Sound Recording and Source Separation of Acoustic Instruments

P. Chouteau H. Desvallées L. Lalay M. Lefebvre I. Meresman Higgs T. Nguyen

Abstract

Music Source Separation (MSS) is the process of separating individual audio signals from a mixed recording containing multiple sound sources, such as different musical instruments, vocals and ambient noise. Its various applications include remixing, transcription and music recommendation. In the context of real acoustic recordings, the separation task is particularly challenging due to the complexity and variability of acoustic instruments and recording conditions such as room acoustics and microphone directivity. We propose the use of Non-negative Matrix Factorization (NMF) algorithms for this task, and in our multi-channel setting, we aim to implement efficient, conditioned versions of this algorithm to be applied to musical recordings performed in a known and controlled context. We investigate methods of informing this algorithm by conditioning on temporal and spectral information from the instruments, that were specifically registered at the time of the recording for this purpose. To this end, we conducted a professional-level recording of a chamber music quintet.

We have compared our results with other state-of-the-art algorithms, obtaining comparable results on benchmark datasets, and we have carried out subjective evaluation according to the MUSHRA protocol, where we see a good performance of our algorithm. We observe a strong effect of the processing of the recording, which helps or hinders the separation depending on the instrument. Our approach confirms the versatility of the FastMNMF algorithm and the possibility of extending and making these algorithms more versatile. Audio results can be heard on our website.

Keywords: Music, Source separation, Non-Negative Matrix Factorization, Live Recording, Acoustics

1 Introduction

Blind Source Separation (BSS) is the process of isolating individual *unobserved* sources in an observed mixture of multiple sources [Cardoso [4]]. When applied to audio signals, it can be practised on musical or non-musical signals, and each case presents a different set of challenges and difficulties. In this paper, we will use musical audio signals, where source separation is a process intended to be applied for tasks as diverse as musical rearrangement, musicological analysis, remixing or sampling, generating a karaoke soundtrack, or creating musical scores.

Musical signals are generated by a complex process that begins in the case of acoustic instruments with the vibration of musical instruments and includes the acoustic radiation from the instrument, the movements that accompany a musical performance, the response of the recording space to the sound waves, the selection, placement, and configuration of microphones, and the transformations that occur in post-production before a final, finished recording is obtained [Bartlett and Bartlett [2]]. It is common for source separation methods to attempt to use *a priori* knowledge of the structure of the source signals and the nature of the transformations they have undergone [Vincent et al. [35]].

The complex and varying nature of these processes does not aid the task of integrating this knowledge to the methods. Our work proposes to address this task and its challenges by implementing a conditionable state-of-the-art (SOTA) algorithm for source separation, and by gaining control over the music recording process to refine the method, extend the *a priori* knowledge, and generate specific material to improve the separation.

The Nonnegative Matrix Factorization (NMF) is a popular machine learning algorithm that aims at decomposing a nonnegative matrix $\mathbf{X} \simeq \mathbf{WH}$ into a product of two matrices $\mathbf{W}$ and $\mathbf{H}$ [Lee and Seung [16]]. In our scenario, it decomposes the power spectrogram of a punctual source, as the product of a frequency base matrix $\mathbf{W}$ and a time activation matrix $\mathbf{H}$. In a multichannel recording condition, an extension called Multi-channel NMF (MNMF) [Ozerov and Fevotte [21]] tries to take advantage of many recordings (*e.g.* from several microphones) of the same sound to better separate the sources by modelling a common source model but with different spatial models.

The parameter optimization is either based on an Expectation-Maximization (EM) algorithm [Dempster et al. [7]] or minimization of Itakura-Saito divergence leading to Multiplicative update rules (MUR) that guarantee the nonnegativity of the matrices [Févotte and Idier [10]]. Another extension by Duong et al. [8] proposes a statistical model of either punctual or diffuse sources by considering a so-called full-rank spatial covariance matrix (SCM) that summarizes the acoustic paths between each source and microphones and solves the parametrization estimation through an EM algorithm. In Sawada et al. [26], a SCM/MNMF combination with a MUR using an auxiliary function technique is designed and outperforms other MNMF extensions. However, all those MNMF algorithms suffer from high computational costs and a strong dependence on their initialisation. Famous lighter versions that were proposed are Auxiliary independent vector (AuxIVA [Ono [20]]) which is a degenerated version of the rank 1 SCM version, Independent Low-Rank Matrix Analysis (ILRMA [Kitamura et al. [14]]) which combines AuxIVA with an NMF model and FastM-

NMF which decomposes the SCMs into a common basis and an NMF decomposition of the power spectrograms.

We focus on FastMNMF, [Sekiguchi et al. [29]] which can be seen as a trade-off between a lighter and more robust version, without considering a degenerate rank 1 model that is less suitable for a reverberant environment. Our proposal for improving the performance of FastMNMF is to incorporate *a priori* knowledge by having a more complete understanding and control of the recording process by which the mixture was generated. To this end, we tested our algorithm on simulated recording situations using the *Pyroomacoustics* system [Scheibler et al. [27]] and the MUSDB18 dataset [Rafii et al. [23]]. We then carefully planned and executed a recording session of a semi-professional chamber music quintet under standard recording conditions for this type of music, as shown in Figure 1. During this session, additional material was played by the musicians and recorded in order to obtain data that would facilitate optimal initialisation schemes for the algorithm.

Finally, due to the limited amount of data that can be recorded in a session and the variety of possible configurations of SOTA systems, and to the fact that we cannot access the ground truth signal from reverberant recordings, we decided to replace the traditional objective evaluation schemes with subjective Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [BS.1534-3 [3]] test in order to evaluate the performance of our system.

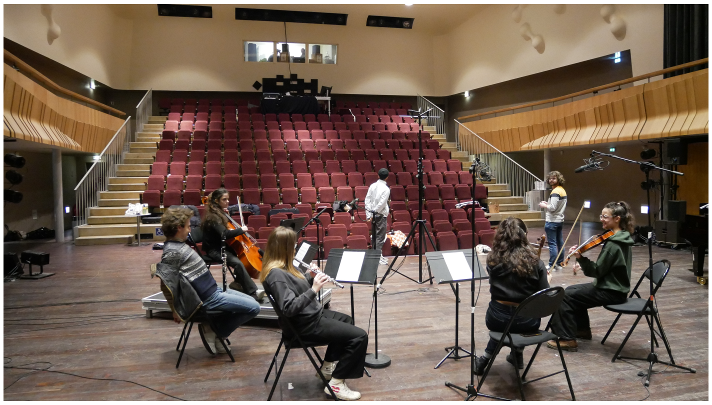


Figure 1: Chamber music quintet recording session

2 State-of-the-Art Review

2.1 Models for Source Separation Algorithms

Source separation as a signal-processing task has received a lot of interest since the turn of the century. Today, deep learning models achieve SOTA results in very challenging contexts, such as underdetermined separation (more sources registered in the mixed signals than signals available) thanks to a supervised framework [Nugraha et al. [19]]. However, they do so in an uninterpretable way, require a large amount of data to work, and while they achieve great results in the traditional objective metrics, they often present numerous artefacts and undesirable interference to the human listener. They also suffer from a poor domain adaptation issue meaning that a “good separation performance” in room A may be highly degraded in room B. On the other hand, unsupervised algorithms such as Maximum Likelihood BSS or Non-Negative Matrix Factorisation (NMF) may require more information about the signal, but under certain conditions can produce excellent performance, while providing a comprehensive explanation of how the algorithm reached such results.

2.1.1 Linear Instantaneous Model

Standard BSS [Cardoso [4]] assumes the existence of n independent signals $s_1(t), \dots, s_n(t)$ and the observation of as many mixtures $x_1(t), \dots, x_n(t)$. These mixtures being linear and instantaneous, i.e. $x_i(t) = \Sigma_{j=1}^{n} a_{ij} s_j(t)$ for each $i = 1, \dots, n$. *i.e.* each receptor x_i is a linear combination of all source signals. BSS assumes independence between the entries of the input vector $\mathbf{s}(t)$ and primarily exploits 'spatial diversity', i.e. different sensors receive different mixtures of the sources.

This is represented compactly by the mixing equation:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

Where $\mathbf{s}(t) = [s_1(t), \dots, s_n(t)]^T$ is an $n \times 1$ column vector collecting the source signals, vector $\mathbf{x}(t) = [x_1(t), \dots, x_n(t)]^T$ similarly collects the n observed signals and the square $n \times n$ ‘mixing matrix’ A contains the mixture coefficients.

The simple model $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ is parameterized by the pair $(\mathbf{A}, q)$ made from the mixing matrix $\mathbf{A}$ and the density q for the source vector s. The density of $x = mathbfAs$ for a given pair $(\mathbf{A}, q)$ is classically given by:

$$p(x; \mathbf{A}, q) = |det\mathbf{A}|^{-1} q(\mathbf{A}^{-1}x) \tag{2}$$

The goal of BSS is then to maximise the probability distribution $p(\mathbf{x}; \mathbf{A}, q)$ to estimate a matrix $\mathbf{A}$.

