IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2021, accepted January 30, 2022, date of publication February 8, 2022, date of current version February 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150248

Multichannel Blind Music Source Separation Using Directivity-Aware MNMF With Harmonicity Constraints

ANTONIO J. MUÑOZ-MONTORO[1], **JULIO J. CARABIAS-ORTI**[2], **PABLO CABAÑAS-MOLERO**[2], **FRANCISCO J. CAÑADAS-QUESADA**[2], **AND NICOLÁS RUIZ-REYES**[2]

[1]Computer Science Department, University of Oviedo, Gijón, 33203 Asturias, Spain
[2]Telecommunication Engineering Department, University of Jaén, Linares, 23700 Jaén, Spain

Corresponding author: Antonio J. Muñoz-Montoro (munozantonio@uniovi.es)

This work was supported in part by the Regional Government of Andalucia under the Framework Programme PAIDI 2020, Proyectos de I+D+i del Plan Andaluz de Investigacion, Desarrollo e Innovacion, through Grant P18-TPJ-4864; and in part by the Ministerio de Ciencia e Innovacion, Spain, under Projects PID2020-119082RB-{C21,C22}.

ABSTRACT In this paper we present a harmonic constrained Multichannel Non-Negative Matrix Factorization (MNMF) method for the task of blind music source separation. In this model, the mixing filter encodes the spatial information in terms of magnitude and phase differences between channels whereas the source variances are modelled using a harmonic constrained NMF structure. In this work, the spatial covariance matrix is obtained from the constant-Q transform to account for the frequency logarithmic scale inherent in music signals and reduce the dimensionality of the parameters. Moreover, to mitigate the strong sensitivity to parameter initialization, we propose to initialize the spatial weights with the output of the steered response power (SRP) with the phase transform (PHAT) algorithm. The proposed method has been evaluated for the task of music source separation using a multichannel classical chamber music dataset with several polyphony and reverberation setups. Furthermore, a comparison with other state-of-the-art signal decomposition methods has been accomplished showing reliable results in terms of BSS_EVAL metrics.

INDEX TERMS Constant-Q transform, harmonicity, multichannel NMF, music source separation.

I. INTRODUCTION

Audio source separation aims to segregate constituent sound sources from an audio signal mixture. This task has been one of the most popular research problems in the music information retrieval community. Since most of the music audio is available in the form of mixtures, there are several applications of a system capable of music source separation – e.g. automatic creation of karaoke, acoustic emphasis, music transcription, music unmixing and remixing, music production and education purposes.

Many approaches have been addressed in the last two decades in order to achieve this separation. A typical approach consists of decomposing a time-frequency representation of the mixture signal using methods such as non-negative matrix factorization (NMF), independent component analysis (ICA), or probabilistic latent component analysis (PLCA). Among these factorization techniques, NMF has been widely used for music audio signals, as it allows to describe the signal as a non-subtractive combination of sound objects (or ''atoms'') over time. However, without further information, the quality of the separation using the aforementioned statistical methods is limited. One solution is to 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 [1]. Speech source spectrogram can be modelled using a source-filter model [2]. Other approaches also used spatial localization of the sources [3], [4].

When training material is available, it is possible to learn the spectro-temporal patterns and the methods are referred to as supervised. In this way, several signal decomposition methods have been presented which provide superior results

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

to the blind scenario. Recently, 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 [5], [6] or directly the time-domain representation [7], [8] to train such a system. Convolutional neural networks (CNN) [6], [9] and long short term memory (LSTM) [10], [11] networks are the popular choices for DNN model architectures adapted for music source separation. Some of the latest top-performing music source separation models are Open-Unmix [11], MMDenseLSTM [12], Demucs [13] and Meta TasNet [14].

The aforementioned approaches are developed for single channel mixtures. When multichannel signals are available, separation can be improved by taking into account the spatial locations of sources or the mixing process. Earlier NMF based approaches relied on stacking magnitude or power spectrograms of all channels into a 3-valence non-negative tensor and decomposing it with non-negative tensor factorisation (NTF) methods [15] or other NTF-like nonnegative structured approximations [16], [17]. However, since the phase information is discarded, these approaches do not allow exploiting the interchannel phase differences (IPDs), but only the interchannel level differences (ILDs). ILD-based approaches perform well in close-miking scenarios [4], [18]. However, in the far-field case (i.e. when the microphone array size is much smaller than the distances between the sources and microphones) the ILDs are practically negligible and, therefore spatial information can only be exploited using IPDs. Multichannel non-negative matrix factorization (MNMF) based approaches model the latent source magnitude- or power-spectrograms with NMF while the spatial mixing system is modelled using a Gaussian probabilistic modelling applied directly to the complex-valued STFTs of all channels [19]–[21]. The spatial properties of the sources can be modelled using a spatial covariance matrix (SCM) which encodes magnitude and phase differences between the recorded channels. Authors in [19] proposed to estimate unconstrained SCM mixing filters together with an 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 [20] proposed an SCM model based on direction of arrival (DoA) kernels to estimate the inter-microphone time delay given a looking direction. Carabias *et al.* [22] proposed an 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. The main drawback of these strategies is the large number of parameters that have to be tuned and thus, without any prior information, these methods are prone to converge to local minima, especially in reverberant environments. Prior information about the localization can be used to reduce the computational cost and the strong sensitivity to parameter initialization [23]. Alternatively, several studies have recently proposed to restrict the SCMs of sources to jointly diagonalize the full-rank matrices for multichannel blind source separation [21], [24]. Similarly, [25] uses a discrete Fourier transform (DFT) matrix as the diagonalizer under the DoA kernel-based model from [20] projecting the signals into the wavenumber domain [25]. Recent works have tried to exploit multichannel audio with deep neural network (DNN) based approaches. Deep-clustering methods are augmented with spatial information in [26] with large improvements over monophonic versions. Alternatively, several works [5], [27] combine DNN-based source spectrogram estimation with multichannel NMF-inspired spatial models. Finally, a fully spatial-spectral factorization DNN is proposed as deep tensor factorization in [28].

In this paper, we present a blind music source separation approach based on MNMF where the signal model is constrained to be harmonic. To account for the inherent logarithmic frequency scale in western music, we propose to use the Constant-Q transform (CQT) [29] as a time-frequency representation. Although CQT is complex-valued and the ILDs and IPDs are encoded similarly to the DFT, existing approaches in the literature only used the amplitude information [30]–[32]. In fact, to our best knowledge, this is the first work that exploited the phase information of the CQT within an MNMF scheme. Using CQT for music source separation is beneficial for three reasons: 1) The frequency scale can be adjusted to a semitone level with a lower dimensionality than STFT representation, 2) The IPDs can be used as spatial cues to enhance the separation results in far-field scenarios, 3) Perfect reconstruction of the time-domain signals from CQT representation is possible [33]. Similar to [23], we propose a prior localization scheme using the Steered response power (SRP) with phase transform (PHAT) [34] algorithm to initialize the model parameters and thus, reduce the computational complexity and increase the robustness w.r.t. the parameter initialization. Several experiments have been performed using a multichannel dataset of classical chamber music with different polyphony levels and reverberant conditions. Moreover, the proposed approach has been compared with other state-of-the-art signal decomposition based approaches showing better results in terms of BSS_EVAL metrics.

The paper is organized as follows. The introduction is presented in Section I. Section II introduces the problem formulation and briefly reviews the background of MNMF, the harmonic signal models and the CQT transform. The proposed harmonic constrained MNMF method is explained in Section III. Section IV presents the experimental results of the proposed methods in comparison with other state-of-the-art methods for the task of multichannel music source separation. Finally, the conclusions are presented in Section V.

II. BACKGROUND

A. PROBLEM SPECIFICATION

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

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

where the mixture $x_m(l)$ consists of $s \in [1, S]$ sources captured by microphones $m \in [1, M]$, and the time-domain sample index is denoted by l. 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 $y_s(l)$.

Assuming the additive of complex spectra, the mixing model in Eq. (1) can be expressed in the frequency domain as:

$$\mathbf{x}_{ft} = \sum_{s=1}^{S} \underbrace{\mathbf{h}_{fs} y_{sft}}_{\mathbf{y}_{sft}} \tag{2}$$

where $\mathbf{x}_{ft} = [x_{1,ft}, \ldots, x_{M,ft}]^T \in \mathbb{C}^M$ is the observed multichannel mixture spectrogram. $f \in [1, F]$ and $t \in [1, T]$ represent both, the frequency and time frame indexes, respectively. $\mathbf{y}_{sft} = [y_{1,sft}, \ldots, y_{M,sft}]^T \in \mathbb{C}^M$ represents the image of source s which is obtained as the product of the single-channel spectrogram for source s, $y_{sft} \in \mathbb{C}$, and the frequency-domain mixing filter $\mathbf{h}_{fs} = [h_{1,fs}, \ldots, h_{Mfs}]^T \in \mathbb{C}^M$.

