The Journal of Supercomputing (2021) 77:12143–12156
https://doi.org/10.1007/s11227-021-03771-y

# Parallel multichannel blind source separation using a spatial covariance model and nonnegative matrix factorization

**A. J. Muñoz-Montoro[2,3] · J. J. Carabias-Orti[1] · R. Cortina[3] · S. García-Galán[1] · J. Ranilla[3]**

Accepted: 22 March 2021 / Published online: 6 April 2021


## Abstract

In this paper, we present a multichannel nonnegative matrix factorization (MNMF) system for the task of source separation. We propose a novel signal model using spatial covariance matrices (SCM) where the mixing filter encodes the spatial information and the source variances are modeled using a NMF structure. Moreover, the proposed model is initialized with the estimated source direction of arrival (DoA) in order to mitigate the strong sensitivity to parameter initialization. The proposed system has been evaluated for the task of music source separation using a multichannel classical chamber music dataset showing that it is possible to reach real time in the tested scenarios by combining multi-core architectures with parallel and high-performance techniques.

**Keywords** Source separation · Multichannel NMF · Real time · Parallel computing

✉ A. J. Muñoz-Montoro
munozantonio@uniovi.es

J. J. Carabias-Orti
carabias@ujaen.es

R. Cortina
raquel@uniovi.es

S. García-Galán
sgalan@ujaen.es

J. Ranilla
ranilla@uniovi.es

[1] Department of Telecommunication Engineering, Universidad de Jaén, Jaén, Spain

[2] Escuela de Ciencias Técnicas e Ingeniería, Universidad a Distancia de Madrid (UDIMA), Madrid, Spain

[3] Department of Computer Science, University of Oviedo, Oviedo, Spain

## 1 Introduction

Source separation is a challenging task in the context of audio signal processing. Separating the sound sources of an audio mixture captured with one or multiple microphones can be useful for a great variety of subsequent audio processing tasks. Some examples of these tasks include spatial audio coding (SAC) [8, 25], music applications [6, 9, 11], 3D sound analysis and synthesis [18], localization [12] and signal enhancement for various purposes, such as automatic speech recognition (ASR) [17, 29].

Over the last two decades, the scientific community has dedicated many efforts to develop approaches that achieve this separation. Typical approaches rely on decomposing a time–frequency representation of the mixture signal using methods such as nonnegative matrix factorization (NMF), independent component analysis (ICA), or probabilistic latent component analysis (PLCA). Among these factorization techniques, NMF has been widely used for speech and music audio signals, as it allows to describe the signal as a nonsubstractive combination of sound objects (or "atoms" ) over time. However, without any prior information, the quality of the separation using the aforementioned statistical methods is limited. In fact, source separation methods can be classified based on the availability of prior information about the sources. Blind source separation (BSS) refers to the situation in which information about the specific sources of the mixture are unknown. On the contrary, informed source separation (ISS) refers to such methods in which information about the specific sources is used to improve the separation [15]. Many ISS approaches exploit the spectro-temporal properties of the sources. For example, spectral harmonicity and temporal continuity can be assumed for several musical instruments while percussive instruments are characterized by short bursts of broadband energy [2]. Speech source spectrogram can be modeled using a source-filter model [5]. Other approaches also used spatial localization of the sources [19, 32]. Recently, the deep neural networks (DNN) have been extensively used for this purpose. The existing methods mostly use DNN with either the spectrogram as the input signal representation [24] or directly the time-domain representation [4] to train such a system.

The aforementioned approaches are developed for single channel signals. In the case of multichannel mixtures, separation can be improved by taking into account the spatial locations of sources or the mixing process. Recent methods have extended NMF to the multichannel case by modeling the latent source magnitude- or power-spectrograms with NMF and estimating the mixing system without the nonnegativity constraint [22, 26]. This strategy is often referred to as MNMF in the literature. For modeling the spatial properties of the sources, many of these approaches use a spatial covariance matrix (SCM) which accounts to the relative inter-microphone phase and amplitude information of the recorded channels. Authors in [26] proposed to estimate unconstrained SCM mixing filters together with a NMF magnitude model to identify and separate repetitive frequency patterns corresponding to a single spatial location. To mitigate the effect of the spatial aliasing, Nikunen and Virtanen [22] proposed a SCM model

based on DoA kernels to estimate the inter-microphone time delay given a looking direction. Carabias et al. [3] proposed a SCM kernel based model where the mixing filter is decomposed into two direction-dependent SCMs to represent and estimate disjointly both time and level differences between array channels. Alternatively, recent works have tried to exploit multichannel audio with deep neural network (DNN) based approaches. Several works [21, 24, 28] combine DNN-based source spectrogram estimation with multichannel NMF-inspired spatial models.

