,l) = [v(f,l)]^i_{F \times L}$ is a matrix of size $F \times L$. NTF approximately represents the tensor as

$$\mathbf{V}^I(f,l) \simeq \mathbf{U}(f,k) \otimes \mathbf{D}^I(k,k) \otimes \mathbf{W}(k,l). \tag{5}$$

where $\otimes$ denotes the tensor multiplication, where each *i-th* slice of the tensor is the multiplication of a $F \times K$ spectral basis matrix $\mathbf{U}(f,k)$ (K is the number of bases), the diagonal *i-th* slice of a $K \times K \times I$ tensor $\mathbf{D}^I(k,k)$, and a $K \times L$ time-varying coefficient matrix $\mathbf{W}(k,l)$. The factorization is achieved by minimizing an error function between $\mathbf{V}^I(f,l)$ and $\mathbf{U}(f,k) \otimes \mathbf{D}^I(k,k) \otimes \mathbf{W}(k,l)$, under the non-negativity constraint of the coefficients of the matrices. The multiplicative update rule of minimizing the Kullback-Leibler (KL) divergence as the error function is given by [16, 17]

$$\begin{aligned}
\mathbf{U}(f,k) &\leftarrow \mathbf{U}(f,k) \frac{\sum_i (\mathbf{V}^i(f,l) \oslash \hat{\mathbf{V}}^i(f,l)) \mathbf{W}(k,l)^T \mathbf{D}^i(k,k)}{\sum_i \mathbf{1}_{FL} \mathbf{W}(k,l)^T \mathbf{D}^i(k,k)}, \\
\mathbf{W}(k,l) &\leftarrow \mathbf{W}(k,l) \frac{\sum_i \mathbf{D}^i(k,k) \mathbf{U}(f,k)^T (\mathbf{V}^i(f,l) \oslash \hat{\mathbf{V}}^i(f,l))}{\sum_i \mathbf{D}^i(k,k) \mathbf{U}(f,k)^T \mathbf{1}_{FL}}, \\
\mathbf{D}^i(k,k) &\leftarrow \mathbf{D}^i(k,k) \frac{\mathbf{U}(f,k)^T (\mathbf{V}^i(f,l) \oslash \hat{\mathbf{V}}^i(f,l)) \mathbf{W}(k,l)^T}{\mathbf{U}(f,k)^T \mathbf{1}_{FL} \mathbf{W}(k,l)^T},
\end{aligned} \tag{6}$$

where $\oslash$ indicates point-wise division, $\mathbf{1}_{FL}$ is a $F \times L$ matrix of ones, and $\hat{\mathbf{V}}^i(f,l)$ is the *i-th* matrix of the estimated tensor. To constitute the library for a number Z of training source signals as in [14], the spectral basis matrices are estimated and sequentially arranged as

$$\mathbf{U}_Z = [\mathbf{U}_1(f,k) | \mathbf{U}_2(f,k) | \cdots | \mathbf{U}_Z(f,k)], \tag{7}$$

where $\mathbf{U}_Z$ is a library of Z spectral matrices and of size $F \times ZK$.

3.2. Detection of the matched spectral basis matrices

Assuming that an estimation of the source spatial image $\tilde{\mathbf{c}}_n(f,l)$ is defined, an empirical covariance matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)$ can be calculated. The diagonal coefficients of the matrix approximately describe the variance at the (f,l) point of the *n-th* source weighted by propagation inter-channel intensities. We arrange the diagonal coefficients in a tensor of size $F \times L \times M$ that is factorized using the predefined redundant library such as

$$\tilde{\mathbf{V}}_n^M(f,l) = \mathbf{U}_Z \otimes \tilde{\mathbf{D}}_n^M \otimes \tilde{\mathbf{W}}_n, \tag{8}$$

where $\tilde{\mathbf{W}}_n$ is a $ZK \times L$ coefficient matrix, and $\tilde{\mathbf{D}}_n^M$ is a diagonal tensor of size $ZK \times ZK \times M$, encoding the contribution of each spectral basis vector of the library in the tensor. To detect the index z of a basis matrix matching the true basis matrix $\mathbf{U}_n(f,k)$ from the library $\mathbf{U}_Z$, the tensor $\tilde{\mathbf{D}}_n^M$ is divided into Z sub-tensors $\tilde{\mathbf{D}}_n^M(k,k)|^z$, each of size $K \times K \times M$. To compensate the indeterminacy of the channel intensities, an averaging operation is performed in the *m-th* direction of the sub-tensors, converting it to a $K \times K$ diagonal matrix, and for each k a quadratic function is computed as

$$\Gamma_z(k) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\tilde{\mathbf{D}}_n^m(k,k)|^z)^2}. \tag{9}$$

The index of the matched basis matrix is detected by computing the likelihood of each sub-tensor, summing over the coefficients of the quadratic function in (9), and selecting the optimal index maximizing the likelihood function as

$$z^* = \arg\max_z \sum_{k=1}^{K} \Gamma_z(k), \;\; z = 1, 2, ..., Z. \tag{10}$$

For each source n, we detect an optimal index z^*.