B. MULTICHANNEL NMF

Multichannel NMF can be formulated based on a so-called local Gaussian model (LGM), which allows modelling and combining spatial and spectral cues in a systematic way. LGM modelling [35] assumes that each source image (M-length complex-valued vector $[y_{1,sft}, \ldots, y_{M,sft}]^T \in \mathbb{C}^M$) is modelled as a zero-mean circular complex Gaussian random vector as follows

$$\mathbf{y}_{sft} \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{H}_{sft} v_{sft}), \tag{3}$$

where the complex-valued covariance matrix is positive definite Hermitian, and it is composed of two factors: 1) a spatial covariance $\mathbf{H}_{sft} \in \mathbb{C}^{M \times M}$ representing the spatial characteristics of the sth source image at the TF point (f, t), and 2) a spectral variance $v_{stf} \in \mathbb{R}$ representing the spectral characteristics of the sth source image at the TF point (f, t). These source variances v_{sft} can be modelled using a classical NMF structure as

$$v_{sft} = \sum_{k=1}^{K_s} b_{skf} g_{skt} \tag{4}$$

where b_{skf} and g_{skt} represent both the basis functions and their corresponding time-varying gains for each source-dependent component $k \in [1, K_s]$.

In the case of static sources, it is reasonable to assume that the spatial covariances are time-invariant, i.e., $\mathbf{H}_{sft} = \mathbf{H}_{sf}$ [17], [35]. In addition, assuming the random vectors $\mathbf{y}_{sft}$ to be independent in time, frequency and between sources, the mixture STFT coefficients in the multichannel mixing in Eq. (3) may be shown distributed as

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_{s=1}^{S} \mathbf{H}_{sf} \sum_{k=1}^{K_s} b_{skf} g_{skt}\right), \tag{5}$$

where $\mathbf{H}_{sf}$ could be modelled as a rank-1 SCM or as a full-rank matrix. Note that the full-rank spatial model can be used even in an under-determined condition $(M < S)$ unlike the rank-1 model (only valid for $(M \geq S)$), but its parameter estimation is harder (and computationally demanding) because of the considerably larger number of parameters.

C. BEAMFORMING INSPIRED DoA-SCM MODEL

When dealing with spectrally similar sources (e.g. several speech signals), without further constraints, NMF-based approaches can lead to the situation where a single NMF component together with the corresponding SCM mixing filter represent multiple sources together at different spatial locations. To enforce SCMs at different frequencies to correspond to the same location, Nikunen and Virtanen [20] proposed to model the spatial covariance as a weighted sum of so-called DoA kernels which are rank-1 spatial covariances modelling plane waves coming from several predefined directions.

$$\mathbf{H}_{kf} = \sum_{o=1}^{O} \mathbf{W}_{fo} z_{ko} \tag{6}$$

where the DoA kernels $\mathbf{W}_{fo} \in \mathbb{C}^{M \times M}$ define the phase differences between every pair of microphones (n, m) for each direction indexed with $o \in [1, O]$ and $z_{ko} \in \mathbb{R}^{\geq 0}$ is the spatial weights matrix that relates NMF components with spatial directions. Note that the spatial weights z_{ko} are frequency independent and estimated directly during the factorization whereas the DoA kernels $\mathbf{W}_{fo}$ are computed beforehand and the phase information is kept fixed. In particular, the DoA kernels are defined as

$$\left[\mathbf{W}_{fo}\right]_{nm} = \mathtt{exp}(j2\pi f \tau_{nm}(\mathbf{k}_o)) \tag{7}$$

where $f = (i-1)F_s/F$ is the frequency in Hz, F_s and F are the sampling frequency and the STFT length, respectively. $\tau_{nm}(\mathbf{k}_o)$ represents the time difference of arrival (TDoA) in seconds between the pair of microphone (n, m) and $\mathbf{k}_o$ is a unit vector pointing towards the look direction from the geometrical center of the microphone array (i.e. the origin of the Cartesian coordinate system $\mathbf{p} = [0, 0, 0]^T$). In the original method in [20], a posterior grouping strategy is required to relate components to sources.

D. HARMONIC NMF-BASED SOURCE MODEL

NMF modelling of each source consists in structuring the source variances v_{sft} in Eq. (4) with NMF structure as in the single-channel NMF case. When dealing with musical instrument sounds, ideally, each basis function can represent a single pitch and the corresponding gains contain information about the onset and offset times of notes having that pitch. Several works [16], [36]–[41] proposed to restrict the source variances in Eq. (4) to be harmonic. The harmonicity constraint is particularly useful for the analysis and separation of musical audio signals since, by using this constraint, each basis can define a single fundamental frequency.

In [38], [39], the basis functions b_{skf} in the model of Eq. (4) are defined as a weighted combination of $n \in [1, N]$ narrowband harmonic spectra (patterns) $p_{knf} \in \mathbb{R}$, which are arbitrarily fixed

$$b_{skf} = \sum_{n=1}^{N} e_{skn} p_{knf}. \tag{8}$$

Each component k is associated with a single pitch (with its corresponding f_0) and the basis functions are obtained as a linear combination of harmonic patterns p_{knf} with different shapes but sharing the same pitch which is fixed. Those harmonic patterns are weighted by a set of coefficients e_{snk} which define the actual spectral representation for component k belonging to source s.

Some approaches [41], [42] use an extension of the model in Eq. (8) using a single flat harmonic excitation where the basis functions for each note and instrument are reduced to a set of coefficients that define the shape of the harmonic spectral pattern:

$$b_{skf} = \sum_{n=1}^{N} a_{skn} \omega(f - nf_0(k)), \tag{9}$$

where each component k represents a single music note and the number of components K_s is usually set to cover the whole dynamic range for source s; $n \in [1, N]$ is the number of harmonics; a_{skn} is the amplitude of harmonic n for note k and instrument s; $f_0(k)$ is the fundamental frequency of note k; $\omega(f)$ is the magnitude spectrum of the window function; and the spectrum of a harmonic component at frequency $nf_0(k)$ is approximated by $\omega(f - nf_0(k))$.

Although the excitation-filter (or source-filter) model has origins in speech processing and sound synthesis, a similar model can be extrapolated to musical instruments [36], [43]. In fact, each instrument can be represented using a single filter λ_{sf} that corresponds to the resonant structure of the body of the instrument whereas the excitations can be represented as frequency components of unity magnitude at integer multiples of a certain fundamental frequency.

$$b_{skf} = \lambda_{sf} \sum_{n=1}^{N} \omega(f - nf_0(k)) \tag{10}$$

Some extensions for this source filter for music signals can be found in the literature. For instance, in [44] instead of defining an excitation for every possible pitch, they are given from a multipitch estimator. Additionally, the filter λ_{sf} is represented as a linear combination of fixed elementary responses. In particular, the authors chose the elementary responses to consist of triangular bandpass magnitude responses uniformly spaced on a Mel frequency scale. Finally, in [40], the authors proposed to decompose the single flat excitation by a combination of a few excitation patterns to better model the changes in the timbre between notes across frequency.

E. THE CONSTANT-Q TRANSFORM

Choosing the proper signal representation for a specific task is not trivial, as there are many aspects to be considered. Depending on the used representation, some signal properties may be hidden or revealed. The advantageous features for a given task are more outstanding under certain representations. Moreover, in some circumstances, a specific representation can provide a priori information related to the sources.

Several signal representations have been used in the literature to adjust the time-frequency resolution to the characteristics of the signals to be analyzed such as ERB [45], MFCC [44], CQT [33] or chroma vectors [46].

In the field of music signal processing, CQT has been widely used with logarithmic spaced frequency bands that can be approximated to have a resolution of a multiple/submultiple of a musical semitone. In fact, CQT computes frequency coefficients similar to DFT but on a logarithmic frequency scale [29] as,

$$x^Q(f, t) = \sum_{j=t-\lfloor|*|\rfloor T_f/2}^{t+\lfloor|*|\rfloor T_f/2} x(j) u_f^*(j - t + T_f/2) \tag{11}$$

where $f \in [1, F]$ indexes the frequency bins of CQT and $u_f^*(t)$ denotes the complex conjugate of $u_k(t)$ and T_f are variable window lengths. The notation $\lfloor| \cdot |\rfloor$ infers rounding down towards the nearest integer. The basis functions $u_f(t)$ are complex-valued waveforms, also called time-frequency atoms, and are defined by

$$u_f(t) = \frac{1}{C} \omega \left(\frac{t}{T_f} \right) e^{i2\pi t \frac{f_c}{f_s}} \tag{12}$$

where f_c is the center frequency of bin f, f_s denotes the sampling rate, and $\omega(t)$ is a continuous window function (e.g. the Hann window) sampled at points determined by t/T_f. The scaling factor C is given by,

$$C = \sum_{r=-\lfloor|*|\rfloor T_f/2}^{\lfloor|*|\rfloor T_f/2} \omega \left(\frac{r + T_f/2}{T_f} \right) \tag{13}$$

In order to have a bin spacing corresponding to an identical frequency ratio for every pair of adjacent notes, the center frequencies f_c obey the following expression,

$$f_c = f_1 \, 2^{\frac{f-1}{B}} \tag{14}$$

where f_1 is the center frequency of the lowest-frequency bin and B determines the number of bins per octave. B is the most crucial parameter in CQT, because it determines the time-frequency resolution tradeoff. The window lengths T_f

are inversely proportional to f_c in order to have the same Q-factor for all f bins and are given by,

$$T_f = \frac{f_s}{f_c(2^{1/B} - 1)} \tag{15}$$

Consequently, CQT can be understood as computing DFT only for specific logarithmic spaced frequency bins. However, it is not invertible and is far less efficient than FFT, which supposes an important drawback. Some approaches have been focused on improving the efficiency and making it quasi-invertible [33], [47] or perfectly invertible [48] under certain implementation constraints.