The main drawback of these strategies is the large number of parameters which have to be estimated, and thus, without any prior information, these methods are prone to converge to local minima, especially in reverberant environments. Moreover, the computational burden of the MNMF-based approaches is heavy and current implementations do not allow a real-time performance. This is because many operations related to matrix inversions and eigenvalue decompositions are involved in the NMF updates. Under moderate echoic conditions, SCMs can be restricted to be rank-1 in a determined scenario, merging-independent vector analysis (IVA) and NMF within a framework called independent low-rank matrix analysis (ILRMA) [13]. Several studies have recently proposed restricting the SCMs of sources to jointly diagonalize the full-rank matrices for multichannel blind source separation [10, 27]. While FastMNMF [27] projects the signals with an optimizable transform matrix, the authors in [20] adopt a fixed projection, namely a discrete Fourier transform (DFT) matrix. In this work, we propose a projection-based multichannel source separation method using SCM and the MNMF algorithm. In particular, we propose a novel MNMF scheme that allows to perform the separation frame by frame. Similar to [23], we propose to initialize the model parameters using prior information from the sources DoA obtained with the Steered Response Power (SRP) with phase transform (PHAT) [30] algorithm in order to reduce the computational complexity and increase the robustness.

In this paper, we make the following technical contributions: (i) a projection-based SCM signal model for the task of multichannel source separation, (ii) an online system that outperforms the state-of-the-art system, and (iii) a novel prototype using a mixed parallelism scheme that allow to perform the separation in real time.

According to the best of our knowledge, there has not yet been presented a holistic, flexible and free system that addresses this problem on parallel shared-memory systems. As a proof of concept, several experiments have been performed using a multichannel dataset of classical chamber music with different polyphony level. The proposed approach has been compared with other online state-of-the-art method showing reliable results in terms of sound quality.

The paper is organized as follows. The introduction is presented in Sect. 1. Section 2 presents the problem formulation of MNMF. The proposed MNMF framework is described in Sect. 3. Section 4 presents the experimental results of the proposal. Finally, the conclusions are outlined in Sect. 5.

## 2 Problem formulation

The problem considered in this work is to separate each source signal from a set of audio mixtures recorded from a microphone array. The observed signal can be expressed as

$$x_m(n) = \sum_{s=1}^{S} \sum_{\tau} h_{ms}(\tau) y_s(n-\tau) \tag{1}$$

where the mixture $x_m(n)$ consists of $s \in [1, S]$ sources captured by microphones $m \in [1, M]$, and the time-domain sample index is denoted by $n$. The spatial response from source $s$ to microphone $m$ is represented by a mixing filter $h_{ms}(\tau)$ and the single-channel source signals are denoted by $y_s(n)$.

Considering the convolutive mixing problem in Eq. 1, the short-time Fourier transform (STFT) of $x_m(n)$ can be written as

$$\mathbf{x}_f(t) = \sum_{s=1}^{S} \mathbf{h}_{fs} y_{fs}(t) \tag{2}$$

where $\mathbf{x}_f(t) = [x_{f1}(t), \ldots, x_{fM}(t)]^{\mathrm{T}}$ is the time–frequency spectrograms of $x_m(n)$, $\mathbf{h}_{fs} = [h_{fs1}, \ldots, h_{fsM}]^{\mathrm{T}}$ denotes the frequency-domain mixing filter and $y_{fs}(t)$ represents the time–frequency spectrograms of source signals. Here, $f \in [1, F]$ and $t \in [1, T]$. As signal representation, the proposed method uses the spatial covariance matrix (SCM) domain [3, 22, 26]. This representation computes the phase and the amplitude difference between every pair of microphones in the multichannel mixture avoiding using the absolute phase of the observed signal.

To obtain the SCM, the magnitude square-rooted matrix $\hat{\mathbf{x}}_f(t)$ for a time–frequency point ($f$, $t$) of the captured signal at each microphone $\mathbf{x}_f(t) = [x_{f1}(t), \ldots, x_{fM}(t)]^T$ is firstly computed as

$$\hat{\mathbf{x}}_f(t) = [|\tilde{x}_{f1}(t)|^{1/2}\mathrm{sgn}(\tilde{x}_{f1}(t)), ..., |\tilde{x}_{fM}(t)|^{1/2}\mathrm{sgn}(\tilde{x}_{fM}(t))]^T, \tag{3}$$

where $\mathrm{sgn}(z) = z/|z|$ is the signum function for complex numbers. Then, the SCM for each single time–frequency point is defined from the multichannel captured vector $\hat{\mathbf{x}}_f(t)$ as the following outer product

$$\mathbf{X}_f(t) = \hat{x}_f(t)\hat{x}_f^H(t) = \begin{bmatrix} |x_{f1}(t)| & \cdots & x_{f1}(t)x_{fM}^*(t) \\ \vdots & \ddots & \vdots \\ x_{fM}(t)x_{f1}^*(t) & \cdots & |x_{fM}(t)| \end{bmatrix}, \tag{4}$$

where $^H$ stands for Hermitian transpose. Matrices $\mathbf{X}_f(t) \in \mathbb{C}^{M \times M}$ for each time–frequency point ($f$, $t$) encode the magnitude spectrum $|\mathbf{x}_f(t)| = [|x_{f1}(t)|, \ldots, |x_{fM}(t)|]^T$ in the main diagonal and the magnitude correlation and phase difference $|x_{fn}(t)x_{fm}(t)|^{1/2}\mathrm{sgn}(x_{fn}(t)x_{fm}^*(t))$ between each microphone pair ($n$, $m$) in the off-diagonal values.