2.1.2 Convolutional Model

If we consider the n previous sources in the short-time Fourier transform (STFT) domain with F frequency bins and T time frames. As mentioned in the introduction, NMF [Lee and Seung [16]] is a type of BSS algorithm, that estimates a given non-negative matrix $\mathbf{X} \in \mathbb{R}^{+^{F \times T}}$ as the product of a base matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and an activation matrix $\mathbf{H} \in \mathbb{R}_+^{K \times T}$, such as $\mathbf{X} \simeq \mathbf{W}\mathbf{H}$. K is the number of basis, its optimal value is discussed below. The Multi-channel NMF (MNMF) extension [Ozerov and Fevotte [21]] estimates $\mathbf{X}_n \in \mathbb{R}_+^{F \times T \times M}$ as the image of source n perceived by all sensors, with the aim to estimate all $\mathbf{X}_n$ with $x_{ftn} = a_{nf} s_{ftn}$, where the steering vectors a_{nf} are estimated together with $\mathbf{W}$ and $\mathbf{H}$.

These algorithms share the same source model, based on a Gaussian mixture $s_{ftn} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{ftn})$ where:

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf} h_{nkt} \tag{3}$$

with w_{nkf} and h_{nkt} being components of matrices $\mathbf{W}$ and $\mathbf{H}$ of the NMF decomposition, but use different spatial models which can be rank-1 ($\mathbf{x}_{ftn} = a_{nf} s_{ftn}$ with a_{nf} the steering vector of source n at frequency f, where the sources can then be derived thanks to the demixing matrix $\mathbf{D}_f = \mathbf{A}_f^{-1} = [a_{1f}, \dots, a_{Nf}]$ and the relation $s_{ft} = \mathbf{D}_f x_{ft}$) or full-rank ($x_{ftn}$ is estimated with Wiener filtering), which can be interpreted physically as related to the acoustic model of the sources. This source model is also used in other SOTA BSS algorithms, such as Independent Low-Rank Matrix Analysis (ILRMA).

Sekiguchi et al. [29]’s FastMNMF proposes extensions to MNMF by using a jointly diagonalisable full-rank spatial model and proposing rank constraints for the spatial covariance matrix of each source, thus adding a new way of incorporating a priori knowledge, the rank is derived from the directivity of the source: A highly directional source will have a one-hot diagonal vector (only one sensor is affected by a source), whereas a diffuse source will have an all-one diagonal vector (each sensor is affected). The rank of the spatial matrices is defined by the number of non-zero diagonal terms in $\mathbf{G}_{nf}$, and since once a parameter is set to zero, the update rules keep it at 0, the rank can be obtained by initialising a given number of diagonal terms to 0. Some of the proposed initialisations are Random, Diagonal, Circular and Gradual.

Other ways of informing these types of algorithms are proposed as a framework generalisation by Ozerov et al. [22], not only can the spatial covariance model change between rank-1 and full-rank spatial covariance matrices, but they propose flexibility in the type of input representation (e.g. STFT or Equivalent Rectangular Bandwidth), the problem dimensionality (related to the number of channels of the observed mixture M and the number of sources to separate N), and the spectral power model (which may depend on the type of source). This is done via N 9-dimensional parameters modelling the spatial covariance and the spectral power (as a source-filter model), with matrices modelling the narrowband spectral patterns (W_i), the spectral pattern weights (U_i), the temporal pattern weights (G_i) and the temporally localised patterns (H_i). These can be initialised to desired values, frozen during part or all of the training, and more.

2.2 Evaluation

A key element in source separation is the evaluation of the separation. This task is complex because the human ear is very finely tuned to unexpected or unusual sounds and noises, and it is difficult to find an evaluation criterion that can correctly detect what our ear perceives. In general, there are two main ways of evaluating the results of a source separation approach: objective and subjective, both of which have their advantages and disadvantages [Zieliński et al. [37], Manilow et al. [17], Gusó et al. [12]].

Objective measures evaluate the quality of source separation by performing a series of calculations that compare the output signals of a source separation system with the "ground truth" isolated sources. The most commonly used are the Source-to-Distortion Ratio (SDR), the Source-to-Interference Ratio (SIR), and the Source-to-Artifact Ratio (SAR), as defined by Vincent et al. [34], and the Scale-Invariant SDR (SI-SDR) as defined by Roux et al. [24]. For these metrics, estimates of a source $\hat{s}_i$ are assumed to consist of four separate components, the target source (a version of s_i modified by an allowed distortion) s_{target}, the interference error e_{interf}, the noise error e_{noise}, and the artefacts e_{artif}.

These four terms represent the part of estimated source $\hat{s}_i$ perceived as coming from the source s_i, from other unwanted sources $s_{j'}$ with $j' \neq j$, and from noises proceeding from sensors, distortions, and artefacts. They are obtained through a decomposition on orthogonal projections: with $\Pi\{y_1, ...y_k\}$ the orthogonal projector onto the subspace spanned by vectors $y_1, ...y_k$ of length T, in the shape of a $T \times T$ matrix. Three orthogonal projectors are defined:

$$P_{s_j} := \Pi\left\{s_j\right\}, \tag{4}$$

$$P_{\mathbf{s}} := \Pi\left\{(s_{j'})_{1\leq j'\leq n}\right\}, \tag{5}$$

$$P_{\mathbf{s},\mathbf{n}} := \Pi\left\{(s_{j'})_{1\leq j'\leq n}, (n_i)_{1\leq i\leq m}\right\} \tag{6}$$

to finally decompose $\hat{s}_i$ as the sum of four terms:

$$s_{\text{target}} := P_{s_j}\widehat{s}_j, \tag{7}$$

$$e_{\text{interf}} := P_{\mathbf{s}}\widehat{s}_j - P_{s_j}\widehat{s}_j, \tag{8}$$

$$e_{\text{noise}} := P_{\mathbf{s},\mathbf{n}}\widehat{s}_j - P_{\mathbf{s}}\widehat{s}_j, \tag{9}$$

$$e_{\text{artif}} := \widehat{s}_j - P_{\mathbf{s},\mathbf{n}}\widehat{s}_j. \tag{10}$$

More details of the computation are available in Vincent et al. [34], but the decomposition now permits the definition of the metrics as:

$$\text{SAR} := 10\log_{10}\left(\frac{||s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2}\right) \tag{11}$$

$$\text{SIR} := 10\log_{10}\left(\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}\right) \tag{12}$$

$$\text{SDR} := 10\log_{10}\left(\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2}\right) \tag{13}$$

Though SDR is unquestionably the most popular of these three, it presents a big problem in that it depends on the amplitude scaling of the signal and so can be artificially inflated. This is addressed by Roux et al. [24]'s SI-SDR metric which rescales the target by finding the orthogonal projection of $\hat{s}_i$ on the line spanned by target s_i, and is defined as:

with $e_{target} = \alpha s$ for $\alpha = argmin_\alpha|\alpha s_i - \hat{s}_i|^2$.

One of the main obstacles for implementation of these metrics is the need for a clearly defined "ground truth" recording of the desired source, restricting this kind of evaluation to sound sources that are previously reduced to a 1-dimensional audio signal. On the other hand, subjective measures are the gold standard for measuring the quality of BSS and do not present this problem.

Subjective evaluation procedures involve human evaluators assigning scores to the output of the source separation system. The SOTA for subjective evaluation is called Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), the test should be developed following the International Telecommunications Union's recommendations [BS.1534-3 [3]]. It is ideally performed by well-trained audio engineers in a sound-treated room and compares the qualities of various conditioned sound files with a reference. Among the files to compare should be a hidden reference and at least one hidden anchor, an intentionally bad variant of the audio which is traditionally obtained by low-pass filtering. These are meant to provide boundaries across evaluators, and alert if any rating may have been incorrectly performed.

Crowd-sourced based variants have been implemented [Cartwright et al. [5],Schoeffler et al. [28]] that can be performed by any fit person with a pair of headphones, and these have been shown to be an effective alternative.

2.3 Datasets

There are more than a few datasets available for source separation, but the undisputed benchmark is the MUSDB18 dataset [Rafii et al. [23]] (which also offers an uncompressed high-quality version), used in the Music Demixing Challenge [Mitsufuji et al. [18]] and many other community staples, it consists of 150 full-length music tracks (approximately 10h duration) of various genres along with their isolated drums, bass, vocals and "other" stems.

One of the drawbacks of this dataset is the limited information about how the instruments were recorded and the processing that was applied to them, although the recordings are relatively clean, allowing for post-processing and data augmentation using classic mixing tools as well as room simulation tools such as *Pyroomacoustics* [Scheibler et al. [27]].