III. PROPOSED DIRECTIONAL MNMF HARMONIC MODEL FOR BSS

In this work, a harmonic multi-excitation model based on SCM and MNMF for blind music source separation is presented. Specifically, we propose a signal model suitable for modelling the harmonic structure of musical instruments. To account for the western music logarithmically spaced frequency bands we adopted the CQT as signal representation.

In our proposal the mixing filter is decomposed as a linear combination of spatial kernels as in [20] and the source magnitude spectrograms as weighted combinations of harmonic constrained basis functions. In this work, we assume far-field (i.e., the microphone array size is much smaller than the distances between the sources and microphones). In this scenario, spatial information is obtained from the phase difference between the complex-valued CQT representation computed for each microphone. As commented in Section I, this is the first MNMF approach exploiting the phase information of the CQT. To alleviate the sensitivity of the system towards the parameter initialization, we propose to use prior information about the sources localization using a well-known source localization method.

The block diagram of the proposed method is depicted in Fig. 1. First, the SCM representation is computed from the CQT of the multichannel mixture. Second, the model parameters are estimated using two steps: 1) Initialization of the spatial weights using the SRP-PHAT algorithm [34], 2) Estimation of the source magnitude spectrogram using MNMF. Finally, a generalized Wiener filtering strategy is used to obtain the source reconstruction.

A. CQT-SCM SIGNAL REPRESENTATION

In this work, we used an SCM signal representation obtained from the complex-valued CQT transform (see Section II-E). From the original formulation of the CQT in Eq. (11), the overlap factor between successive window functions in the time domain is constant while window lengths T_f are decreasing with increasing f. Consequently, the number of time frames varies as a function of the frequency bin. A way to overcome this issue is to use a "rasterized" CQT representation using a fixed hop size for all the center frequencies, that is, the smaller hop size in the representation. As a drawback, this procedure produces a highly redundant representation. To mitigate this problem, we propose to downsample the rasterized CQT representation to a fixed hop size of 32 ms for all the frequency bins. This downsampled representation will be used until the estimation of the Wiener-like masks (see Fig. 1). Then, an interpolation process will be carried out to obtain the inverse transform.

For each frequency bin f and time frame t, the magnitude square-rooted matrix CQT transform $\mathbf{x}^Q_{ft}$ of the captured signal at each sensor $\mathbf{x}^Q_{ft} = [x^Q_{ft1}, \ldots, x^Q_{ftM}]^T$ is given by

$$\hat{\mathbf{x}}_{ft} = [|x^Q_{ft1}|^{1/2} sgn(x^Q_{ft1}); \ldots ; |x^Q_{ftM}|^{1/2} sgn(x^Q_{ftM})], \tag{16}$$

where $sgn(z) = z/|z|$ is the signum function for complex numbers. Then, the CQT-SCM $\mathbf{X}_{ft}$ for a single time-frequency point (f, t) is defined from the multichannel captured signal $\hat{\mathbf{x}}_{ft}$ as the outer product

$$\mathbf{X}^Q_{ft} = \hat{\mathbf{x}}_{ft}\hat{\mathbf{x}}^H_{ft} = \begin{bmatrix} \left|x^Q_{ft1}\right| & \cdots & x^Q_{ft1} {x^Q}^*_{ftM} \\ \vdots & \ddots & \vdots \\ x^Q_{ftM} {x^Q}^*_{ft1} & \cdots & \left|x^Q_{ftM}\right| \end{bmatrix}, \tag{17}$$

where ^{H} and * stand for Hermitian transpose and complex valued conjugate, respectively. With the above definitions, the magnitude spectrum of each channel is at the diagonal of $\mathbf{X}^Q_{ft}$, while the spatial properties of the mixture are restricted to its off-diagonal values, which encode the square magnitude cross correlation and phase difference between each microphone pair. Note that the SCM representation is independent of the absolute phase of the recorded signals.

B. PROPOSED MULTICHANNEL HARMONIC MULTI-EXCITATION MODEL

The SCM model in Eq. (5) allows estimating the source magnitude spectrograms v_{sft} and the corresponding SCM mixing filter $\mathbf{H}_{fs}$ yielding the desired BSS properties. Although the SCM mixing filter $\mathbf{H}_{fs}$ considers the phase and amplitude differences between channels, estimating it jointly over all frequencies does not provide any explicit relation to the spatial location of the sources. Following the beamforming-inspired SCM mixing model in [22] and inspired in the original model in [20], [23] (see Eq. (6)), the SCM mixing filter $\mathbf{H}_{fs}$ is modelled as a linear combination of 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. The SCM mixing filter is described as

$$\mathbf{H}_{sf} = \sum_{o=1}^{O} \mathbf{W}_{fo} z_{so} \tag{18}$$

Note that the direct relation between sources to directions in Eq. (18) avoids the need for the grouping strategy after the factorization.

Regarding the source magnitude spectrogram v_{sft}, in this work we propose to restrict the source basis functions to be harmonic. In fact, musical notes, excluding transients,

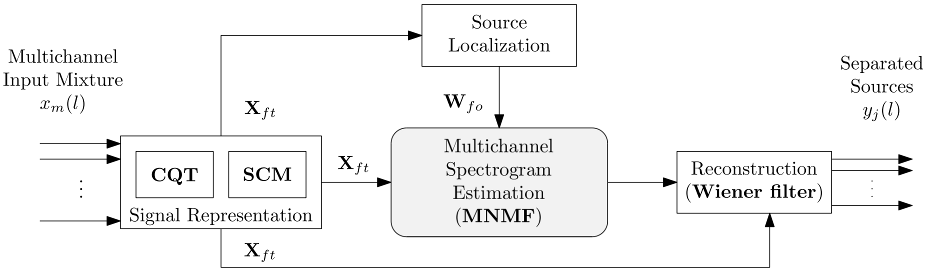

Multichannel Input Mixture $x_m(l)$
CQT
SCM
Signal Representation
$\mathbf{X}_{ft}$
Source Localization
$\mathbf{W}_{fo}$
$\mathbf{X}_{ft}$
Multichannel Spectrogram Estimation (MNMF)
Reconstruction (Wiener filter)
Separated Sources $y_j(l)$
$\mathbf{X}_{ft}$

FIGURE 1. Block diagram of the proposed system.

are pseudo-periodic, and their spectra consist of regularly spaced frequency peaks. In this way, we propose to model the magnitude time-frequency spectrogram v_{sft} in Eq. (5) of each harmonic instrument (or source) s in frame t and frequency f as a weighted sum of spectral bases with harmonic shape as

$$v_{sft} = \sum_{k=1}^{K} g_{kts} \sum_{n=1}^{N} a_{ksn}\, \omega^Q(f - nf_0(k)) \tag{19}$$

where g_{kts} denotes the time-varying activation of each musical note (i.e. the gains for each component). $n \in [1, N]$ is the number of harmonics, $f_0(k)$ is the fundamental frequency of note k and a_{ksn} is the amplitude of the partial (or harmonic) n corresponding to the note k and instrument s. $\omega^Q(f - nf_0(k))$ is the magnitude spectrum of the window function in the CQT domain of the n-th harmonic component at frequency $nf_0(k)$.

The proposed CQT-SCM model is obtained by combining the DoA kernel based SCM mixing filter from Eq. (6) with the harmonic source spectrogram model from Eq. (19). Thus, the proposed signal model is given by

$$\tilde{\mathbf{X}}^Q_{ft} = \sum_{s=1}^{S}\sum_{o=1}^{O} \mathbf{W}_{fo}\, z_{so} \sum_{k=1}^{K} g_{kts} \sum_{n=1}^{N} a_{ksn}\, \omega^Q(f - nf_0(k)) \tag{20}$$

C. SPATIAL WEIGHTS INITIALIZATION

For the sake of reducing the algorithm complexity and increasing the robustness w.r.t. to the parameter initialization, we constraint the spatial weights $\mathbf{Z}$ using prior information about the sources DoAs obtained using the well-known SRP-PHAT algorithm. Similar to [23], the DoA estimation is computed from the STFT of the multichannel signal. For each time frame, the maximum of the SRP function is scaled to one. Assuming stationary sources, the source directions can be estimated by averaging the SRP functions over time. In this work, we select all the peaks with SRP higher than 75% of the SRP value of the highest peak and we also impose a minimum distance of 10 degrees between peaks. Note that this scheme allows reducing the number of directions to be analyzed during the decomposition process (i.e., to reduce the size of $\mathbf{Z}$), while allowing to relate several locations o to a single source s. Thus, a reverberant condition can be modelled as a weighted combination of anechoic responses similar to the scheme proposed in [20].

D. PARAMETER ESTIMATION

For the estimation of the signal model parameters in (20), we propose to use the majorization-minimization algorithm presented in [19], [20], [22]. Using this approach, the cost function can be described using both Euclidean (EUC) and Itakura Saito (IS) divergence. However, in this work, we use the IS divergence since it is better suited for audio modelling in comparison to EUC [49]. Following a similar approach to [19], we obtain the multiplicative updates (MU) via auxiliary functions for the case of the IS divergence. Other methods in [20], [23] derived the expectation-maximization (EM) algorithms for the IS divergence. However, as demonstrated in [19], MU updates provide faster convergence than the EM algorithms.