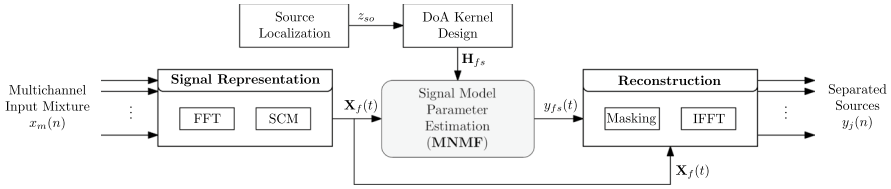


**Fig. 1** Block diagram of the proposed system

The convolutive mixing model in Eq. 2 can be expressed in terms of the SCM domain as

$$\mathbf{X}_f(t) \approx \tilde{\mathbf{X}}_f(t) = \sum_{s=1}^{S} \mathbf{H}_{fs} \bar{y}_{fs}(t) \tag{5}$$

where $\bar{y}_{fs}(t)$ denotes the magnitude spectrogram for each source $s$ and $\mathbf{H}_{fs} \in \mathbb{C}^{M \times M}$ is the SCM representation of the spatial frequency response $\mathbf{h}_{fs}$.

## 3 Proposed MNMF algorithm for multichannel source separation

In this work, we propose a multichannel source separation system based on the SCM domain and MNMF algorithm. In particular, we propose a practical and versatile framework that can perform the multichannel separation in real time. For this purpose, we have developed a beamforming inspired efficient and fast implementation that able to estimate the source variances using a projection based MNMF procedure where the source DoAs is estimated a priori. As a result, we have developed a software solution that satisfies two essential requirements: mobility and real time. Therefore, our design takes the low memory resources and low computational power of cheap and handheld devices into account. This has been possible using and deeply exploiting the possibilities offered by parallel architectures.

The block diagram of the proposed framework is depicted in Fig. 1. As can be observed, the full system combines different stages: (1) signal representation, (2) signal model parameter estimation, and (3) signal reconstruction. In the following subsections, we detail and describe the main function of each stage.

### 3.1 Proposed MNMF signal model

In this section, we introduce the signal model that enables to estimate the spectrogram of the sources using the MNMF algorithm [26]. Although the SCM mixing filter $\mathbf{H}_{fs}$ in Eq. 5 takes amplitude and phase differences between channels into consideration, it does not have any explicit relation to spatial locations. To overcome this, [22] proposed a beamforming-inspired SCM model based on DoA kernels. The main idea relies on decomposing the mixing filter $\mathbf{H}_{fs}$ as a linear combination of

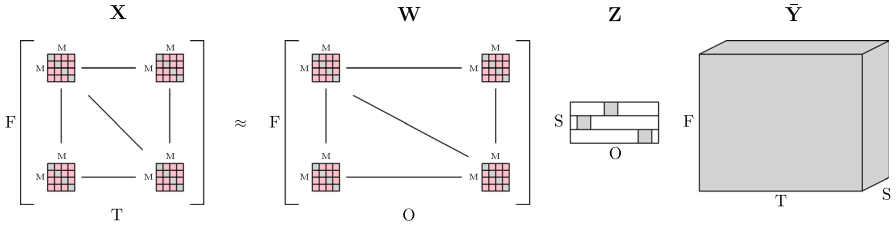


**Fig. 2** Proposed MNMF signal model parameters. Complex values are displayed in red, positive real values in gray and zero values in white

DoA kernels $\mathbf{W}_{fo} \in \mathbb{C}^{M \times M}$ multiplied by a spatial weights matrix $\mathbf{Z} \in \mathbb{R}_{+}^{S \times O}$ which relates sources $S$ with spatial directions $O$.