2.4 Sound Recording

When it comes to the recording of musical sound, there is no defined consensus as to what are the best practices that lead to an optimal recording. There are, in fact, many ways to make a good music recording, and the difficulty in standardisation lies in the variability of factors such as instrument quality, room characteristics, room conditions at the time of recording, and the processing preferences of the recording engineer. In the case of classical chamber music, although it is possible to find recordings made in isolated (between instruments) studio conditions, it is common practice to record such groups playing together. It is also relevant to record in a concert hall or reverberant recording studio, so as to emulate the experience of the musicians playing live, and the experience of the listener who would normally hear the music under these conditions [Lang [15], Spiro and Schober [30]].

Although various microphone arrangements have been proposed over the years, especially in recent years for recordings that focus on spatialisation [Alexandridis et al. [1], Zhang et al. [36]], the traditional approach of placing "spot" microphones close to each individual source, combined with coupled microphones that capture segments or the whole ensemble and provide a natural spatial component [Streicher and Dooley [31], Theile [32], Sarkar et al. [25]], is often still the preferred method.

Recommendations and common practices for microphone placement when recording classical instruments can be found in Hugonnet and Walder [13] and Valentine [33]. Streicher and Dooley [31] provides a complete overview of all common stereo recording methods, suggesting the main configurations with 2, 3 or 4 microphones, their placements and characteristics, and presenting the advantages and disadvantages of common methods such as Coincident Stereo Techniques (XY, Blumlein, MS Stereo), Near Coincident Techniques (O.R.T.F, N.O.S, Faulkner, Binaural) and Spaced Techniques.

It also exists more complex recording techniques used for three or five loudspeakers restitution. Those practices as INA3, near-coincident LCR, Widely Spaced Omni or Optimised Cardioid Triangle (OCT) are configuration explained in Theile [32].

3 Methodology

3.1 Algorithm

This section presents our approach to the development of the separation algorithm. It starts by explaining certain specifics of Sekiguchi et al. [29]'s FastMNMF2 on which we based our model, then introduces our improvements which were obtained by adding constraints such as those brought by Ozerov et al. [22].

3.1.1 FastMNMF2

FastMNMF2 is introduced by Sekiguchi et al. [29] as a computationally efficient algorithm based on multichannel NMF for source separation. FastMNMF2 takes the multichannel observed spectrogram $\mathbf{X}$ as argument, and outputs the separated multichannel spectrograms corresponding to each source as if the source alone was observed by the microphones. The multichannel NMF updates sources matrices $\mathbf{W}$ and $\mathbf{H}$, and spatial matrices $\mathbf{Q}$ and $\tilde{\mathbf{G}}$, in order to estimate the separated spectrograms. $\mathbf{X} \in {\mathbb{R}^+}^{F\times T\times M}$ is the observed spectrogram, $\mathbf{W} \in {\mathbb{R}^+}^{N\times F\times K}$ is the spectral matrix, $\mathbf{H} \in {\mathbb{R}^+}^{N\times K\times T}$ is the activation matrix, $\mathbf{Q} \in \mathbb{C}^{N\times F\times K}$ is the diagonaliser of $\mathbf{G}$, and $\tilde{\mathbf{G}} \in {\mathbb{R}^+}^{\mathbf{N\times F\times K}}$ the diagonal components of the spatial matrix.

The outputs are the source images estimated by a Wiener filter written as follows:

$$\mathbb{E}[x_{ftm}|\mathbf{x}_{ft}] = \mathbf{Q}_f^{-1} Diag\left(\frac{\lambda_{ftn}\tilde{\mathbf{g}}_{nf}}{\sum_n \lambda_{ftn}\tilde{\mathbf{g}}_{nf}}\right)\mathbf{Q}_f\mathbf{x}_{ft} \tag{14}$$

with $\lambda_{ftn} = \sum_k w_{nkf}h_{nkt}$.

We obtain the parameters needed in Equation 14 by maximizing the following log-likelihood:

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \tilde{\mathbf{G}}, \mathbf{Q}) = &- \sum_{f,t,n}\left(\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftm}} + \log \tilde{y}_{ftm}\right) \\ &+ T\sum_f \log|\mathbf{Q}_f\mathbf{Q}_f^H|\end{aligned}$$

with $\tilde{x}_{ftm} = |\mathbf{q}_{fm}^H\mathbf{x}_{ft}|^2$, $\tilde{y}_{ftm} = \sum_{n,k} w_{nkf}h_{nkt}\tilde{g}_{nm}$.

The update rules derived from the log-likelihood are shown in Equation 15 to Equation 18.

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m} h_{nkt}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m} h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}} \tag{15}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m} w_{nkf}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m} w_{nkf}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}} \tag{16}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm}\sqrt{\frac{\sum_{f,t,k} w_{nkf}h_{nkt}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,t,k} w_{nkf}h_{nkt}\tilde{y}_{ftm}^{-1}}} \tag{17}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{V}_{fm})^{-1} e_m \tag{18}$$

with $\mathbf{V}_{fm} = \sum_t \mathbf{x}_{ft}\tilde{y}_{ftm}^{-1}$.

Between each step, the matrices are normalized.

3.1.2 Split source model

In Ozerov et al. [22], a further factorization of the source model is proposed. The matrices $\mathbf{W}$ and $\mathbf{H}$ are decomposed as in Equation 20[1].

$$\mathbf{W} = \mathbf{EU} \tag{19}$$

$$\mathbf{H} = \mathbf{TP} \tag{20}$$

with $\mathbf{E} \in {\mathbb{R}^+}^{N\times F\times L}$,$\mathbf{U} \in {\mathbb{R}^+}^{N\times L\times K}$ $\mathbf{T} \in {\mathbb{R}^+}^{N\times K\times O}$,$\mathbf{P} \in {\mathbb{R}^+}^{N\times O\times T}$. $\mathbf{E}$ and $\mathbf{P}$ are basis matrices holding prior information

[1] This notation differs from the original paper to avoid conflict with previous names

and are not updated in the main loop, while $\mathbf{U}$ and $\mathbf{T}$ are weight matrices updated in the main loop. The goal of this new factorization is to use prior information in $\mathbf{W}$ and $\mathbf{H}$ while keeping some degree of freedom.

Under the assumption of $\mathbf{E}^+$ and $\mathbf{P}^+$ (pseudo inverses of $\mathbf{E}$ and $\mathbf{P}$) being non-negative matrices, we can modify the previous update rules as in Equation 21 and Equation 22, with $\odot$ the element-wise product, N_1, N_2, D_1, D_2 the update coefficients.

$$\mathbf{U} \leftarrow \mathbf{E}^+ \left((\mathbf{EU}) \odot \sqrt{\frac{N_2(\tilde{\mathbf{G}}, \mathbf{H})}{D_2(\tilde{\mathbf{G}}, \mathbf{H})}} \right) \tag{21}$$

$$\mathbf{T} \leftarrow \left((\mathbf{TP}) \odot \sqrt{\frac{N_1(\tilde{\mathbf{G}}, \mathbf{W})}{D_1(\tilde{\mathbf{G}}, \mathbf{W})}} \right) \mathbf{P}^+ \tag{22}$$

3.1.3 Use of prior information

We can use 3 types of prior information: spatial ($\mathbf{Q}$ and $\tilde{\mathbf{G}}$), temporal ($\mathbf{H}$) or frequential ($\mathbf{W}$).

- Frequential: $\mathbf{W}$ or $\mathbf{E}$ is filled with a dictionary of notes (composed of the squared FFT of each individual note played by an instrument). The matrix $\mathbf{W}$ is kept frozen for a given number of iterations to let the algorithm adapt W at the end, whereas $\mathbf{E}$ is always kept frozen.
- Temporal: $\mathbf{H}$ is filled with information related to the impulse response. This part was not implemented.
- Spatial: delays and attenuation are given to $\tilde{\mathbf{G}}$ and $\mathbf{Q}$ from the sources and microphones positions. This part was not implemented due to lack of precise information and the movement of the musicians.

The dictionaries are needed for the frequential prior information. Each source needs its own dictionary, so every instrument played a chromatic scale covering all the notes played in the piece, including pizzicato for the strings. Then the notes were extracted individually from the scales and time-shifted to start at 0 seconds. Finally, the squared FFT of each note was concatenated to form a matrix and either stand as $\mathbf{W}$ or $\mathbf{E}$ in the algorithm.

3.2 Simulated Data

To test our implementation of the FastMNMF2 algorithm, we created a room simulation with the *Pyroomacoustics* [Scheibler et al. [27]] python library. Set on obtaining a simulation as close as possible to a real situation, we simulated a room with sources and microphones similar to those used in the recording. The sources are placed in a circular arc and the microphones used are cardioid. Two microphones are placed in an AB configuration, the others are used as spot (close) microphones.

This simulation allowed us to test the performance of our implementation on the *MUSDB18* dataset [Rafii et al. [23]]. For all these experiments, 4 different sources were used (bass, drums, others and vocals). Using 6, 4 or 2 microphones for all cases of source separation, overdetermined, determined and underdetermined respectively, the algorithm parameters were also explored for optimal computation.