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

$$D_{IS}(\mathbf{X}^Q, \tilde{\mathbf{X}}^Q) = \sum_{ft} \mathrm{tr}(\mathbf{X}^Q_{ft}\tilde{\mathbf{X}}^{Q^{-1}}_{ft}) - \log\det(\mathbf{X}^Q_{ft}\tilde{\mathbf{X}}^{Q^{-1}}_{ft}) - M \tag{21}$$

where $\mathrm{tr}(\mathbf{X}) = \sum_{m=1}^{M} x_{mm}$ is the trace of a square matrix $\mathbf{X}$. Note that Eq. (21) reduces to IS NMF when $M = 1$. For a given observation $\mathbf{X}^Q$, the cost function in (21) together with the proposed model in (20) can be written as

$$f(\mathbf{Z}, \mathbf{G}, \mathbf{A}) = \sum_{ft}\left[\mathrm{tr}\left(\mathbf{X}^Q_{ft}\tilde{\mathbf{X}}^{Q^{-1}}_{ft}\right) - \log\det\left(\tilde{\mathbf{X}}^Q_{ft}\right)\right] \tag{22}$$

where constant terms are omitted. To minimize this complex-valued function $f(\mathbf{Z}, \mathbf{G}, \mathbf{A})$, we follow the optimization scheme of majorization in [19] using an auxiliary positive definite function f^+ which allows the factorization under non-negativity of the parameters restriction. Then, the derivation of the algorithm updates is achieved via partial derivation of function f^+ w.r.t. each model parameter and setting these derivatives to zero. For the sake of brevity, we write directly the multiplicative update rules for each free parameter. Further information about the derivation of these rules can be found in [19], [22].

Similar to [25], in this work, we assume far-field and therefore, we assume the DOA kernels $\mathbf{W}_{fo}$ to be fixed during the factorization. Therefore, from the proposed signal model in Eq. (20) only the spatial weights z_{so} and the source variance parameters, a_{ksn} and g_{kts} are the free parameters to be optimized. The update rules for the non-negative parameters in (20) are

$$z_{so} \leftarrow z_{so} \sqrt{\frac{\sum_{ftkn} g_{kts}\, a_{ksn}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{-1} \mathbf{X}_{ft}^{Q} \tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}{\sum_{ftkn} g_{kts}\, a_{ksn}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}} \tag{23}$$

$$g_{kts} \leftarrow g_{kts} \sqrt{\frac{\sum_{fno} z_{so}\, a_{ksn}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{-1} \mathbf{X}_{ft}^{Q} \tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}{\sum_{fno} z_{so}\, a_{ksn}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}} \tag{24}$$

$$a_{ksn} \leftarrow a_{ksn} \sqrt{\frac{\sum_{fto} z_{so}\, g_{kts}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{X}_{ft}^{Q} \tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}{\sum_{fto} z_{so}\, g_{kts}\, \omega(f - nf_0(k))\, \mathrm{tr}\left(\tilde{\mathbf{X}}_{ft}^{Q^{-1}} \mathbf{W}_{fo}\right)}} \tag{25}$$

In addition, scaling the parameters is required for eliminating trivial scale indeterminacies and avoids numerical instabilities. Then, the scaling procedure is presented as follows,

$$z_{so} \leftarrow \frac{z_{so}}{\sqrt{\sum_o z_{so}^2}}, \quad g_{kts} \leftarrow \frac{g_{kts}}{\sqrt{\sum_{ks} g_{kts}^2}}, \quad a_{ksn} \leftarrow \frac{a_{ksn}}{\sqrt{\sum_n a_{ksn}^2}} \tag{26}$$

Note that this scaling procedure is carried out after the updates of z_{so}, g_{kts} and a_{ksn}.

E. SOURCE RECONSTRUCTION

The source signals are reconstructed using the model parameters previously estimated. To ensure that the reconstruction process is conservative, we propose to follow a generalized Wiener filtering strategy. The generalized Wiener masks represent the relative energy contribution of each source with respect to the energy of the mixture. The estimated CQT magnitude spectrogram for each sound source s and microphone m can be defined from our proposed model in Eq. 20 as

$$\breve{y}_{ftsm}^{Q} = \sum_o tr(\mathbf{W}_{fo})_m\, z_{so} \sum_{k=1}^{K} g_{kts} \sum_{n=1}^{N} a_{ksn}\, \omega^{Q}(f - nf_0(k)) \tag{27}$$

Then, we apply the generalized Wiener mask to reconstitute each source from the mixture based on the power spectrum ratio between the reference signals in the CQT domain as

$$\tilde{y}_{ftsm}^{Q} = \frac{\breve{y}_{ftsm}^{Q}}{\sum_{so} tr(\mathbf{W}_{fo})_m\, z_{so}\, \sum_{k=1}^{K} g_{kts}\, \sum_{n=1}^{N} a_{ksn}\, \omega^{Q}(f - nf_0(k))}\, x_{ftm}^{Q} \tag{28}$$

where $x_{ftm}^{Q} \in \mathbb{C}$ is the time-frequency CQT transform of the input multichannel mixture. Finally, the multichannel time-domain signals are obtained by the inverse CQT transform [33] of $\tilde{y}_{ftsm}^{Q}$.

The whole proposed MNMF algorithm for music source separation is summarized in Algorithm 1.

Algorithm 1 Pseudo-Code of the Proposed Harmonic MNMF System Algorithm

```
Compute the CQT transform of the multichannel input signal.
Compute the signal CQT-SCM observation by using Eq. (17).
Initialize $z_{so}$ by using the SRP-PHAT algorithm.
Compute the signal model by using Eq. (20).
while not convergence and iter $\leq$ no. of iters do
    Update $z_{so}$ according to Eq. (23).
    Recompute the signal model by using Eq. (20).
    Update $g_{kts}$ according to Eq. (24).
    Recompute the signal model by using Eq. (20).
    Update $a_{ksn}$ according to Eq. (25).
    Scale $z_{so}$, $g_{kts}$ and $a_{ksn}$ to $\ell_2$-norm as specified in Eq. (26).
    Recompute the signal model by using Eq. (20).
end while
Compute the CQT magnitude spectrograms $\breve{y}_{ftsm}^{Q}$ by using the Eq. (27).
Compute the reference signals $\tilde{y}_{ftsm}^{Q}$ in the CQT domain by using Eq. (28).
Reconstruct the source signal by using the inverse CQT transform of $\tilde{y}_{ftsm}^{Q}$.
```

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the proposed method in Section III is evaluated for the task of multichannel instrumental music source separation using a popular dataset of small ensembles. Moreover, the performance of our framework has been compared to other state-of-the-art algorithms to demonstrate the reliability of our proposal.

A. DATASETS

In this study, we have evaluated our method using the University of Rochester Multimodal Music Performance (URMP) dataset presented in [50]. This dataset contains single-channel audio recordings and ground-truth annotations for 44 classical chamber music pieces ranging from duets to quintets (11 duets, 12 trios, 14 quartets, and 7 quintets) and played by 14 different common instruments in orchestra. For each piece, the musical score in MIDI format and the high-quality individual instrument audio recordings are provided.

The multichannel mixtures were generated by simulating the spatial position of the sources. In this regard, mixing filters were simulated with the Roomsim Toolbox [51] for a rectangular room of dimensions 3.55 m × 4.45 m × 2.5 m and a linear array of eight omnidirectional microphones. The reverberation time RT_{60}[1] of the room was set to 65 ms, 250 ms and 500 ms, and the inter-microphone distance to 5 cm. The distance between the sources and the geometrical center of the array was fixed to 2 m and the source directions of arrival varied between 0° and 180° with a minimal spacing of 30°. The overview of the recording configuration and room layout is illustrated in Fig. 2.

The anechoic material from the URMP dataset was convolved with the obtained IRs resulting in two-channels, four-channels and eight-channels datasets for the three RT_{60} combinations. Therefore, this process provided nine different setups for each set of audio signals. Note that each audio excerpt has a duration of 30 seconds and is sampled at 44.1 kHz.

B. EXPERIMENTAL SETUP

In this paper, the time-frequency representation is obtained from CQT using 12 bins per octave. Regarding the harmonic decomposition scheme, the number of harmonic components per note N and the number of notes per instrument K are set to 20 and 115, respectively. Note that the entire range of MIDI notes possible for any instrument has to be covered, i.e. from the MIDI note 21 to the MIDI note 135.

As for the mixing filter decomposition, the number of look directions is equal to the number of sources that compound the audio mixture (see Section III-B). Note that only azimuthal angles are considered. Finally, the maximum number of iterations for the parameters estimation loop is set to 300, since we have observed that this value is adequate for the convergence of the reconstruction error.