In this work, we propose a projection-based SCM model that enables to estimate directly the spectrogram of the sources using their spatial location as prior information. The proposed signal model for SCM observation is defined in Eq. 6 and illustrated in Fig. 2.

$$\mathbf{X}_f(t) \approx \hat{\mathbf{X}}_f(t) = \sum_{s=1}^{S} \underbrace{\sum_{o=1}^{O} \mathbf{W}_{fo} z_{so}}_{\mathbf{H}_{fs}} \bar{y}_{fs}(t) \tag{6}$$

Here, the spatial weights matrix $\mathbf{Z}$ is initialized a priori in order to reduce the number of free parameter. In this way, the SCM DoA kernel matrix $\mathbf{W}$ is computed for the source spatial positions as

$$\left[\mathbf{W}_{fo}\right]_{nm} = e^{j\theta_{nm}(f,o)} \tag{7}$$

where $\theta_{nm}(f, o) = 2\pi f \tau_{nm}(\mathbf{x}_o)$ is the phase difference computed from the TDoA between sensors $n$ and $m$ for the frequency in Hz at bin $f$ and the spatial position $o$. Remember that the number of spatial position $O$ is equal to the number of sources $S$ and is known a priori. Each spatial position $\mathbf{x}_o$ can be translated to a TDoA (in seconds) for a pair of microphones $(n, m)$ using the following expression

$$\tau_{nm}(\mathbf{x}_o) = \frac{||\mathbf{x}_o - \mathbf{x}_n||_2 - ||\mathbf{x}_o - \mathbf{x}_m||_2}{c} \tag{8}$$

where $||\cdot||_2$ denotes the $\ell_2$-norm, $\mathbf{x}_o$ is the source spatial position, $\mathbf{x}_m$ and $\mathbf{x}_n$ are the microphone $m$ and $n$ locations, all of them expressed in the Cartesian coordinate system, and $c$ is the speed of sound.

As in [26], after computing $\mathbf{W}_{fo}$, some post-processing is required to make it Hermitian and positive semidefinite. For the sake of brevity, this processing has been omitted here, regardless refer to [26] for more details. After that, $\mathbf{W}_{fo}$ is kept fixed during the factorization.

### 3.2 MNMF parameter estimation

For the estimation of the source magnitude spectrograms $\bar{\mathbf{Y}}$, we used the majorization-minimization algorithm proposed in [3, 22, 26]. Using this approach, the cost function can be described using both Euclidean and Itakura Saito (IS) divergence. In this work, we use the IS divergence, since it is better suited for audio modeling in comparison to EUC [7].

The IS divergence of the observed and estimated multichannel signal using the SCM domain can be expressed as

$$D_{IS}(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{ft} \text{tr}(\mathbf{X}_f(t)\tilde{\mathbf{X}}_f(t)^{-1}) - \log\det(\mathbf{X}_f(t)\tilde{\mathbf{X}}_f(t)^{-1}) - M \tag{9}$$

where $\text{tr}(\mathbf{X}) = \sum_{m=1}^{M} x_{mm}$ is the trace of a square matrix $\mathbf{X}$. Then, the source spectrogram can be obtained from the projection of the fixed averaged DoA-kernels over the observed SCM signal mixture using

$$\bar{y}_{fs}(t) \leftarrow \bar{y}_{fs}(t)\sqrt{\frac{\sum_{so} z_{so}\ tr(\hat{\mathbf{X}}_f(t)^{-1}\mathbf{X}_f(t)\hat{\mathbf{X}}_f(t)^{-1}\mathbf{W}_{fo})}{\sum_{so} z_{so}\ tr(\hat{\mathbf{X}}_f(t)^{-1}\mathbf{W}_{fo})}} \tag{10}$$

and repeating Eq. 6 followed by Eq. 10 until convergence. Further information about the derivation of Eq. 10 can be found in [23, 26].

Finally, once the source magnitude spectrograms are estimated, the spectrogram of each source can be computed using a soft-filter strategy.

### 3.3 Source reconstruction

The reconstruction of the source signals is performed using a generalized Wiener filtering strategy. Firstly, the estimated MNMF magnitude spectrogram for each sound source $s$ and microphone $m$ can be defined from our proposed model in Eq. 6 as

$$\breve{y}_{fsm}(t) = \sum_{o} tr(\mathbf{W}_{fo})_m z_{so}\bar{y}_{fs}(t) \tag{11}$$

Then, we apply the generalized Wiener mask to reconstitute different sources of the mixture based on the power spectrum ratio between the reference signals as

$$\tilde{y}_{fsm}(t) = \frac{\breve{y}_{fsm}(t)}{\sum_{os} tr(\mathbf{W}_{fo})_m z_{so} y_{fs}(t)} \cdot x_{fm}(t) \tag{12}$$

where $x_{fm}(t) \in \mathbb{C}$ is the time–frequency spectrogram of the input multichannel mixture (see Sect. 2). Finally, the multichannel time-domain signals are obtained by the inverse STFT of $\tilde{y}_{fsm}(t)$ and frames are combined by weighted overlap-add.

The procedure of the whole system is summarized in Algorithm 1.