3.3 Recordings

The sound recording session took place in the auditorium of the regional conservatory in Aubervilliers (see Figure 1 and Figure 2). This venue was chosen because of its pleasant acoustic response to live classical music. The recorded quintet was comprised of two violins, a cello, a clarinet, and a flute, playing an original piece named *Perdrix* by composer Inès Lassègue.

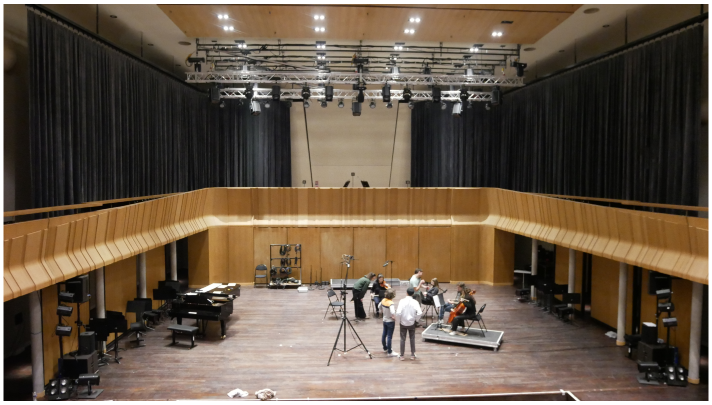


Figure 2: Auditorium for the sound recording

An eight-microphone setup was used to record the quintet. Bidirectional microphones (reference in Table 1) were used as spot microphones to minimise bleeding between instruments. Three other microphones with cardioid polar patterns were added to record the sound field used as the basis for the soundtrack. A three-mic setup called LCR (Left-Centre-Right) allows the use of the phantom centre to be avoided and allows for better panning between sources. A total of eight microphones makes it possible to test different configurations with the source separation program: under-determined, determined and over-determined.

Reference	Directivity	Instrument
Neumann U87	bidirectional	Violin 1
Neumann U87	(spot)	Violin 2
Audio-Technica 4050		Cello
Audio-Technica 4050		Clarinet
Schoeps MK6		Flute
DPA 4011 (L)	cardioid	Group
DPA 4011 (R)	(setup LCR)	Group
Schoeps MK4 (C)		Group

Table 1: Reference of the microphones for the recording

The position of the LCR setup and the musicians was chosen prior to the recording to avoid any symmetry effect when placed in the centre. The final positions were determined after listening to the room acoustics. The exact position on the stage (Figure 3) and the order of the musicians were chosen to allow comfortable and optimal communication during the recording.

Once the musicians were positioned, the bidirectional spot microphones were positioned so that one lobe of the directional pattern was aimed at the instrument. Bidirectional microphones are better at rejecting close instrument bleed. The disadvantage of these microphones is that they pick up as much from the front as from the back. Therefore, in any setup of source and spot microphones, the other four sources must be placed in the rejection area of the directional pattern. For the violins, one of the two directional patterns of the microphone was angled towards the floor, so that the other was directed towards the ceiling to avoid bleeding. Also, to avoid positions where acoustically symmetrical effects could be heard (based on the radiation directional pattern of the violin [Chaigne and Kergomard [6]]), the microphone was placed slightly off-centre in front of the body of the violin. For the cello, one of the radiation patterns of the spot was also directed towards the instrument, off-centre (see Figure 4).

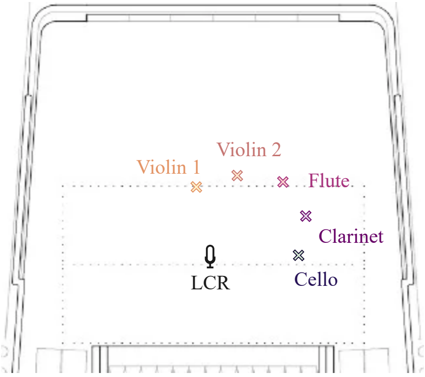


Figure 3: Positions of musicians and microphones in the room

For the position of the two wind instruments, the spot microphones were placed to record the sound radiating from the first holes of the body of the instrument.

Procedure

There were three steps in the recording. First, the five musicians played the original piece together. Then, each musician was asked to play his part of the piece alone, this audio was meant to be compared with the different results in the subjective test. Lastly, the musician played a chromatic scale in the pitch range used in the piece, with a metronome beat to ease the cutting that would be done to generate the dictionary. The recording setup was maintained the same throughout.

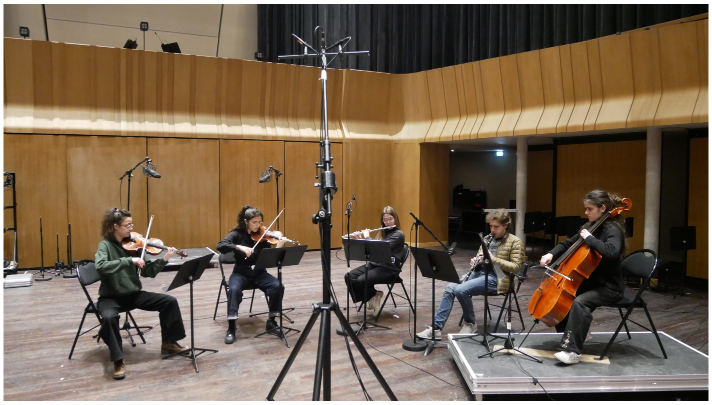


Figure 4: Positions of the musicians and microphones

An impulse response measurement was carried out in the auditorium. The goal is to use this room response information to improve the quality of source separation. Keeping the same setup without the musicians and using a speaker *JBL LSR2325P*, five impulse room responses were measured, one for each position of the musicians. These room responses were measured by reproducing a chirp audio ascending from $50Hz$ to $10000Hz$. The chirp audio signal was generated by the *Aurora* a plug-in [Farina [9]] for *Audacity*, that also generated the inverse filter use to compute the impulse room response by convoluting the audio recorded and the inverse filter.

Figure 5 presents the results obtained from the measurements using the analysis tool *Acoustical Parameters Calculation Module*. These results are directly computed from measurements made at the centre microphone of the LCR pair, for each source. To have a value for the room response we chose the position of the flute as the reference position. Since the microphone and the loudspeaker are not omnidirectional, the following results include the influence of the impulse responses and polar characteristics of both speakers and microphones.

All the recordings, including the room response, were made using *Ableton*, which was also used to mix the final audio.

Acoustic mixture

The reverberation time or decay time, called RT30, is the time required for sound to decay by 60 dB after the sound source has stopped [11]. It is a measure of how long sound energy continues to reverberate in a room after the sound source has ceased. RT30 is often used to evaluate the acoustic quality of a space, as it can affect speech intelligibility, music clarity, and overall sound quality.

EDT stands for Early Decay Time. It is defined as the time it takes for sound in a room to decay by 10 dB after the sound source has been turned off [11]. Unlike RT30, EDT focuses on the initial decay of sound in a room, rather than the overall decay. EDT is used to describe the clarity and definition of sound in a room. Rooms with lower values have clearer and more distinct sounds, while rooms with higher values have more diffuse and less clarity.

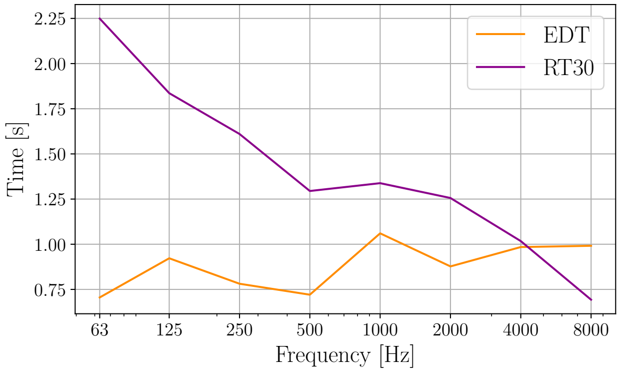


Figure 5: EDT and RT30 of the auditorium by frequency bands

Both indices are used to characterise the response of the room to an acoustic source (such as musicians playing). Information about the room response permits, together with the positions of the musicians, to describe the acoustic mix occurring during the recording. First, values of the RT30 (Figure 5) indicate the delay for the reverb to arrive after the sound is emitted. Then, EDT values (Figure 5) give the delay when the reverb sound is under 10dB from the direct sound. A difference of 10dB between two sources emitting at the same time, tends to make the quietest source not audible. Therefore, the impact of the reverb isn't audible during moments when two or more musicians are playing together with a recording microphone close to the source (under 3 meters). Only frequencies over $4kHz$ have an EDT index higher than the RT30.