3.3. Estimation of the model parameters

As in [11], we reformulate the model parameters to estimate to be $\hat{\theta} = \{\tilde{v}_1(f,l), ..., \tilde{v}_N(f,l), \bar{\mathbf{R}}_1(f), ..., \bar{\mathbf{R}}_N(f)\}$, where $\tilde{v}_n(f,l)$ is an estimated source variance corrupted by spatial information, defined as

$$\tilde{v}_n(f,l) = v_n(f,l) ||\mathbf{R}_n(f)||_F, \tag{11}$$

where $||.||_F$ indicates the Frobenius norm of a square matrix. $\bar{\mathbf{R}}_n(f)$ is a normalized spatial covariance matrix, defined as

$$\bar{\mathbf{R}}_n(f) = \frac{\mathbf{R}_n(f)}{||\mathbf{R}_n(f)||_F}. \tag{12}$$

Computing the estimated covariance matrix of the *n-th* source spatial image as the multiplication of the parameters in (11) and (12), leads to an estimation of the matrix as in (3), which is rewritten as

$$\mathbf{\Sigma}_{\mathbf{c}_n}(f,l) = \tilde{v}_n(f,l)\bar{\mathbf{R}}_n(f). \tag{13}$$

This reformulation gives us the opportunity to directly estimate $\tilde{v}_n(f,l)$, using the empirical covariance matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)$, regardless of the estimated spatial information. In order to estimate the parameters in the ML sense [11], the error between $\mathbf{\Sigma}_{\mathbf{c}_n}(f,l)$ in (13) and $\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)$ is minimized with respect to each parameter. The empirical covariance matrix can be computed as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l) = \frac{\sum_{\tilde{f},\tilde{l}} \gamma(\tilde{f}-f,\tilde{l}-l)\tilde{\mathbf{c}}_n(\tilde{f},\tilde{l})\tilde{\mathbf{c}}_n(\tilde{f},\tilde{l})^H}{\sum_{\tilde{f},\tilde{l}} \gamma(\tilde{f}-f,\tilde{l}-l)}, \tag{14}$$

where γ is a bi-dimensional window describing the shape of the neighborhood, and $.^H$ indicates matrix conjugate transposition.

3.3.1. Estimation of $\tilde{v}_n(f,l)$

The singular value decomposition represents a complex matrix by sigular values and two unitary matrices. As a specific case, a rank-1 complex matrix can be described by one singular value. Modeling $\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)$ as a rank-1 matrix, the decomposition can be specified such as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l) = \sigma_n(f,l)\mathbf{A}_n(f), \tag{15}$$

where $\sigma_n(f,l)$ is a scalar singular value, and $\mathbf{A}_n(f)$ is a $M \times M$ rank-1 unitary matrix. In [11] the spatially corrupted source variance is estimated to equal $\sigma_n(f,l)$. However, the estimation does not exploit the spectral-temporal redundancy between the time-frequency points of the estimated source spatial image. We propose a new estimation algorithm that does not just consider the spectral-temporal redundancy, but also the spatial redundancy between propagation channels. As an approximation, the diagonal coefficients of $\mathbf{A}_n(f)$ describe the inter-channel intensities, and the off-diagonal coefficients represent the cross-channel spatial information. We arrange the singular value $\sigma_n(f,l)$ times the diagonal coefficients of $\mathbf{A}_n(f)$ in the tensor $\tilde{\mathbf{V}}_n^M(f,l)$; where each (f,l) coefficient of the *m-th* matrix of the tensor is $\sigma_n(f,l)$ weighted by the (m,m) coefficient of $\mathbf{A}_n(f)$. The *n-th* tensor is decomposed using the detected spectral basis matrix $\mathbf{U}_n(f,k)$ introduced in section 3.2 such as

$$\tilde{\mathbf{V}}_n^M(f,l) = \mathbf{U}_n(f,k) \otimes \tilde{\mathbf{D}}_n^M(k,k) \otimes \tilde{\mathbf{W}}_n(k,l), \tag{16}$$