**Algorithm 1** Pseudo Code of the Proposed MNMF system algorithm

```
Initialize Z with the position of the sources.
Initialize W using Eq. 7.
Apply post-processing to enforce W to be hermitian and semipositive definite.
while audio stream do
    Read an audio frame.
    Compute the input signal phase SCM using Eq. 4.
    Compute the signal model using 6.
    for iter = 1 to N_iter do
        Update Ȳ according to Eq. 10.
        Recompute the signal model using Eq. 6.
    end for
    for S = 1 to S do
        for m = 1 to M do
            Compute ỹ_fsm(t) using the Eq. 11.
            Compute the time-frequency spectrogram estimation of ỹ_fsm(t) using Eq. 12.
            Reconstruct the source signal using the inverse STFT of ỹ_fsm(t).
        end for
    end for
end while
```

## 4 Evaluation and experimental results

This section presents the experimentation carried out in the evaluation of the proposed system in Sect. 3. In this evaluation, we have exploited a subset of the University of Rochester Multimodal Music Performance (URMP) dataset presented in [14]. We have selected some classical chamber music pieces ranging from duets to quartets and played by 9 different common instruments in orchestra. Note that the musical score, the audio recordings of the individual tracks, the audio recordings of the assembled mixture and ground-truth annotation files are available for each piece. The multichannel mixtures were generated by simulating the spatial position of the sources. In this regard, mixing filters were simulated with the Roomsim Toolbox [1] for a rectangular room and a linear array of eight omnidirectional microphones. The reverberation time $RT_{60}$[1] of the room was set to either 10 *ms* or 400 *ms*.

Regarding the used testbed, we have focused our interest on two different systems. Firstly, we have used a server with an Intel® Xeon® Silver 4110 processor with 8 cores. It operates at 2.1 GHz and HyperThreading and Turbo Boost are both deactivated. Secondly, the experiments were conducted on a NVIDIA Jetson AGX Xavier development kit, which is an embedded system-on-chip (SoC) with an ARM v8.2 64-bit CPU. Xavier supports different kinds of running modes (configurable with the *NVPModel* command tool). This setup allows to simulate a wide range of mobile devices such as smartphones, laptops, tablets and other embedded systems under controlled conditions.

Concerning the used software, Xavier runs Ubuntu Linux 18.04.1 LTS and the server runs CentOS Linux 7. Both systems use the OpenBlas[2] library (release 0.3.9,

[1] $RT_{60}$ is the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

[2] https://www.openblas.net

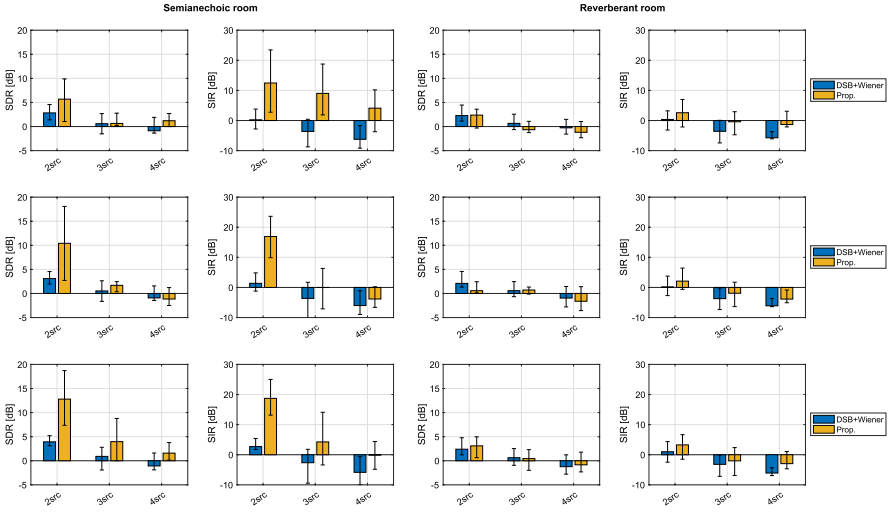


**Fig. 3** Objective results using the BSS_EVAL metrics [31] for the proposed dataset. Results for two channels are displayed in the upper row, for three channels in the middle row and for four channels in the lower row. Each bar indicates the median values of the obtained results

March 2020), the FFTW[3] library (release 3.3.8, May 2018) and the GNU C Compiler 7 with the specification 4.5 of OpenMP. OpenBLAS is an optimized BLAS library based on GotoBLAS2 1.13 BSD. Note that both packages have been built in our system from source codes. Finally, it should be remarked that the used data type is "*double*" (i.e., IEEE 754 double-precision binary floating-point format).

### 4.1 Results

Firstly, we have tested the reliability of our separation system in terms of sound quality by using the BSS_Eval toolbox [31]. These metrics are commonly accepted and represent a standard approach in the specialized scientific community for testing the quality of separated signals, allowing a fair comparison with other methods. In this paper, we compare the separation performance of our beamforming-inspired proposal with a well-known spatial beamforming [30] method from the literature. In particular, we have implemented a Delay and Sum Beamforming (DSB) design which consists of time aligning and summing the microphone signals. This technique uses the geometrical information of the microphone array to filter and enhance the sources coming from a specific direction. To allow a fair comparison with our NMF-based approach, a postprocessing Wiener filtering stage is applied to the output of DSB [16] as well.