C. EVALUATION METRICS

In this work, we have objectively evaluated the performance of the separation method by comparing each separated signal to the spatial images of the original sources and using objective measures by BSS_Eval toolbox [52]. 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 state-of-the-art methods. These metrics assume that each separated signal produces a distortion model that can be expressed as:

$$\hat{s}_j(l) - s_j(l) = e_j^{target}(l) + e_j^{interf}(l) + e_j^{artif}(l) \qquad (29)$$

where $\hat{\mathbf{s}}_j$ is the estimated source signal for instrument j, $\mathbf{s}_j$ is the original signal of the instrument j, $\mathbf{e}_j^{target}$ is the error term associated with the target distortion component, $\mathbf{e}_j^{interf}$ is the error term due to interference of the other sources, and $\mathbf{e}_j^{artif}$ is the error term attributed to the numerical artifacts of the separation algorithm. In this way, BSS_EVAL provides the following metrics based on energy ratios for each separated signal: the *source to distortion ratio* (SDR), the *source to interference ratio* (SIR), the *source to artifacts ratio* (SAR), and the *source image spatial distortion ratio* (ISR) [52].

D. ALGORITHMS FOR COMPARISON

In this subsection, we present the state-of-the-art algorithms used for comparing the separation performance of our proposal. Note that we also include two *unrealistic* baseline methods to show the extreme separation performances, here denoted as oracle mask separation and energy distribution. The different approaches compared here are the following:

1) ORACLE MASK SEPARATION

This method computes the optimal value of the Wiener mask at each frequency and time component using the downsampled CQT representation and assuming that the signals to be separated are known in advance. Therefore, this approach represents the upper bound for the best separation that can be reached with the used time-frequency representation.

2) ENERGY DISTRIBUTION (ED)

This procedure uses the mixture signal divided by the number of sources as input for the evaluation. This evaluation provides a starting point for the separation algorithms.

3) FASST

In our evaluation we have included the results obtained by the multichannel method presented in [45]. This method decomposes the magnitude spectrogram of the mixture signal into a sum of basis spectra representing individual pitches scaled by time-varying amplitudes. In this approach, each basis spectrum is modelled as a weighted sum of narrowband spectra representing a few adjacent harmonic partials, thus enforcing harmonicity and spectral smoothness while adapting the spectral envelope to each instrument. The author used a generalized expectation–maximization (GEM) algorithm to estimate the model parameters.

4) MULTICHANNEL NMF

The approach presented in [17] decomposes the multichannel audio spectrogram using NMF with the Itakura Saito divergence. The method considers two variants, instantaneous and convolutive mixing that are compared here and referred to as *Mult. NMF inst.* and *Mult. NMF conv.*, respectively. In order to estimate the mixing and source parameters, we have used the implementation provided by the authors using the multiplicative update (MU) algorithm.

5) ILRMA

In this approach, the independent frequency vectors in Independent Vector Analysis (IVA) are extended to low-rank matrices, which correspond to the power spectrograms

[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.

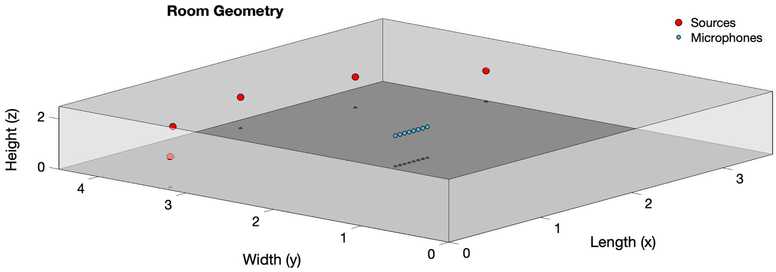


FIGURE 2. Room, source positions and microphone array placement.

of estimated sources, using NMF decomposition [53]. This signal model (independence between sources and low-rank decomposition of source spectrograms) is theoretically equivalent to conventional MNMF only when the spatial covariance matrix of each source in MNMF is constrained to a rank-1 matrix, which yields a computationally efficient algorithm for so-callen independent low-rank matrix analysis (ILRMA). Note that ILRMA is applicable to the determined case ($M = S$). In the overdetermined case ($M > S$), dimensionality reduction using principal component analysis (PCA) should be applied so that $M = S$. Therefore, only those signals for the dataset that satisfy the condition $M \geq S$ have been considered for the evaluation of this method.

6) HARMONIC NTF

We have included in the evaluation results for the MNMF-based method that relies only on the amplitude information of the multichannel signal. In this case, only the magnitude spectrograms are considered and the phase information of the multichannel signal is discarded. This signal model follows the same harmonic structure described in Eq. (9) and includes a panning matrix within the model that models the contribution of each source to each channel of the input signal as in [18], [54].

7) DSB+Wiener

We have also reported results for a spatial beamforming [34] method. Specifically, 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 form a specific direction. To allow a fair comparison with NMF-based approaches, a postprocessing Wiener filtering stage is applied to the output of DSB [55].

8) DoA-MNMF

We have included as the baseline for our experiments the results of the beamforming-inspired SCM model in [20] using the implementation provided by the authors.

E. VARIANTS OF THE PROPOSED METHOD

We have also presented results of two variants of our own model. In that sense, we have been considered both the CQT and STFT signal representations when comparing the models in order to know the influence of using these two different representations. Thus, we have the following two scenarios:

- *Proposal-STFT*: This variant refers to our proposed signal model presented in Eq. 20 where the SCMs are obtained from the STFT of the multichannel mixture.
- *Proposal-CQT*: This variant refers to our proposed signal model presented in Eq. 20 where the SCMs are obtained from the CQT of the multichannel mixture (see Section III-A).

F. RESULTS

We start by analyzing the performance of the method in a semi-anechoic scenario. Fig. 3 shows the separation results averaged over all audio excerpts for the room with $\mathrm{RT}_{60} = 65$ ms. Each column corresponds to a different array size (2, 4 and 8 microphones). Note that the *DoA-MNMF* and both proposed variants are evaluated using two different initializations for the spatial weights (ground-truth source DoA represented in light color vs SRP-PHAT estimation represented in dark color).

The best results are obtained with the *Oracle mask separation* method (SDR $\approx$ 11.5 dB, SIR $\approx$ 17 dB, SAR $\approx$ 12.5 dB, ISR $\approx$ 17 dB). This measure informs us about the best separation that can be achieved using the selected Wiener mask with the downsampled CQT representation. Moreover, the results show that the proposed system achieves significantly better average SDR and ISR than the other compared methods. Observe that only the $SB + Wiener$, $DoA - MNMF$ and both proposed variants are able to surpass the SDR score obtained by *ED* (i.e. the mixture signal divided by the number of sources), which is a clear indication of the difficulty of the task at hand.

Concerning our variants, *Proposal-CQT* achieves better results than *Proposal-STFT* regardless of the score and the number of channels used. This is due to the robustness of the CQT representation against not perfectly tuned notes of the instruments composing the music excerpts (see Section II-E), especially string instruments. In comparison to the *DoA-MNMF* method, our *Proposal-CQT* obtains a clearly higher SDR result (2.8 dB vs 1.1 dB with 2 channels, 3 dB vs 1.2 dB with 4 channels and 3.1 dB vs 1.8 dB with 8 channels), while maintaining a similar level of artifacts and interferences. This results indicate that the harmonic constrain of our model improves the separation of harmonic instruments in