As it was previously stated, the diagonal coefficients of each $K \times K$ matrix of the tensor $\tilde{\mathbf{D}}_n^M(k,k)$ encode the contribution of each spectral basis vector. For the (k,k) vector of length M in $\tilde{\mathbf{D}}_n^M(k,k)$, we propose to select the *m-th* channel index that maximizes the contribution of each basis vector in $\mathbf{U}_n(f,k)$, and the optimal index of the coefficient is selected as follows

$$m^* = \arg\max_m \tilde{\mathbf{D}}_n^m(k,k), \;\; m = 1,2,...,M. \tag{17}$$

Then the spatially corrupted source variance is reconstructed as

$$\tilde{\mathbf{V}}_n(f,l) = [\tilde{v}_n(f,l)]_{F\times L} = \mathbf{U}_n(f,k)\tilde{\mathbf{D}}_n^{m^*}(k,k)\tilde{\mathbf{W}}_n(k,l). \tag{18}$$

3.3.2. Estimation of $\bar{\mathbf{R}}_n(f)$

Up to the factorization error, the modeled covariance matrix in (13) can be rewritten in the decomposition form as

$$\mathbf{\Sigma}_{\mathbf{c}_n}(f,l) = \mathbf{U}_n(f,k)\tilde{\mathbf{D}}_n^{m^*}(k,k)\tilde{\mathbf{W}}_n(k,l)\bar{\mathbf{R}}_n(f). \tag{19}$$

Rewriting the empirical covariance matrix is in the polar form $(\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l) = |\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)|\angle\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l))$, a tensor of size $F \times L \times M^2$ can be built from absolute values of the matrix coefficients. The tensor is factorized using the pre-defined spectral basis matrix, and the full representation of the matrix is written as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l) \Leftarrow \mathbf{U}_n(f,k) \otimes \tilde{\mathbf{D}}_n^{M^2}(k,k) \otimes \tilde{\mathbf{W}}_{cn}(k,l)\angle\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l), \tag{20}$$

where the arrow means that the tensor representation of the right hand side is rearranged in a matrix form in the left hand side. Minimizing the error with respect to $\bar{\mathbf{R}}_n(f)$ between $\mathbf{\Sigma}_{\mathbf{c}_n}(f,l)$ and $\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l)$ in (19) and (20), respectively, leads to an estimation of the normalized spatial covariance matrix such as

$$\bar{\mathbf{R}}_n(f) \Leftarrow \frac{1}{KL}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\tilde{\mathbf{D}}_n^{M^2}(k,k) \otimes \tilde{\mathbf{W}}_{cn}(k,l)}{\tilde{\mathbf{D}}_n^{m*}(k,k)\tilde{\mathbf{W}}_n(k,l)}\angle\tilde{\mathbf{R}}_{\mathbf{c}_n}(f,l), \tag{21}$$

where the division is a point-wise operation.

4. EXPERIMENTAL ANALYSIS AND RESULTS

A room with size $4.45 \times 3.35 \times 2.5$ meters and an array of 2 omni directional microphones spaced of 0.2 m are considered. The microphones are located in the middle of the room and have the same height (1.4 m) as the sources. The distance between source positions and a center point between the microphones is either 0.5 or 1 m. The source directions of arrival of three mixed sources are 35, 90, and 145 degrees. Synthetic room impulse responses (RIRs) are simulated through ISM [20] with a sampling frequency of 16 kHz for three reverberation times: $T_{60} = 200, 350$, or 500 ms. 6 native Italian speakers are considered as our audio sources, 3 males and 3 females, from the clean speech dataset of the DIRHA project (Distant-speech Interaction for Robust Home Applications). For each speaker, we have 20 signals, each signal of length 8.75 s. The clean speech signals are divided into 5 speech signals of test data and 15 of training data. 4 male-female combinations of mixtures of $N = 3$ speech sources (3 males, 3 females, 2 females and 1 male, and 2 males and 1 female) are generated by individually convolving the full length of the simulated RIRs with the original source signals and adding the source image contributions to each microphone. This resulted in a total of 20 test mixtures for each T_{60} and source-microphone distance. The discrete time-frequency representation of the mixture $\mathbf{X}(k,l)$ is obtained through STFT with a Hanning analysis window with length of 128 ms , or 2048 samples, with a shift factor of 64 ms ($L = 137$). The window γ for the computation of the empirical covariance matrix of the source image is of size 1×1. We evaluated the separation performance via the signal-to-distortion ratio (SDR) and source image-to-spatial distortion ratio (ISR) criteria in decibels (dBs) [21] using the spatial images of true sources $\mathbf{c}_n(f,l)$ and the estimated ones $\tilde{\mathbf{c}}_n(f,l)$.