3.4 Mixing

In order to obtain a recording of standard quality, post-production steps were carried out on the *Ableton Live* software to obtain from the sound recording a mixed and mastered sound in stereo. The audio processing chain (presented in figure Figure 6) consisted of three steps. First, the signals captured during the recording went through a processing chain including the following elements: a delay, an equalizer, a compressor, and a gain. The next step was to convert the eight monophonic signals into a stereo audio signal with panning. Finally, this stereo signal passes through a process-

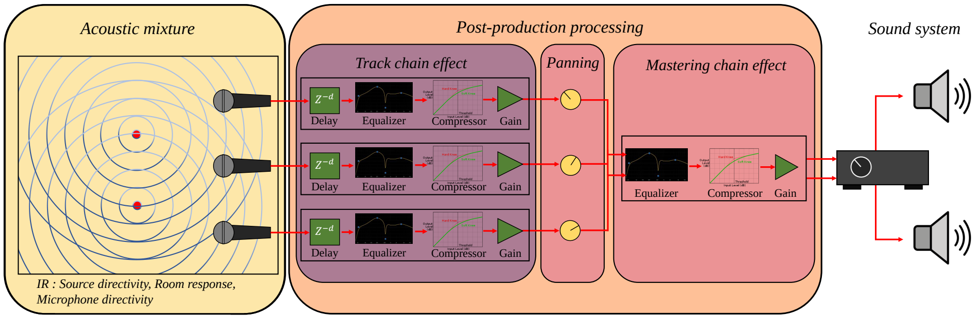


Figure 6: Audio processing chain

ing chain for mastering, including an equalizer, a compressor and a gain.

The monophonic tracks used at the beginning of the chain were the tracks of the eight microphones of the recording session in the auditorium, before any post-processing or trimming.

Mixing for the three tracks of the sound field microphones LCR, consisted of the same effects regarding the filtering. However, adjusting the volume of the centre microphone is different. Automation on the volume level of the LCR microphones has also been added, the goal being to reduce the noise level when the instruments are silent. No dynamic compression is used in this processing step. The aim being to limit the processing of the LCR microphones as much as possible, the added value of the mix will be essentially provided by the modifications on spot microphones.

The signals of spot microphones were time shifted to compensate for natural delays caused by their distance from the LCR microphones. Those signals are then processed through equalizers and compressors. Filters are adjusted to highlight certain characteristics of the instrument in its spectrum to enhance the desired frequency components of its sound. The threshold of the compressor is set to decrease the crest factor for high playback levels and to ignore the influence of low playback ones. When all the musicians play loudly, more emphasis is put on the sounds captured by the LCR than the auxiliary microphones. The level of the auxiliary microphones was adjusted by listening to the balance of each musician in the ensemble recorded by the couple and heard during the recording session.

Processing all monophonic tracks to one stereophonic audio included some panning. The musicians have been positioned considering their layout as being in a single plane, with a single level of depth. Concerning the stereophonic panoramic rendering, the right and left signals of the LCR pair were placed at their respective ends, [13]. The central microphone was then placed in the centre of the panoramic space. The spot microphones were placed perceptually from right to left, taking as reference the stereophonic sound of the LCR microphones. The stereo tracks then go through a mastering effect chain consisting of an equalizer, a compressor, and a gain. The equalizer is used to adapt the spectral balance. The compressor is used to reduce the overall dynamic range in order to reduce the differences in amplitude between the *piano* section of the piece and *forte* passages. Finally, a gain is applied to the signal to normalize its volume over the entire available dynamic range.

3.5 Evaluation

The heterogeneous nature of the available data meant that different manners of evaluation were implemented so as to obtain a more complete understanding of the quality of separation.

3.5.1 Objective Evaluation

Firstly, the classical SOTA objective metrics for source separation (SAR, SIR, SDR, and SI-SDR) were implemented for the algorithm results of the algorithm on the simulated data using implementations based on Scheibler et al. [27] and Roux et al. [24]. The artificially generated acoustic mixtures were separated and compared to the original stems as targets.

When attempting to evaluate the chamber music data, we were confronted by the issue of not having a recording of each individual instrument playing at the same time but isolated from the other instruments, which would be our target when calculating the objective metrics. We did have the spot microphones, but as expected, they presented a lot of bleeding from the other instruments because of the way they were intentionally set up. We also had recordings of each musician playing alone, but they were all at different tempos and any time change, even a minimum shift, greatly affects the objective metrics. Because of this, we decided to implement a subjective evaluation technique.

3.5.2 Subjective Evaluation

We decided to implement a MUSHRA-type crowd-sourced evaluation of the separation, via a web platform based on the *web-MUSHRA* platform by Schoeffler et al. [28]. The interface can be found in the Appendix or on our website. The test was hosted on private servers and was open to the general public for five days. The requirements were that the evaluator have a laptop, personal computer, or smartphone with a stable connection, headphones, a moderately quiet environment, and 15 minutes to dedicate to the test.

The workflow of the evaluation interface was as follows: When accessing the website, each evaluator was randomly directed to one of the five instruments that was separated and the whole test was done on that instrument. The test was introduced with a brief explanation of the context and by stating the previously specified evaluation conditions, the evaluator would then advance to a consent page that stated data privacy conditions to which the evaluator had to consent in order to perform the test, and the final step before beginning the test was a volume setup page. The test did not give the option to go back once a step (configuration step or evaluation step) was completed.

The test consisted of the evaluation of the quality of five audios, while focusing on four different factors in the following order: the overall quality, the interference, the distortion, and the artefacts. The five audios were used in the four cases, though their order was randomized and their condition was hidden. The five tracks of the test were :

- **a reference :** the recording of the corresponding musician playing by themselves, recorded on the spot microphone
- **an anchor :** created by adding distortion and equalization to the recording of the spot microphone
- **the separation on the instrument's spot microphone before processing the stems**
- **the separation on the instrument's spot microphone after processing the stems** with the mix's volume adjustment, equalization, and compression
- **the separation on the centre microphone** before processing

The objective is to be able to study the effect of the mixing process on our separation algorithm, as well as the effect of the spatial disposition of the microphones and the acoustic mixing in the recording room.

The first characteristic that was asked was to evaluate the overall quality ("focus on the overall quality of the conditioned audios and judge any and all detected differences between the reference and the conditioned."). The scale from 0 to 100 was explained as "0 means that the conditioned audio is much worse than the reference audio and 100 means that the conditioned audio is equal or better than the reference audio", and the evaluators were asked to "Please take your time to listen (and re-listen) and rate the audios". The interface only allows rating a track while it is playing and tracks are played in loop until stopped. A specific segment can be selected by the evaluator to listen to more closely. The waveform shown in the graphic interface is that of the reference track.

After rating the overall quality, an informative page was shown, explaining and defining the three characteristics that the evaluator should now focus on while rating the quality.

The second characteristic was the interference, it was explained as "the amount of sounds coming from instruments that are NOT the main instrument" and the rating was clarified as "best score (100) should be given if no secondary instruments are heard, and the worst score (0) should be given if a secondary instrument is heard as if it was the main instrument". The definition of the third characteristic (distortion) was "clipping, or modification of the instrument's sound ("fuzzy", "growling", or "gritty" sounds, noticeable absence of high or low frequencies, timbre modification)", and finally the artefacts were defined as "unexpected/non-musical sounds that seem to have been generated artificially, or that don't belong to the instruments' domain".

A final page with a questionnaire asked the evaluator to specify age, gender, years of musical training, and an email address, before finishing the evaluation. The option of going on to evaluate another instrument was also given if the evaluator so wished.

4 Experiments and results

4.1 Separation with MUSDB18 dataset

As explained in subsection 3.2, variations of the main parameters F, K were explored to observe their impact on the source separation task. These tests were also performed to understand the role of the parameters and how they improve or degrade the separation. We also tested different source separation configurations, alternating the use of different microphones to be in the overdetermined, determined or underdetermined cases, so as to assess the performance of the algorithm in each case.

Main parameters of the algorithm: A first experiment was carried out focusing on the main parameters F, K and their impact on the source separation task, in order to understand their role and how they improve or worsen the separation.

For these experiments, we first decided to set the parameter $K = 32$, following Sekiguchi et al. [29] who tested different values for speech separation and show that optimal values are $K = 16$ and $K = 64$ for almost all cases. However, since the higher the value, the longer the computation time, we decided that $K = 32$ was a good starting point. We then varied the parameter F to find the optimal pair with this value of K.

Note: All the following values (in Table 2, Table 3, Table 4) result from the average of 5 runs of the algorithm per song and for 5 songs from *MUSEDB18* [Rafii et al. [23]].

	SDR	SI_SDR	SIR	SAR
$F = 512$	2.41	0.04	8.52	6.04
$F = 1024$	4.18	2.13	11.16	7.23
$F = 2048$	3.88	2.49	11.65	7.23
F = 4096	**4.27**	**2.99**	**13.14**	**7.6**

Table 2: Impact of F on the source separation.

We then set $F = 4096$ (the optimal value for $K = 32$) and varied the parameter K to observe how the separation performance was affected. The number of iterations was set to 300.

	SDR	SI_SDR	SIR	SAR
$K = 2$	1.05	-3.19	8.01	5.34
$K = 4$	2.06	-0.54	9.25	6.04
$K = 8$	3.41	1.68	11.47	7.04
$K = 16$	4.15	2.58	13.16	7.47
K = 32	**4.27**	**2.99**	**13.14**	**7.6**
K = 64	**4.18**	**3.02**	**13.12**	**7.41**
$K = 128$	3.78	2.47	12.52	7.3

Table 3: Effect of K (n_components) on the source separation.