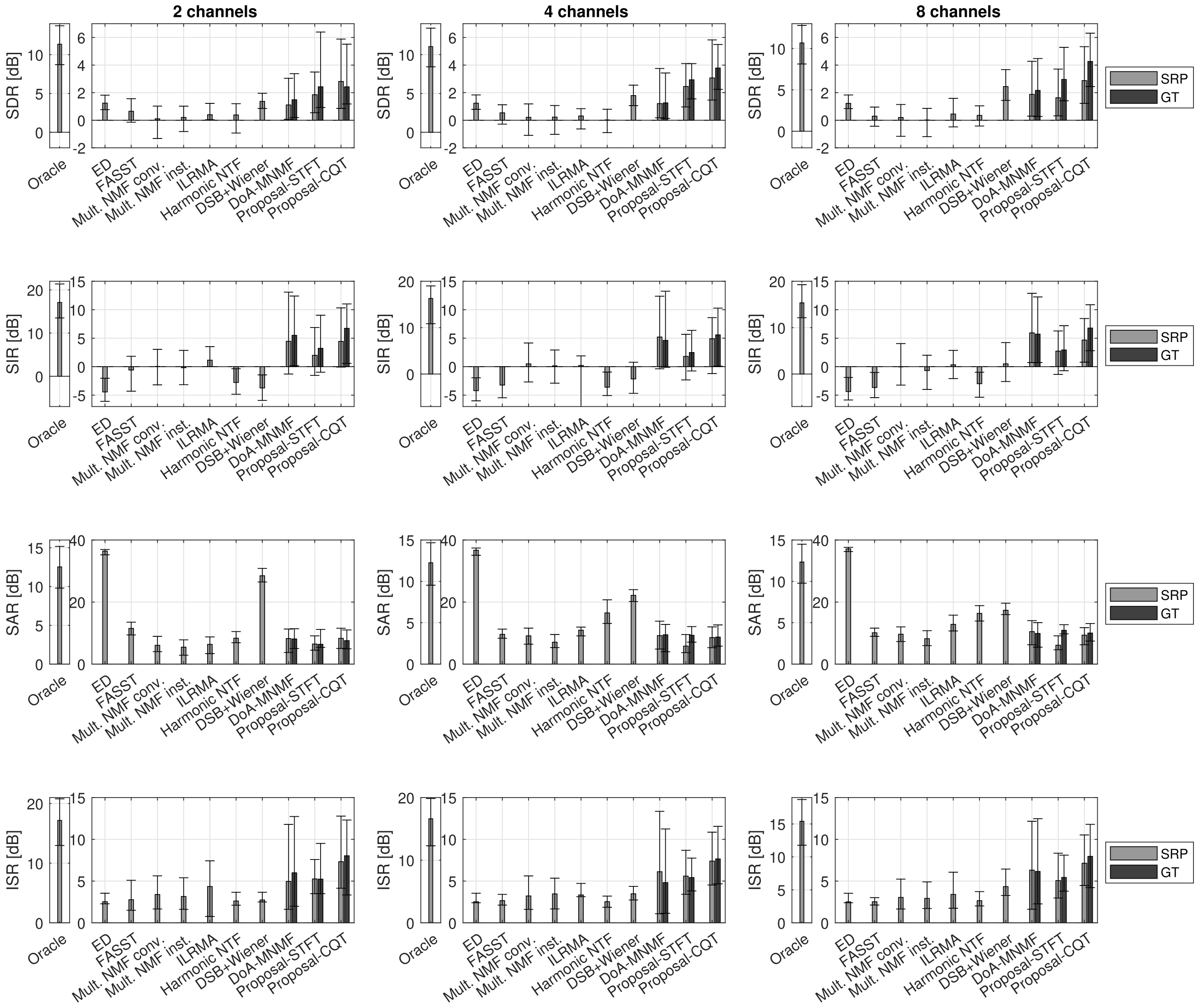


FIGURE 3. Source separation metrics averaged over the URMP dataset [50] under semi-anechoic conditions (RT_{60} = 65 ms) with the simulated room. Each column corresponds to a different array size (2, 4 and 8 microphones). Method abbreviated as (GT) uses ground-truth source DoA (represented in light color) whereas (SRP) uses SRP-PHAT estimation (represented in dark color). The error bars represent 95% confidence intervals.

classical music mixtures, and does not introduce additional artifacts. It is also worth noting that our model has fewer parameters due to the use of the harmonicity constraint, which may be helping in the convergence to a better solution. Both the *DoA-MNMF* and the proposed variants perform similarly well when initialized with ground-truth and SRP-PHAT source locations. Consequently, the method can be successfully used in a blind fashion in combination with SRP-PHAT.

As expected, the *DSB + Wiener* approach outperforms all the other methods in terms of added artifacts, as demonstrated by the SAR score. However, this approach fails to provide a meaningful isolation of the sound sources, resulting in a poor SIR value. Compared to the remaining decomposition methods, our approach performs much better in terms of SDR and SIR metrics, while offering a comparable SAR score. Although *FASST* and *Mult. NMF conv.* also make use of a convolutive mixing model, they suffer to discriminate the sources because the phase differences are not constrained, and the amplitude differences are small when the microphones are close to each other. Phase constraining is a clear advantage of the *DoA-MNMF* and both proposed variants, in which the beamforming-based MNMF model imposes a small set of directions of arrival to each source and benefits from the spatial diversity of the sources. *ILRMA*, which uses an instantaneous mixing model, performs better than the other decomposition methods in terms of SDR and SIR with 8 channels, and offers a better SAR metric than our method. This may be attributed to the more strict mathematical constraints of our model, which provides much better isolation at the expense of slightly higher artifacts. Also, note that *ILRMA* is determined, so only in

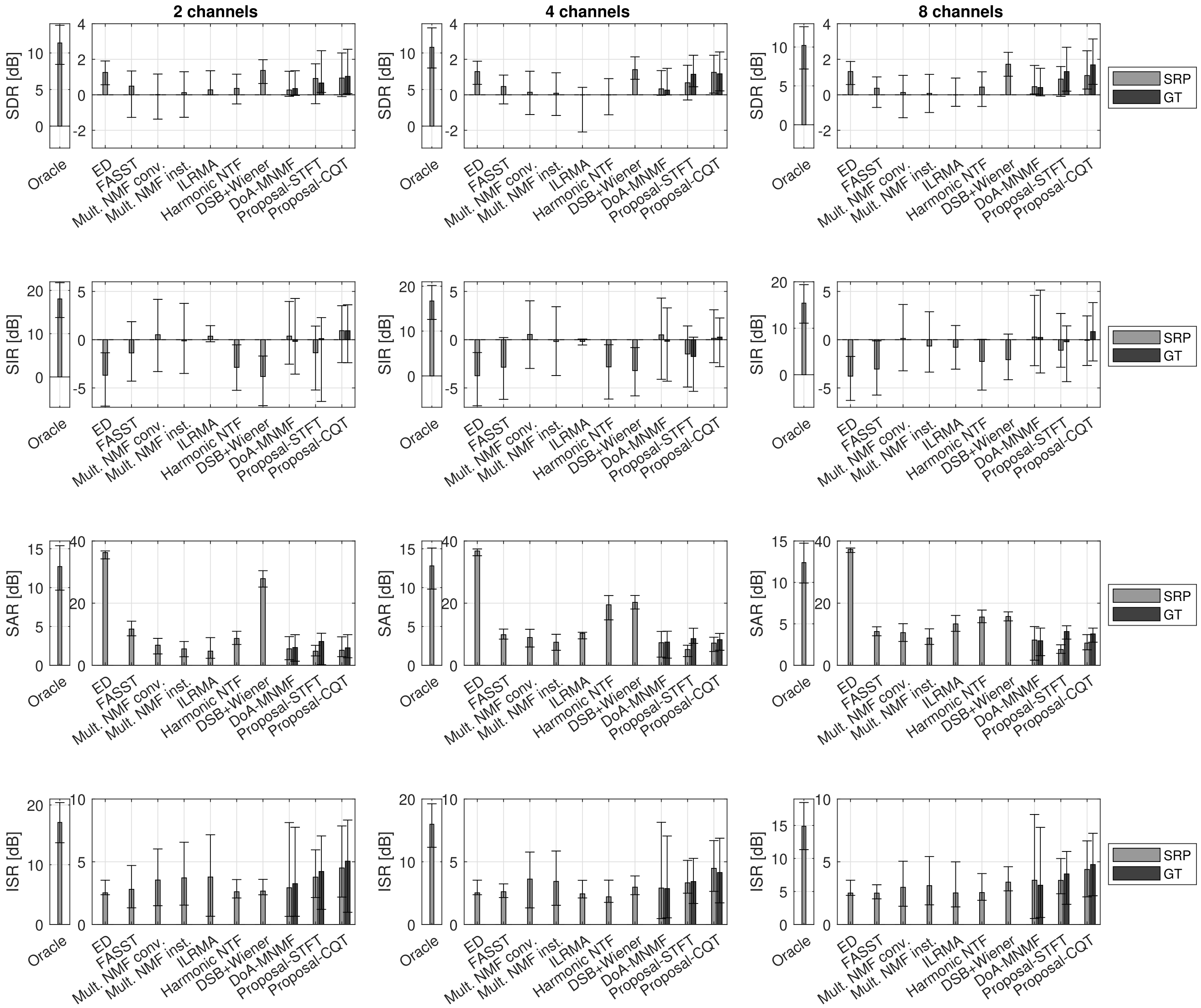


FIGURE 4. Source separation metrics averaged over the URMP dataset [50] under moderate reverberation (RT_{60} = 250 ms) with simulated room. Each column corresponds to a different array size (2, 4 and 8 microphones). Method abbreviated as (GT) uses ground-truth source DoA (represented in light color) whereas (SRP) uses SRP-PHAT estimation (represented in dark color). The error bars represent 95% confidence intervals.

the 8-channel array case the method is evaluated over all excerpts. *Harmonic NTF* obtains slightly better SDR than *ILRMA*, probably due to its harmonic constraint. However, it does not provide proper isolation from the sources, as can be observed from the SIR value obtained.

Fig. 4 illustrates the separation results in a moderate reverberant scenario (i.e., for the room with $\mathrm{RT}_{60} = 250$ ms). In this condition, the SDR performance of all methods, except *DSB + Wiener*, goes below the reference value provided by *ED*. When initialized with SRP-PHAT, our proposals still manage to obtain better SDR than the other decomposition methods, especially with 4 and 8 channels. Again, compared to the *DoA-MNMF* approach, our harmonic-constrained model based on the CQT obtains a much better approximation to the real sources according to the SDR metric. In terms of interferences, our method performs best with 2 channels, whereas *Mult. NMF conv.* offers the best SIR score with 4 and 8 channels. Nonetheless, *Proposal-CQT* always seems to benefit from ground-truth initialization of the source directions, reaching the best SIR score with 2 channels and 8 channels. *Proposal-CQT* variant also performs best in reconstructing the spatial image of the sources (ISR) with 8 channels, and slightly outperforms *Mult. NMF conv.* and *Mult. NMF inst.* with 2 and 4 channels.