4.1. Analysis of the detection algorithm

For each signal example ($I = 15$) of the training data of the speech folders ($Z = 6$), the source power spectrum is computed and arranged in the tensor $\mathbf{V}^{15}(f,l)$ of size $1025 \times 137 \times 15$. The tensor is factorized with a number K of bases equals to 50, and a spectral basis matrix $\mathbf{U}(f,k)$ of size 1025×50 is estimated to build the library $\mathbf{U}_Z$ of 6 spectral basis matrices of size 1025×300. On source-to-microphone distance of 1 m, and in a reverberant environment with reverberation time of 200 ms, a mixture of two males and one female speech signals was generated. As a function of separation iterations, we computed the cost function of calculating the likelihood of the sub-tensor coefficients associated with each spectral basis vector in (9), as well as the likelihood of each sub-tensor, summing over the likelihoods of the coefficients of the sub-tensor.

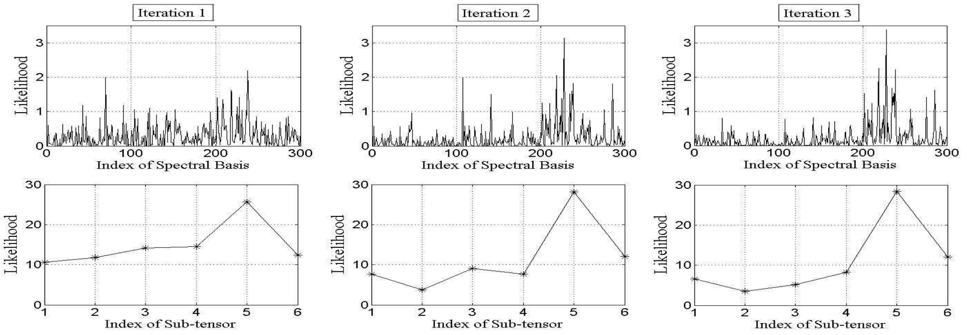

Fig. 1: Likelihood functions of the detection algorithm as a function of the separation iterations.

Table 1: Separation performance in dB for $K = 60$.
Source-to-microphone distance is $0.5\ m$.

T_{60}	SS-NTF		SS-NMF		Improvement(%)	
(ms)	SDR	ISR	SDR	ISR	SDR	ISR
200	9.06	15.40	8.35	14.80	8.50	4.05
350	7.91	13.60	7.26	13.07	9.00	4.05
500	6.90	12.12	6.30	11.70	9.52	3.60

Source-to-microphone distance is $1\ m$.

T_{60}	SS-NTF		SS-NMF		Improvement(%)	
(ms)	SDR	ISR	SDR	ISR	SDR	ISR
200	7.11	12.58	6.53	12.11	8.90	3.90
350	5.20	9.90	4.46	9.23	16.60	7.26
500	4.11	8.47	3.55	8.04	15.77	5.35

As we observe in Fig. 1, the likelihood of spectral bases associated with a certain target source (source number 5), increases iterating the separation algorithm, and the optimal index of the sub-tensor becomes more identifiable with respect to the other indexes.

In terms of source-to- microphone distances and reverberation times, the accuracy of the detection algorithm was tested on several female-male mixture combinations. For a distance of $0.5\ m$ and for each of the tested reverberation times ($200, 350$, and $500\ ms$), the algorithm detects with 100% the matched basis matrices. Moreover, the detection was successfully achieved with 100% on a distance of $1\ m$ and reverberation times of 200 and $350\ ms$. However, for what concerns highly reverberant environments, the accuracy of the detection reduced to around 65% when the distance was $1\ m$ and the reverberation time was $500\ ms$.

4.2. Source separation

As it was previously mentioned, the source separation is achieved by estimating the model parameters using the detected spectral basis matrices as in (18) and (21), computing the covariance matrix of the spatial image as in (13), and applying the Wiener filtering as in (4). The algorithm is initialized as in [11], and it converges after 4 or 5 iterations. We compared the performance of the proposed algorithm, called source separation using non-negative tensor factorization (SS-NTF), with an efficient source separation algorithm informed by spectral basis matrices factorized by non-negative matrix factorization (SS-NMF) in [11]. For the SS-NMF, the matrices are predefined and fixed for each source in advance with a number of bases K equals to 15, tuned to the best performed value. The SS-NMF algorithm outperforms the blind ML in [22] by around 4 dBs of SDR and 5 dBs of ISR. Besides the flexibility of SS-NTF in detecting the matched spectral basis matrices, the separation performance is further improved, as it is reported in Table1. Fixing the number of spectral bases K at 60, the table shows an example of the calculated values of the separation performance. When the spatial positions of the sources to the microphones is $0.5\ m$, there is an improvement in the SDR by around 9%, and a suppression of the amount of spatial distortion by around 4%. However, the performance increases in reverberant environments and far talking by around 16% of SDR.