[3] http://www.fftw.org

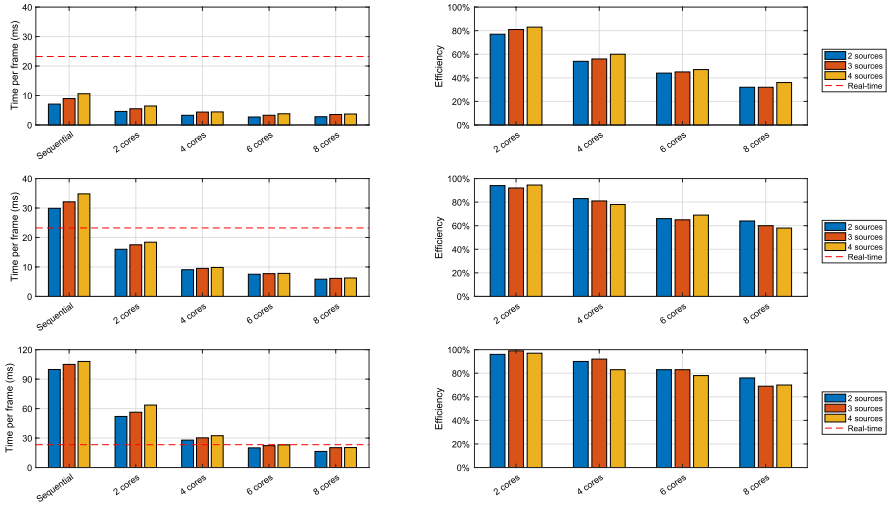


**Fig. 4** Execution times measured in milliseconds per frame and efficiency on the Intel® Xeon® Silver 4110. Results for two channels are displayed in the upper row, for three channels in the middle row and for four channels in the lower row

Figure 3 depicts the median values of SDR and SIR obtained in the evaluation of the proposed database for each approach. We start by analyzing the values obtained in the semianechoic room. As can be seen, the proposed framework provides superior results in terms of SDR and SIR compared with the DSB+Wiener method for all cases. As expected, the separation performance decreases as the number of sources increases. On the other hand, better results are obtained when the number of channels increase, since the DoA estimation performed by SRP-PHAT algorithm is more accurate and, therefore, the initialization of the source location is more reliable.

Concerning the reverberant room, a similar performance can be observed. The proposed framework slightly outperforms the DSB+Wiener method for most of the cases. However, under higher reverberant conditions, the results dramatically decrease due to localization errors. Note that despite the simplicity of the DSB+Wiener algorithm (which allows its implementation in real time), the method suffers from the leakage of other sources into the extracted source resulting in a poor interference-related metrics (SIR) with respect to the proposed method.

Secondly, we have explored the limits of our proposed system. In this second experiment, we have measured the complexity per frame and the efficiency of the algorithm as a function of the number of computing cores used simultaneously, the number of sources of the mixture and the number of channels of the input signal. The results obtained using an Intel® Xeon® Silver 4110 are depicted in Fig. 4.

As can be observed, the computational complexity of the algorithm increases as the number of channels and sources increases. For two-channels mixtures, real time is achieved regardless of the number of simultaneous computing cores used, including the sequential version. Note that real time is guaranteed when the execution time per frame is lower than 23.2 *ms* (displayed as a dashed red line). However,

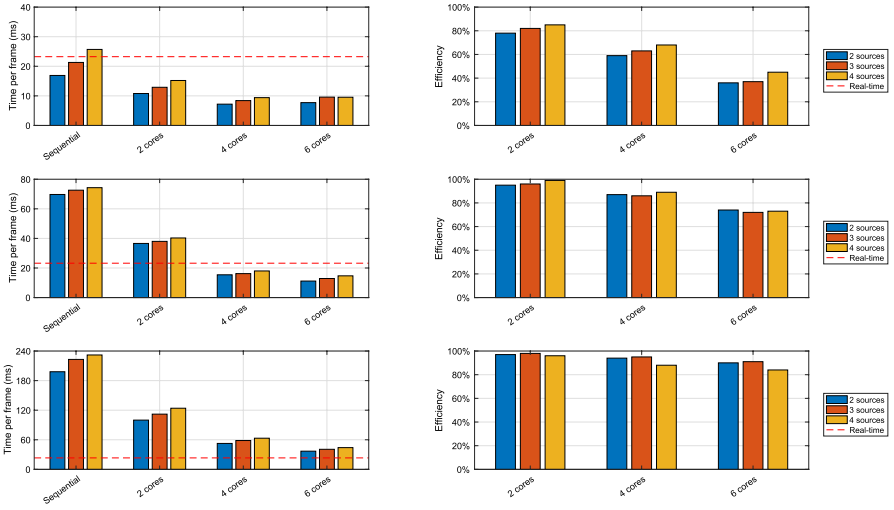


**Fig. 5** Execution times measured in milliseconds per frame and efficiency on the NVIDIA Jetson AGX Xavier. Results for two channels are displayed in the upper row, for three channels in the middle row and for four channels in the lower row