The separation results for the room with $\mathrm{RT}_{60} = 500$ ms are shown in Fig. IV-D7. This is an extremely challenging scenario, since the mixture at each channel is affected by a strong reverberation tail and echos coming from multiple directions. As can be seen, the behavior is similar to the $\mathrm{RT}_{60} = 250$ ms case. *DSB + Wiener* obtains the best results in terms of SDR and SAR performance. However, it provides the worst SIR value, resulting in a really poor isolation of

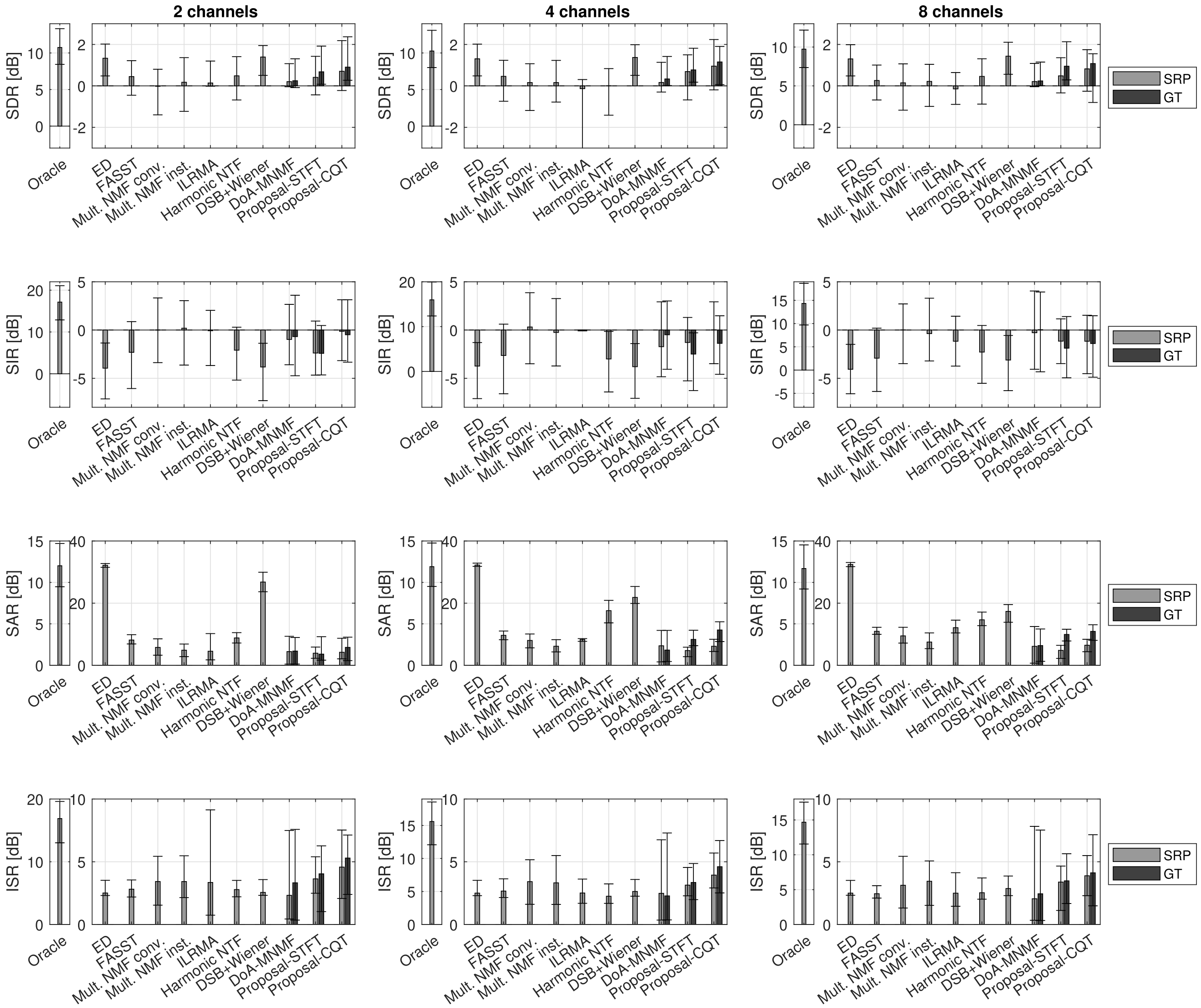


FIGURE 5. Source separation metrics averaged over the URMP dataset [50] under reverberant conditions (RT_{60} = 500 ms) with simulated room. Each column corresponds to a different array size (2, 4 and 8 microphones). Method abbreviated as (GT) uses ground-truth source DoA (represented in light color) whereas (SRP) uses SRP-PHAT estimation (represented in dark color). The error bars represent 95% confidence intervals.

the sound sources. Concerning our proposals, better SDR and ISR are obtained compared to the other decomposition methods. In terms of interferences, *Mult. NMF conv.* again offers the best SIR score. Even so, our method initialized with SRP-PHAT shows competitive results regarding this score. In this case, the ground-truth initialization does not benefit our proposals. This can be due to the fact that, with high reverberation and small arrays, constraining each source to a single direction of arrival is not significantly beneficial in terms of interferences due to the strong reflections coming from other sources.

Fig. 6 reports the separation metrics of our CQT variant (initialized with SRP-PHAT) as a function of the number of sources in the mixture. As shown, the results scale as expected according to the complexity of the mixture. For two and three sources under slight reverberation, our system gets a very high level of isolation with very low artifacts even in the 2-channel case. With four and five sources, the separation scores decrease significantly, but the method is still able to offer acceptable results under low reverberation, especially using the 8-channel array (around 1.6 dB SDR and 2.5 dB SIR for five sources). We provide sound samples of separated signals at the web page of results.[2]

Table 1 shows the computational time of our proposed method variants. Note that the computation time increases as the number of sources in the mixture increases. As can be seen, the CQT variant takes much less time than the STFT variant. Regarding the spatial weights initialization, it can be clearly observed how initializing with SRP-PHAT algorithm

[2] https://antoniojmm.github.io/Harmonic_CQT_MNMF.github.io/

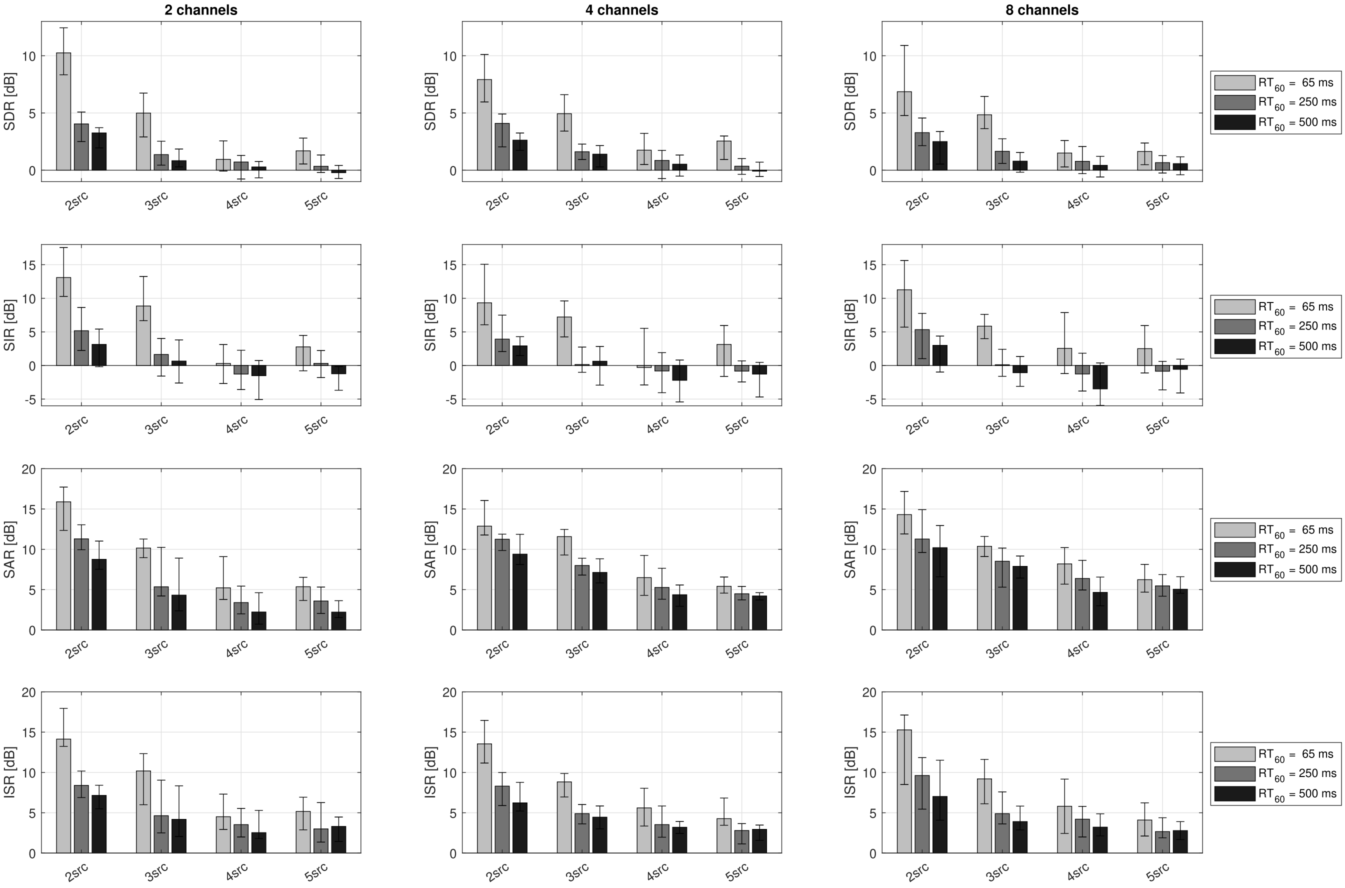


FIGURE 6. Separation metrics of our method for different numbers of sources in URMP dataset [50] with the simulated room. Initialization is done with SRP-PHAT. Each column corresponds to a different array size (2, 4 and 8 microphones). The error bars represent 95% confidence intervals.