At the cost of a longer computation time, it seems that the higher the F, the better the separation (Table 2). However, the number of components K is more complex, although the intuition is the same as for the F parameter. Experience shows that a value of K that is too low imposes too many constraints on the algorithm, which then does not perform sufficient separation (Table 3). Conversely, a value of K that is too high gives the algorithm too much freedom and creates too many minima, which means that the global minimum will be harder to find and therefore will not result in the best possible separation (Table 3).

3 Setup Configurations: In this experiment, different source separation configurations were used to test the performance of the algorithm in each. The main parameters used were $F = 4096$, $K = 32$, and the number of iterations was set to $n\_iter = 300$. As discussed earlier, these parameters were considered sufficient and optimal to achieve convergence and the best separation with the algorithm.

	SDR	SI_SDR	SIR	SAR
overdetermined	**3.95**	**2.58**	**12.82**	**7.27**
determined	**3.44**	**-6.5**	**10.56**	**6.66**
underdetermined	-4.25	-13.17	0.19	1.69

Table 4: Performance of source separation in 3 configurations

In the Table 4, we observe that the algorithm does not work well in the underdetermined case. This result seems consistent, knowing that this is not the intended use of FastMNMF2. However, when the number of microphones is equal to or greater than the number of sources, the algorithm performs well and gives a clear separation of the different sources at each microphone.

Throughout the rest of this paper, experiments have been carried out with the main parameters $F = 4096$, $K = 32$, $n\_iter = 300$, and in the overdetermined case, as the results with FastMNMF2 are best in this configuration.

4.2 FastMNMF2 on real data

We were also able to test our algorithm on the real data of our recording session. Thanks to our configuration and the control we had over the various aspects (number of microphones, musical repertoire, mixing/mastering information), it was possible to understand and test the limits of the algorithm when modelling the recording process of a classical piece of music.

The first experiment we did was to test the behaviour of a blind separation in the different configurations: overdetermined, determined and underdetermined. To keep the MUSHRA test short enough, we started by evaluating some separations ourselves. As we observed with the experiment on MUSDB18 (Table 4), the overdetermined configuration performs well, while the others struggle to separate the different sources. So the only configuration that really performed a source separation task was the overdetermined one. The resulting audio can be found on our GitHub page.

We also tested the algorithm at different stages of the mixing process to understand which part of the mixing has the biggest impact on the performance of the separation and whether it can be improved (in blind separation). In particular, we found that compression and EQ effects tended to make the separation more complex while adjusting the volume between the microphones greatly improved it. Finally, we were surprised to find that the separation in the final mix was significantly better for some instruments, which we will analyse further in subsection 4.3.

4.2.1 With prior information in the dictionary

We used the dictionaries constructed from the scales played by the musicians to inform our algorithm. The two use cases of the dictionaries described in subsubsection 3.1.2 involving $\mathbf{E}$ and $\mathbf{W}$, proved ineffective because the former required $\mathbf{E}^{+}$ to be non-negative, but this requirement is not met regardless of the rank of $\mathbf{E}$. So we could not use Equation 21 and Equation 22.

Instead of using a split version, we constrained $\mathbf{W}$ itself with the dictionaries. We forced the number of bases K to be equal to the number of notes in the dictionaries (78 in our case). For a given number of iterations, $\mathbf{W}$ was fixed and equal to the dictionaries. We tried freezing it for different numbers of iterations without success. When freezing $\mathbf{W}$ for 20, 40, 100, or 300 iterations (out of 300), the timbre of one of the two instruments was too distorted, while the separation was not as precise as in the blind separation. We speculate that this is due to an overly constrained model (as $\mathbf{W}$ is less updated) and a lack of variety in the dynamics in the dictionaries.

4.3 Subjective Evaluation

The online MUSHRA platform was active for five days and received 69 evaluations from 63 evaluators. The average age of evaluators was 32.29 years, with the youngest being 21 years old, and the oldest 72. The average years of musical training was 10.44 years, with 16 people having less than 5 years of musical training.

During the test, participants rate the audio quality of different audio recordings compared to a reference. The MUSHRA methodology implies that in the audios to evaluate, participants rate a hidden reference and an anchor. These results allow us to sort the participant and exclude any incoherent responses from the results; if participants rated the hidden reference less than the other four audios, or if they rated the anchor at the maximum score, their answers were removed. In the Figure 7, comparing the figure with all evaluators and the figure without incoherent responses, the anchor ratings are generally lower, the higher rating goes down and the standard deviation is also less. In addition, this test was designed for participants with a trained ear for music listening. A comparison of the ratings for the anchor audios shows that when participants with less than 5 years of musical training are excluded, the rating of the anchor for overall quality descends. Looking at the Figure 7, it can be seen that without the inexperienced participants, the higher rate and the standard deviation also go down.

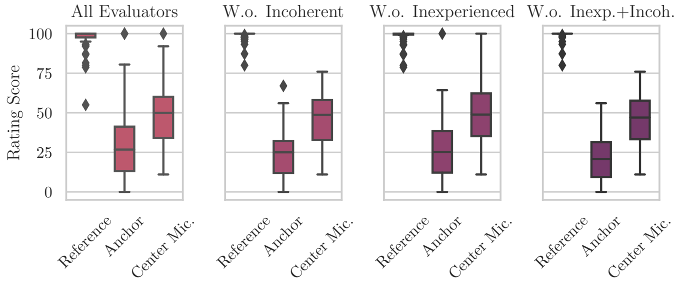


Figure 7: MUSHRA results: evaluators choice

Excluding participants with incoherent answers and inexperienced participants, the ratings of the anchor and the reference are more consistent. The standard deviations are lower in both cases, and the higher quartile of the anchor ratings becomes lower than the lower quartile of the centre microphone ratings. In Figure 7 only the rating of the centre microphone is compared to the anchor, because these are the lowest ratings of the different microphones (centre microphone, spot microphone and spot microphone with processing), as can be seen in Figure 8.

For the following observation, the results are calculated for the participants excluding those with incoherent answers and those without at least five years of musical training. This new subset consists of 41 evaluations from 39 evaluators with an average age of 31.02

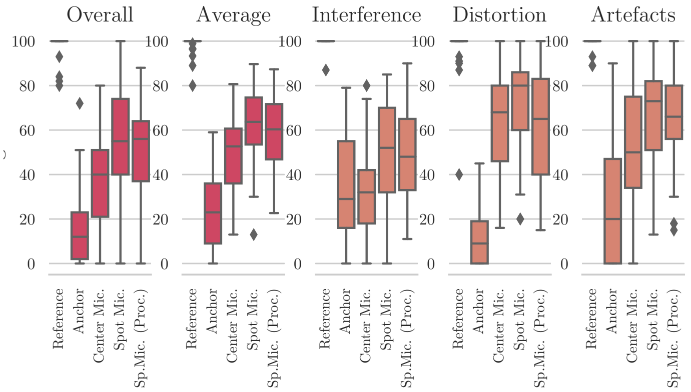


Figure 8: MUSHRA results: mean across instruments. Average is mean across Interference, Distortion & Artefacts

years old (Oldest: 58. Youngest: 21.), and an average of 14.54 years of musical study.

Firstly, we have chosen to focus on the results for the different tests, considering the average across all instruments, in order to assess the global capacity of our model. The Figure 8 compares the statistical scores for overall quality, quality with a focus on interference, quality with a focus on distortion, quality with a focus on artefacts and the average of these scores with a focus on characteristics. This average could be seen as a "proxy" for the overall quality, and it's important to see that both the overall and its "proxy" follow the same pattern, but the average scores slightly higher throughout, which can be analysed as the fact that the perceived quality is not covered by the traditional characteristics used in the objective metrics.

These results confirm that the anchor is rated worse than the other tracks, the separation at the centre microphone, the separation at the spot microphone and the separation at the spot microphone after mixing. In most of the results where all the instruments were responded to, the centre microphone separation was judged to be of inferior or equal quality.

A surprising result comes from comparing the rating of the quality of separation on the spot microphone before and after processing. Intuitively, adding filters, delays and other effects to most of the tracks used for multichannel separation would degrade the results. However, none of the results obviously show a significant difference in quality between the two audios of the spots.

These conclusions should be treated with caution. Firstly, we only have a limited number of participants. Secondly, the Figure 8 includes all five separations of the instrument as being of equal quality, but when listening to the audios and considering the disposition of the recording, the quality of the separation is not the same, nor is the acoustic mixture, the distance to the spot microphone or to the other musicians. The conditions are therefore not the same and the comparison is not fully possible.

Looking at the Figure 9, the ratings for each source show noticeable variations. For example, the separation quality of the cello is rated better with the spot microphone without effects, while the best result for the flute is obtained with the spot microphone with effects. This result is not surprising when you consider that the equaliser for the cello had a strong effect on the timbre of the instrument, whereas the flute was left almost untouched by the equaliser, although its volume was greatly reduced.