as observed in Fig. 3, the separation results increase considerably as a function of the number of channels. In the cases of three and four channels, the sequential version of the algorithm does not run in real time. In the three-channels case, parallel computing allows execution times in real time regardless of the number of sources and computing cores. Finally, real time is only achieved in four-channel mixtures for six and eight cores.

Regarding the efficiency, again better results are obtained when the number of channels increases. Note that some memory-bound operations of the system are performed as matrix–vector, such as the Wiener filter. Therefore, the sequential approach maximizes the performance, taking advantage of the whole memory bandwidth, while a parallel approach is limited by this fact.

As for the NVIDIA Jetson AGX Xavier, the experimental results obtained in the evaluation are provided in Fig. 5. In this case, real time is not reached for four-channels mixtures in any case. For the three-channels case, the separation is performed in less than 23.2 *ms* when four and six cores are used. Finally, note that the results obtained for the efficiency are always above 75% for three and four channels.

## 5 Conclusion

In this paper, we proposed a projection-based multichannel source separation method using SCM and the MNMF algorithm. In particular, we proposed a novel MNMF prototype using a mixed parallelism scheme that allow to perform the separation frame-by-frame in real time. The proposed signal model uses SCM to encodes the spatial information and the source variances are modeled using a NMF structure.

Moreover, the signal model is initialized with the estimated source DoA in order to mitigate the strong sensitivity to parameter initialization.

The proposed framework has been implemented for multi-core architectures allowing that the application can be executed in a wide range of devices. Furthermore, we have shown the robustness of the proposed algorithm in comparison with other state-of-the-art method using various types of microphone array setups. Results showed that real time is reached in most of the cases. To our best knowledge, our proposal is the first MNMF implementation in real time that obtains reliable results in terms of sound quality.

**Acknowledgements** This work was supported by the Regional Ministry of the Principality of Asturias under grant *FC-GRUPIN-IDI/2018/000226*, by the Ministry of Economy, Knowledge and University of the Government of the "Junta de Andalucía" under project *P18-RT-1994*, by the "Programa Operativo FEDER Andalucía 2014-2020" under project with reference *1257914*, and by Pre-doctoral Fellowship Program from the "Ministerio de Ciencia, Innovación y Universidades" of Spain under the reference BES-2016-078512.

## References

1. Campbell DR, Palomaki KJ, Brown G (2005) A MATLAB simulation of "shoebox'' room acoustics for use in research and teaching. Comput Inf Syst 9:48–51
2. Canadas-Quesada F, Fitzgerald D, Vera-Candeas P, Ruiz-Reyes N (2017) Harmonic-percussive sound separation using rhythmic information from non-negative matrix factorization in single-channel music recordings. DAFx 2017 - Proceedings of the 20th International Conference on Digital Audio Effects (i), 276–282
3. Carabias-Orti JJ, Nikunen J, Virtanen T, Vera-Candeas P (2018) Multichannel blind Sound source separation using spatial covariance model With level and time Differences and nonnegative matrix factorization. IEEE/ACM Trans Audio Speech Lang Process 26(9):1512–1527. https://doi.org/10.1109/TASLP.2018.2830105
4. Défossez A, Bach F, Usunier N, Bottou L (2019) Music source separation in the waveform domain (2019)
5. Durrieu JL, Richard G, David B, Fevotte C (2010) Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Trans Audio Speech Lang Process 18(3):564–575. https://doi.org/10.1109/TASL.2010.2041114
6. Ewert S, Muller M (2011) Estimating note intensities in music recordings. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 385–388. IEEE. https://doi.org/10.1109/ICASSP.2011.5946421
7. Févotte C, Bertin N, Durrieu JL (2009) Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput 21(3):793–830. https://doi.org/10.1162/neco.2008.04-08-771
8. Herre J, Falch C, Mahne D, Del Galdo G, Kallinger M, Thiergart O (2010) Interactive teleconferencing combining spatial Audio Object Coding and DirAC technology. In: 128th Audio Engineering Society Convention 2010, vol. 3, pp. 1579–1590
9. Huang PS, Chen SD, Smaragdis P, Hasegawa-Johnson M (2012) Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 57–60
10. Ito N, Nakatani T (2019) FastMNMF: Joint Diagonalization Based Accelerated Algorithms for Multichannel Nonnegative Matrix Factorization. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. https://doi.org/10.1109/ICASSP.2019.8682291
11. Itoyama K, Goto M, Komatani K, Ogata T, Okuno HG (2008) Instrument equalizer for query-by-example retrieval: improving sound source separation based on Integrated harmonic and Inharmonic Models. Ismir. https://doi.org/10.1136/bmj.324.7341.827