TABLE 1. Computational time (in seconds) for 300 iterations of the proposed method variants.

Method	Initial.	Sources			
		2	3	4	5
Proposal-STFT	Full-rank	17493	18261	19416	20186
	Low-rank	10956	11919	12809	14128
Proposal-CQT	Full-rank	1428	1531	1685	1773
	Low-rank	995	1079	1207	1327

offers a great advantage in terms of computational time with respect to considering the full-rank version of **Z**.

V. CONCLUSION

In this paper, we present a harmonic constrained MNMF-based method for the task of blind music source separation. In the proposed signal model, the mixing filter encodes the spatial information in terms of magnitude and phase differences between channels, whereas the source variances are modelled using a harmonic constrained NMF structure. In order to reduce the dimensionality of the signal model parameters, we propose to use the CQT as time-frequency representation. Thus, the SCM is obtained from the CQT to account to the frequency logarithmic scale inherent in music signals. To our best knowledge, this is the first work that exploited the phase information of the CQT within an MNMF scheme. Moreover, the SRP-PHAT algorithm is used to initialize the model parameters and thus, reduce the computational complexity and increase the robustness.

The proposed method has been evaluated for the task of multichannel music source separation of classical chamber music ensembles with several polyphony and reverberation setups. The results obtained in the evaluation showed a reliable performance in terms of BSS_EVAL metrics in comparison with other signal decomposition approaches.

Finally, as future work, we would investigate the integration of the score information within the signal model. Thus, we would expect to improve the separation results and reduce the computational complexity by initializing the time-varying gains of the model. Moreover, we will try to improve the results in high reverberation scenarios by evolving the phase model. Finally, since the current system needs to know the microphone arrangement, we will work on developing a blind system respect to the microphone array.

REFERENCES

[1] F. Canadas-Quesada, D. Fitzgerald, P. Vera-Candeas, and N. Ruiz-Reyes, "Harmonic-percussive sound separation using rhythmic information from non-negative matrix factorization in single-channel music recordings," in *Proc. 20th Int. Conf. Digit. Audio Effects (DAFx)*, 2017, pp. 276–282.

[2] J.-L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[3] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, no. 1, pp. 1–13, Dec. 2010.

[4] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 71–75.

[5] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.

[6] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3339–3343.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2018, pp. 334–340.

[8] A. Défossez, F. Bach, N. Usunier, and L. Bottou, "Music source separation in the waveform domain," 2019, *arXiv:1911.13254*.

[9] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. 19th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2018, pp. 289–296.

[10] P. Seetharaman, G. Wichern, S. Venkataramani, and J. L. Roux, "Class-conditional embeddings for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 301–305.

[11] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10891. Cham, Switzerland: Springer, 2018, pp. 293–305.

[12] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 106–110.

[13] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," 2019, *arXiv:1909.01174*.

[14] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 816–820.

[15] D. FitzGerald, "Non-negative tensor factorisation for sound source separation," in *Proc. IEE Irish Signals Syst. Conf.*, Sep. 2005, pp. 8–12.

[16] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Comput. Intell. Neurosci.*, vol. 2008, pp. 1–15, May 2008.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[18] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 184, Dec. 2013.

[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[20] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[21] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2610–2625, 2020.

[22] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas, "Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1512–1527, Sep. 2018.

[23] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6677–6681.

[24] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 371–375.

[25] Y. Mitsufuji, S. Uhlich, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari, "Multichannel non-negative matrix factorization using banded spatial covariance matrices in wavenumber domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 49–60, 2020.

[26] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 815–826, Aug. 2019.

[27] A. J. Munoz-Montoro, A. Politis, K. Drossos, and J. J. Carabias-Orti, "Multichannel singing voice separation by deep neural network informed DOA constrained CMNMF," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[28] J. Casebeer, M. Colomb, and P. Smaragdis, "Deep tensor factorization for spatially-aware scene decomposition," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 180–184.

[29] J. C. Brown, "Calculation of a constant *Q* spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[30] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 245–248.

[31] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 2654–2658.

[32] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "SpecTNT: A time-frequency transformer for music audio," 2021, *arXiv:2110.09127*.

[33] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound Music Comput. Conf.*, Jan. 2010, pp. 3–64.

[34] I. J. Tashev, *Sound Capture and Processing*. Chichester, U.K.: Wiley, 2009.

[35] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[36] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Proc. Adv. Models Acoustic Process., Neural Inf. Process. Syst. Workshop*, vol. 18, 2006, pp. 1–5.

[37] S. Raczyński and N. Ono, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. 8th Int. Conf. Music Inf. Retr.*, 2007, pp. 381–386.

[38] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[39] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.

[40] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Canadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram factorization," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1144–1158, Oct. 2011.

[41] J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes, "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1671–1680, Aug. 2013.

[42] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 888–891.

[43] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 3073–3076.

[44] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. 10th Int. Soc. Music Inf. Retr. Conf.*, 2009, pp. 327–332.

[45] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[46] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. 7th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2006, pp. 314–319.

[47] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.

[48] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Audio pitch shifting using the constant-Q transform," *J. Audio Eng. Soc.*, vol. 61, no. 7, pp. 562–572, 2013.

[49] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[50] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[51] D. R. Campbell, K. J. Palomaki, and G. Brown, "A MATLAB simulation of 'Shoebox' room acoustics for use in research and teaching," *Comput. Inf. Syst.*, vol. 9, no. 3, pp. 48–51, 2005.

[52] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

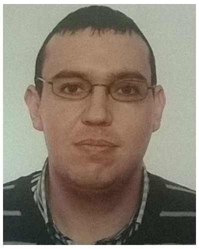
[53] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[54] M. Miron, J. J. Carabias-Orti, J. J. Bosch, E. Gómez, and J. Janer, "Score-informed source separation for multichannel orchestral recordings," *J. Electr. Comput. Eng.*, vol. 2016, pp. 1–19, Dec. 2016.

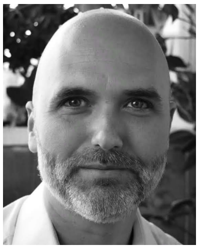
[55] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.

ANTONIO J. MUÑOZ-MONTORO received the M.Sc. degree in telecommunications engineering from the University of Malaga, Spain, in 2015, and the Doctor of Science degree from the University of Jaén, Spain, in 2020. He is currently a Postdoctoral Researcher with the Computer Science Department, University of Oviedo. His research interests include signal processing and machine learning. In particular, his contributions are related to music signal processing applications, including source localization, source separation, and audio-to-score alignment.

JULIO J. CARABIAS-ORTI received the M.Sc. degree in computer science and the Doctor of Science degree from the University of Jaén, Jaén, Spain, in 2006 and 2011, respectively. He is currently a Postdoctoral Researcher with the Telecommunication Engineering Department, University of Jaén. His research interests include signal processing and machine learning, with a focus on signal decomposition methods for music signal processing applications including music transcription, source separation, and audio to score alignment.



PABLO CABAÑAS-MOLERO received the M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Jaén, Spain, in 2008 and 2016, respectively. He is currently a Postdoctoral Researcher with the Telecommunication Engineering Department, University of Jaén. His research interests include sound source separation, automatic sound classification, speech enhancement, and audio-to-score alignment.



FRANCISCO J. CAÑADAS-QUESADA was born in Linares, Jaén, Spain, in 1977. He received the Ph.D. degree from the University of Jaén, Spain, in 2009. Since 2006, he has been with the Telecommunication Engineering Department, University of Jaén. He is currently an Associate Professor in signal processing and communications area. His research interests include signal processing applied to sound source separation, factorization algorithms, and machine learning. He is the coauthor of about 30 technical publications, including more than ten journal citation reports publications covering *Signal Processing* and *Acoustics and Ultrasonics*. He has been involved in research projects of the Spanish Ministry of Science and Education (MEC) and private companies.

NICOLÁS RUIZ-REYES was born in Linares, Jaén, Spain, in 1967. He received the M.Sc. degree from the Technical University of Madrid (UPM), Madrid, Spain, in 1993, and the Ph.D. degree in telecommunication engineering from the University of Alcala, Madrid, in 2001. Since 2010, he has been a Full Professor in signal processing and communications with the Telecommunication Engineering Department, University of Jaén. His research interests include signal processing and its applications to communications, speech and audio analysis, electrical and biomedical engineering, and ultrasonic NDT. He is the coauthor of more than 100 papers in prestigious journal and has been involved in research projects of the Spanish Ministry of Science and Education, European Commission, and private companies.