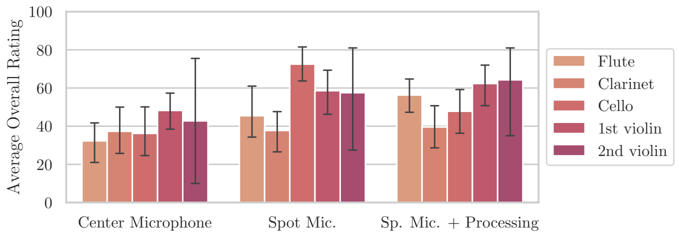


Figure 9: MUSHRA results: average overall ratings per instrument

The Figure 9 also shows that the source separation ratings for each microphone are always lower for the clarinet than for the cello and the two violins.

Contrary to our first hypothesis, adding a modification for mixing does not drastically worsen the quality of source separation. A more detailed study of each modification and effect added to the tracks would be necessary to judge the degradation of the source separation after the addition of processing.

Conclusion

Our work evaluates the performance of FastMNMF2 in a real acoustic recording scenario. The influence of microphone type and position has been studied in order to find an optimal arrangement that allows to determine the limits of the algorithm in this situation and during the mixing process. The algorithm was tested in blind (without *a priori* information) and constrained (with *a priori* information) situations. Based on our listening to the tracks, the use of *a priori* information (dictionary) in our implementation did not help the separation and introduced distortion.

Subjective evaluation showed great value in assessing the separation algorithm, proving that good separation quality was achieved. It also provided interesting insights, such as the fact that contrary to expectations that post-recording effects would degrade the quality of the separation, with careful mixing and good use of the effects chain, the quality of the separation with the algorithm in a blind situation is neither improved nor degraded, as explained in the subsection 4.3.

Future directions include improving the performance of the algorithm, both in a blind and in constrained situations. A better implementation of the dictionaries (e.g. changing the update rules) or the use of more complex dictionaries can be a good start. Similarly, informing the algorithm about the acoustic mixture that occurs during the recording should improve the performance of the algorithm. A possible approach would be to constrain the algorithm by adding the impulse room response as *a priori* information to inform the algorithm. Further tests on the mixing process could also help to clearly understand how each effect affects the separation, which looks promising for further research.

Regarding the evaluation of our algorithm, we can say that in order to compare our work more accurately with other algorithms, objective evaluations can be performed on several algorithms, such as ILRMA and AuxIVA, using the MUSDB18 dataset. It might also be possible to develop different sound recording setups, such as using loudspeakers in a concert hall or contact microphones on the instruments, to try to obtain a "ground truth" that could be used for objective evaluation.

Acknowledgments

We would like to express our sincere gratitude to Benoit Fabre, Professor of Musical Acoustics and Sound Techniques at CRR93, and Teacher and Researcher at the Institut ∂'Alembert of the Sorbonne University for support throughout the project. Special thanks for his invaluable assistance in recording and mixing the audio content used in this study. The smooth running of the audio recordings was made possible by Prof.Fabre's in-depth acoustic knowledge and expertise, and the research team was very grateful for his willingness to share his knowledge and experience. He has made significant contributions to this project and we are grateful for his kindness, wisdom and support throughout the process.

We would also like to thank Mathieu Fontaine, Professor and Researcher in the ADASP group at Ecole Télécom Paris, for his invaluable help and knowledge in the field of source separation. His contributions to this project have been instrumental in achieving the results presented in this paper. In addition, we would like to thank him for his insightful advice on the writing of this paper and for his help in troubleshooting the webMUSHRA platform, both of which greatly facilitated the experimental work described in this study.

Ines Lassègue kindly gave us permission to record her composition for use in this study, and for this we would like to express our sincere gratitude. This project would not have been successful without her musical knowledge and cooperation. We would also like to express our sincere gratitude to all the musicians who contributed to this project, including Ariane Alix, Axelle Renaux, Yonathan Navia and Anaïs Dumond-Gomez, for their exceptional dedication and professionalism in recording and performing the musical pieces. The excellent results presented in this work could not have been achieved without their commitment and diligence.

We are very grateful to the CRR93 Aubervilliers Conservatoire for their support and assistance throughout this project. In particular, we would like to thank Catherine Navarro, who was crucial to the success of the recording session in the auditorium. The recording would not have been possible without her invaluable help and knowledge. We are incredibly grateful to the Conservatoire for their willingness to work with us on this project and for kindly providing the necessary resources and support.

References

[1] Anastasios Alexandridis, Anthony Griffin, and Athanasios Mouchtaris. Capturing and reproducing spatial audio based on a circular microphone array. *Journal of Electrical and Computer Engineering*, 2013, 01 2013. doi: 10.1155/2013/718574.

[2] Bruce Bartlett and Jenny Bartlett. *Practical Recording Techniques: The step-by-step approach to professional audio recording*. Routledge, 2016.

[3] Recommendation ITU-R BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems. *R BS.*, 2014.

[4] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998. doi: 10.1109/5.720250.

[5] Mark Cartwright, Bryan Pardo, Gautham J. Mysore, and Matt Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 619–623, 2016. doi: 10.1109/ICASSP.2016.7471749.

[6] Antoine Chaigne and Jean Kergomard. *Acoustique des instruments de musique*. Belin, 2008. URL `https://hal.science/hal-00455011`. Pages: 712 pages.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

[8] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010-09. ISSN 1558-7924. doi: 10.1109/TASL.2010.2050716. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[9] Angelo Farina. simultaneous measurement of impulse response and distortion with a swept-sine technique. *journal of the audio engineering society*, february 2000.

[10] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011. doi: 10.1162/NECO_a_00168.

[11] Anders Gade. Investigations of musician's room acoustics conditions in concert halls, part ii. *Acta Acustica united with Acustica*, 69:249–262, 01 1989.

[12] Enric Gusó, Jordi Pons, Santiago Pascual, and Joan Serrà. On loss functions and evaluation metrics for music source separation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 306–310, 2022. doi: 10.1109/ICASSP43922.2022.9746530.

[13] Christian Hugonnet and Pierre Walder. *Théorie et pratique de la prise de son stéréophonique*. Eyrolles, 1994.

[14] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. *Determined Blind Source Separation with Independent Low-Rank Matrix Analysis*, pages 125–155. Springer International Publishing, Cham, 2018. ISBN 978-3-319-73031-8. doi: 10.1007/978-3-319-73031-8_6. URL `https://doi.org/10.1007/978-3-319-73031-8_6`.

[15] Istvan Lang. The effect of the recording process on european classical music. *The World of Music*, 27(3):114–121, 1985. ISSN 00438774. URL `http://www.jstor.org/stable/43562726`.

[16] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, 1999-10. ISSN 1476-4687. doi: 10.1038/44565. URL `https://www.nature.com/articles/44565`. Number: 6755 Publisher: Nature Publishing Group.

[17] Ethan Manilow, Prem Seetharman, and Justin Salamon. *Open Source Tools & Data for Music Source Separation*. https://source-separation.github.io/tutorial, October 2020. URL `https://source-separation.github.io/tutorial`.

[18] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, Fabian-Robert Stöter, Alexandre Défossez, Minseok Kim, Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk. Music demixing challenge 2021. *Frontiers in Signal Processing*, 1, 2022. ISSN 2673-8198. URL `https://www.frontiersin.org/articles/10.3389/frsip.2021.808395`.

[19] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Deep neural network based multichannel audio source separation. *Audio Source Separation*, pages 157–185, 2018.

[20] Nobutaka Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 189–192, 2011. doi: 10.1109/ASPAA.2011.6082320.

[21] Alexey Ozerov and Cédric Fevotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010-03. ISSN 1558-7924. doi: 10.1109/TASL.2009.2031510. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[22] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, 2012-05. ISSN 1558-7916, 1558-7924. doi: 10.1109/TASL.2011.2172425. URL `http://ieeexplore.ieee.org/document/6047568/`.

[23] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. MUSDB18-HQ - an uncompressed version of MUSDB18, 2019-08-01. URL `https://zenodo.org/record/3338373`. Type: dataset.

[24] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

[25] Saurjya Sarkar, Emmanouil Benetos, and Mark Sandler. Ensembleset: A new high quality synthesised dataset for chamber ensemble separation. *Proceedings of the 23rd ISMIR Conference*, 2022.

[26] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982, 2013. doi: 10.1109/TASL.2013.2239990.

[27] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355, 2018-04. doi: 10.1109/ICASSP.2018.8461310. ISSN: 2379-190X.

[28] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. webMUSHRA — a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1):8, 2018-02-05. ISSN 2049-9647. doi: 10.5334/jors.187. URL `http://openresearchsoftware.metajnl.com/article/10.5334/jors.187/`. Number: 1 Publisher: Ubiquity Press.