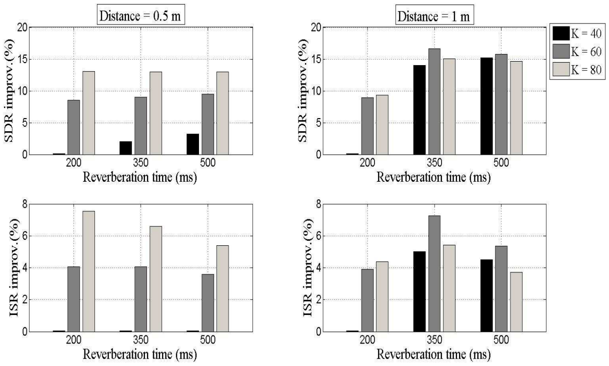

Fig. 2: Percentages of SDR and ISR improvements as a function of the reverberation times and the number of spectral bases K.

The amount of improvement is also shown in Fig. 2 as a function of the number of spectral bases K. For the proposed algorithm, we can observe that with a dense spectral basis matrix, where K is large, the separation performance is improved, when the distance is $0.5\ m$ and under low reverberant conditions. However, in case of high reverberant environments and distant talking, there is not a big difference in the percentage of improvement as a function of K.

5. CONCLUSION

This paper proposed a new framework of source separation using source-based prior information. Source power spectra of a set of training speech signals are decomposed by applying NTF, and a redundant library of spectral basis matrices, extracted in the decomposition, is built. Observing the mixture, the matched spectral basis matrices are detected using NTF. To perform the separation by following the local Gaussian modeling of the mixing process, the detected matrices are used to estimate a set of parameters of the model. Exploiting the redundancy of the observed mixture, the parameters are estimated in the ML sense using NTF. The contribution of each source signal in the mixture is extracted, applying the Wiener filtering process. The proposed algorithm was compared with our NMF based source separation algorithm and provided more flexibility and better separation performance.

6. REFERENCES

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[2] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 624–639, 2011.

[3] M. Parvaix and L. Girni, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE TASLP*, vol. 19, no. 6, pp. 1721–1733, 2011.

[4] S. Gorlow and S. Marchand, "Informed source separation: underdetermined source signal recovery from instantaneous stereo mixture," 2011, pp. 309–312.

[5] A. Liutkus, J. Pinel, R. Badeau, L. Girni, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[6] A. Liutkus, J. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *14th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2013, Paris, France, July 3-5, 2013*, 2013, pp. 1–4.

[7] M. Fakhry and F. Nesta, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," in *International Workshop on Acoustic Signal Enhancement, IWAENC 2012, Proceedings, Aachen, Germany, September 4th - 6th*, 2012, pp. 1–4.

[8] F. Nesta and M. Fakhry, "Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31*, 2013, pp. 86–90.

[9] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP J. Adv. Sig. Proc.*, p. 149, 2013.

[10] 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, 2012.

[11] M. Fakhry, P. Svaizer, and M. Omologo, "Reverberant audio source separation using partially pre-trained nonnegative matrix factorization," in *Proceedings of IWAENC 2014, Juan-les-Pins, France*, 2014, pp. 273 – 277.

[12] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral GMM estimation for underdetermined instantaneous audio source separation," in *ICA*, vol. 5441, 2009, pp. 751–758.

[13] A. Ozerov and C. Fevotte, "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, 2010.

[14] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *ICASSP2013*, 2013, pp. 141–145.

[15] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 293 – 297, 2015.

[16] A. Carmi, L. Mihaylova, and S. Godsill, *Compressed sensing and sparse filtering*, ser. Signals and Communication Technology. Springer, 2013.

[17] A. Cichocki, R. Zdunek, A. H. Phan, and S. ichi Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, 2009.

[18] Y. Mitsufuji and A. Roebel, "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization," *EURASIP J. Adv. Sig. Proc.*, 2014.

[19] C. Fevotte and J. F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian models," in *WASPAA*, 2005, pp. 78–81.

[20] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

[21] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[22] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined convolutive blind source separation using spatial covariance models," in *Proc. ICASSP2010, Dallas, TX*, 2010, pp. 9–12.