12. Jensen JR, Christensen MG, Jensen SH (2013) Nonlinear least squares methods for joint DOA and pitch estimation. IEEE Trans audio Speech Lang Process 21(5):923–933. https://doi.org/10.1109/TASL.2013.2239290
13. Kitamura D, Ono N, Sawada H, Kameoka H, Saruwatari H (2016) Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE/ACM Trans Audio Speech Lang Process 24(9):1626–1641. https://doi.org/10.1109/TASLP.2016.2577880
14. Li B, Liu X, Dinesh K, Duan Z, Sharma G (2019) Creating a multitrack classical music performance dataset for multimodal music analysis: challenges, insights, and applications. IEEE Trans Multimedia 21(2):522–535. https://doi.org/10.1109/TMM.2018.2856090
15. Liutkus A, Durrieu JL, Daudet L, Richard G (2013) An overview of informed audio source separation. In: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4. IEEE. https://doi.org/10.1109/WIAMIS.2013.6616139
16. Marro C, Mahieux Y, Simmer K (1998) Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. IEEE Trans Speech Audio Process 6(3):240–259. https://doi.org/10.1109/89.668818
17. McDonough J, Kumatani K (2012) Microphone Arrays. Techniques for Noise Robustness in Automatic Speech Recognition. Wiley, Chichester, UK, pp 109–157. https://doi.org/10.1002/9781118392683.ch6
18. Merimaa J, Pulkki V (2005) Spatial impulse response rendering I: analysis and synthesis. AES J Audio Eng Soc 53(12):1115–1127
19. Mitsufuji Y, Roebel A (2013) Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 71–75. IEEE. https://doi.org/10.1109/ICASSP.2013.6637611
20. Mitsufuji Y, Uhlich S, Takamune N, Kitamura D, Koyama S, Saruwatari H (2020) Multichannel non-negative matrix factorization using nanded spatial covariance matrices in wavenumber domain. IEEE/ACM Trans Audio Speech Lang Process 28:49–60. https://doi.org/10.1109/TASLP.2019.2948770
21. Munoz-Montoro AJ, Politis A, Drossos K, Carabias-Orti JJ (2020) Multichannel Singing Voice Separation by Deep Neural Network Informed DOA Constrained CMNMF. In: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6. IEEE. https://doi.org/10.1109/MMSP48831.2020.9287068
22. Nikunen J, Virtanen T (2014) Direction of arrival based spatial covariance model for blind sound source separation. IEEE/ACM Trans Audio Speech Lang Process 22(3):727–739. https://doi.org/10.1109/TASLP.2014.2303576
23. Nikunen J, Virtanen T (2014) Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 6677–6681. IEEE. https://doi.org/10.1109/ICASSP.2014.6854892
24. Nugraha AA, Liutkus A, Vincent E (2016) Multichannel audio source separation with deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 24(9):1652–1664. https://doi.org/10.1109/TASLP.2016.2580946
25. Pulkki V (2007) Spatial sound reproduction with directional audio coding. AES: J Audio Eng Soc 55(6):503–516
26. Sawada H, Kameoka H, Araki S, Ueda N (2013) Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans Audio Speech Lang Process 21(5):971–982. https://doi.org/10.1109/TASL.2013.2239990
27. Sekiguchi K, Bando Y, Nugraha AA, Yoshii K, Kawahara T (2020) Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. IEEE/ACM Trans Audio Speech Lang Process 28:2610–2625. https://doi.org/10.1109/TASLP.2020.3019181
28. Sekiguchi K, Nugraha AA, Bando Y, Yoshii K (2019) Fast Multichannel Source Separation Based on Jointly Diagonalizable Spatial Covariance Matrices. In: 2019 27th European Signal Processing Conference (EUSIPCO), pp. 1–5. IEEE. https://doi.org/10.23919/EUSIPCO.2019.8902557
29. Smaragdis P (2012) Extraction of Speech from mixture signals. Techniques for noise robustness in automatic speech recognition. Wiley, Chichester, UK, pp 87–108. https://doi.org/10.1002/9781118392683.ch5
30. Tashev IJ (2009) Sound capture and processing. Wiley, Chichester, UK. https://doi.org/10.1002/9780470994443

31. Vincent E, Gribonval R, Fevotte C (2006) Performance measurement in blind audio source separation. IEEE Trans Audio Speech Lang Process 14(4):1462–1469. https://doi.org/10.1109/TSA.2005.858005
32. Wang L, Ding H, Yin F (2010) Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals. EURASIP J Audio Speech Process 2010(1):1–13. https://doi.org/10.1155/2010/797962