[29] Kouhei Sekiguchi, Yoshiaki Bando, Aditya Arie Nugraha, Kazuyoshi Yoshii, and Tatsuya Kawahara. Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2610–2625, 2020. ISSN 2329-9304. doi: 10.1109/TASLP.2020.3019181. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[30] Neta Spiro and Michael F. Schober. Discrepancies and disagreements in classical chamber musicians' characterisations of a performance. *Music & Science*, 4:20592043211011091, 2021. doi: 10.1177/20592043211011091. URL `https://doi.org/10.1177/20592043211011091`.

[31] Ron Streicher and Wes Dooley. Basic stereo microphone perspectives-a review. *Journal of the Audio Engineering Society*, 33(7/8): 548–556, 1985.

[32] Günther Theile. multichannel natural recording based on psychoacoustic principles. *journal of the audio engineering society*, february 2000.

[33] Nantho Valentine. L'enregistrement des cordes, July 2020. URL `https://fr.audiofanzine.com/prise-de-son-mixage/editorial/dossiers/l-enregistrement-des-cordes.html`.

[34] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006. URL `https://hal.inria.fr/inria-00544230`.

[35] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[36] Wen Zhang, Parasanga N. Samarasinghe, Hanchi Chen, and Thushara D. Abhayapala. Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences*, 7(5), 2017. ISSN 2076-3417. doi: 10.3390/app7050532. URL `https://www.mdpi.com/2076-3417/7/5/532`.

[37] Sławomir Zieliński, Philip Hardisty, Christopher Hummersone, and Francis Rumsey. Potential biases in MUSHRA listening tests. *Audio Engineering Society - 123rd Audio Engineering Society Convention 2007*, 2, 2007-01-01.

Appendix A - Results Comparison

Comparing Results: To compare the performance of this algorithm with other existing algorithms, the average of each objective metric on the 4 sources has been taken into account. The overall SDR, overall SI-SDR, overall SIR and overall SAR values were calculated for 10 songs. The average scores were calculated using only the four spot microphones.

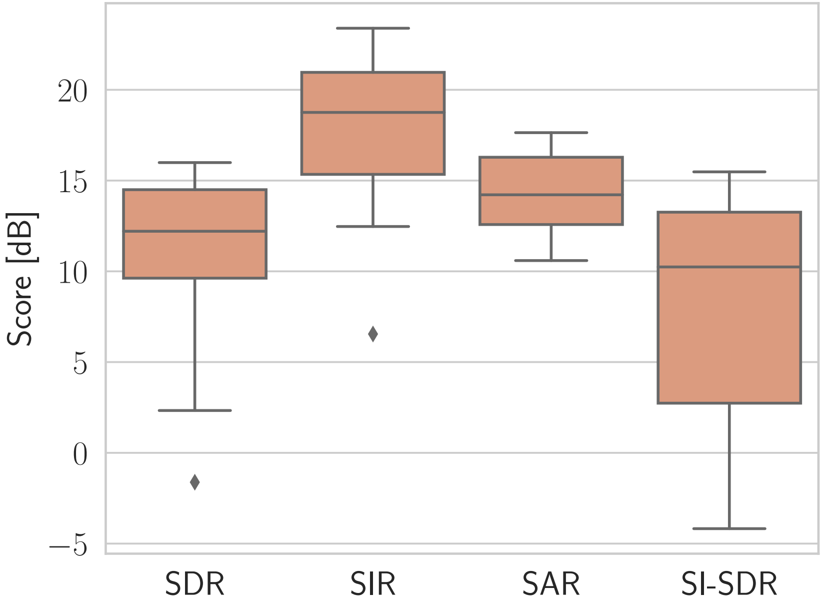
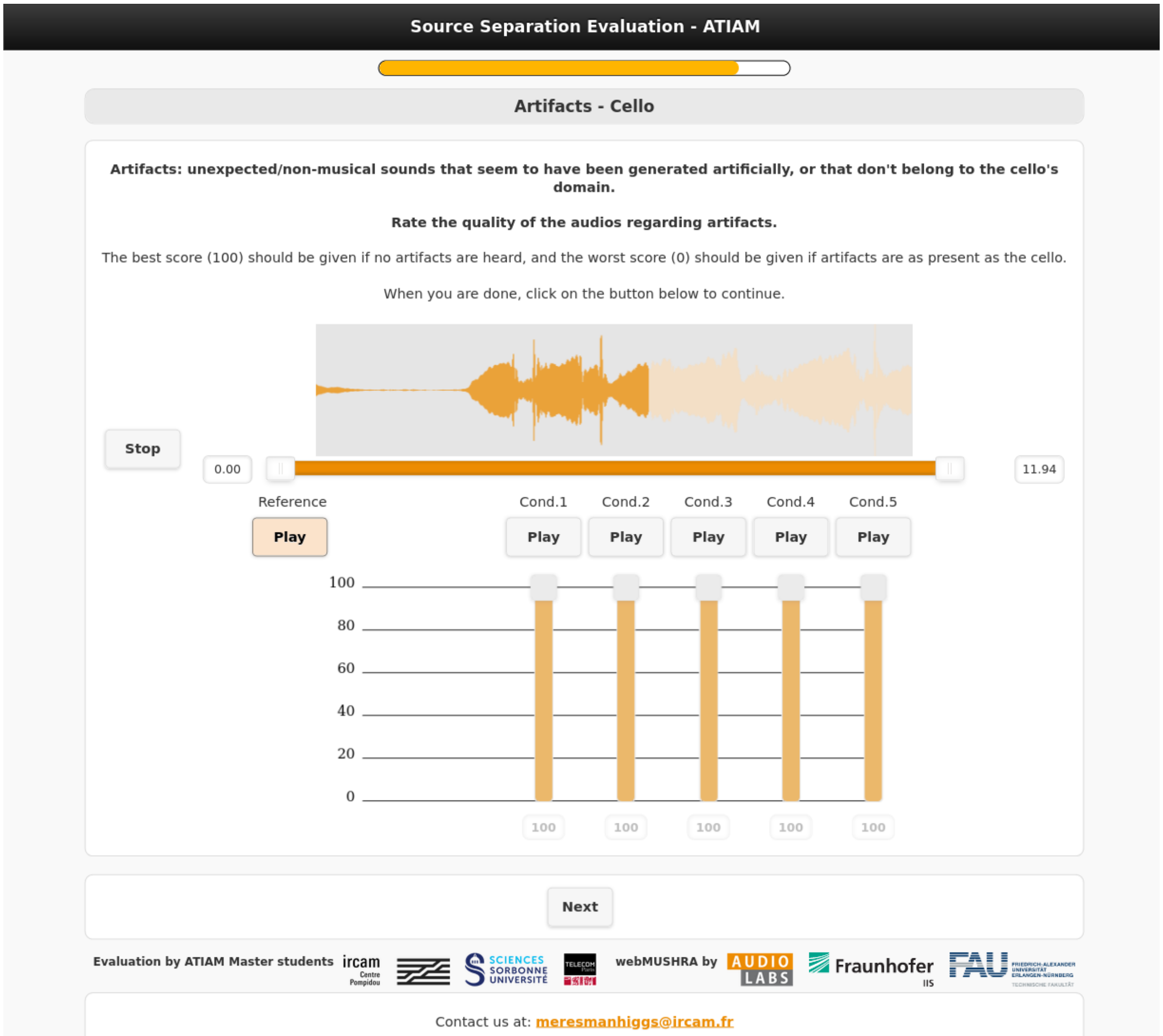


Figure 10: Objective evaluation on overall scores

The range of overall scores is wide across the songs, these results are due to the fact that the 4 sources are not necessarily played on all the songs. The calculation of these scores on empty tracks is very low (negative), which is why the overall scores can be strongly influenced by songs with only one voice and one guitar.

In order to compare our work with other algorithms with more accuracy, objective evaluations on multiple algorithms such as ILRMA and AuxIVA, with the MUSDB18 dataset, could have been performed.

Appendix B - WebMushra



Artifacts: unexpected/non-musical sounds that seem to have been generated artificially, or that don't belong to the cello's domain.

Rate the quality of the audios regarding artifacts.

The best score (100) should be given if no artifacts are heard, and the worst score (0) should be given if artifacts are as present as the cello.

When you are done, click on the button below to continue.

Evaluation by ATIAM Master students

webMUSHRA by

Contact us at: **meresmanhiggs@ircam.fr**

Figure 11: Subjective evaluation interface

	Distortion			**Equalization**			
Instrument	**Clip [dB]**	**Drive**	**Makeup Gain**	**Inverse notch filter [Frequential Center[Hz]——Gain[dB]]**			
1st Violin	-5	80	50	63——20	2k——13	2.5k——13	
2nd Violin	-5	80	50	63——20	2k——13	2.5k——13	
Clarinet	-3	60	50	80——20	1.25k——14	1.6k——14	
Cello	-3	60	50	80——20	1k——1	1.25k——20	1.6k——1
Flute	-3	60	50	63——20	80——20	1k——20	

Table 5: MUSHRA anchor creation parameters

Appendix C - Subjective Evaluation

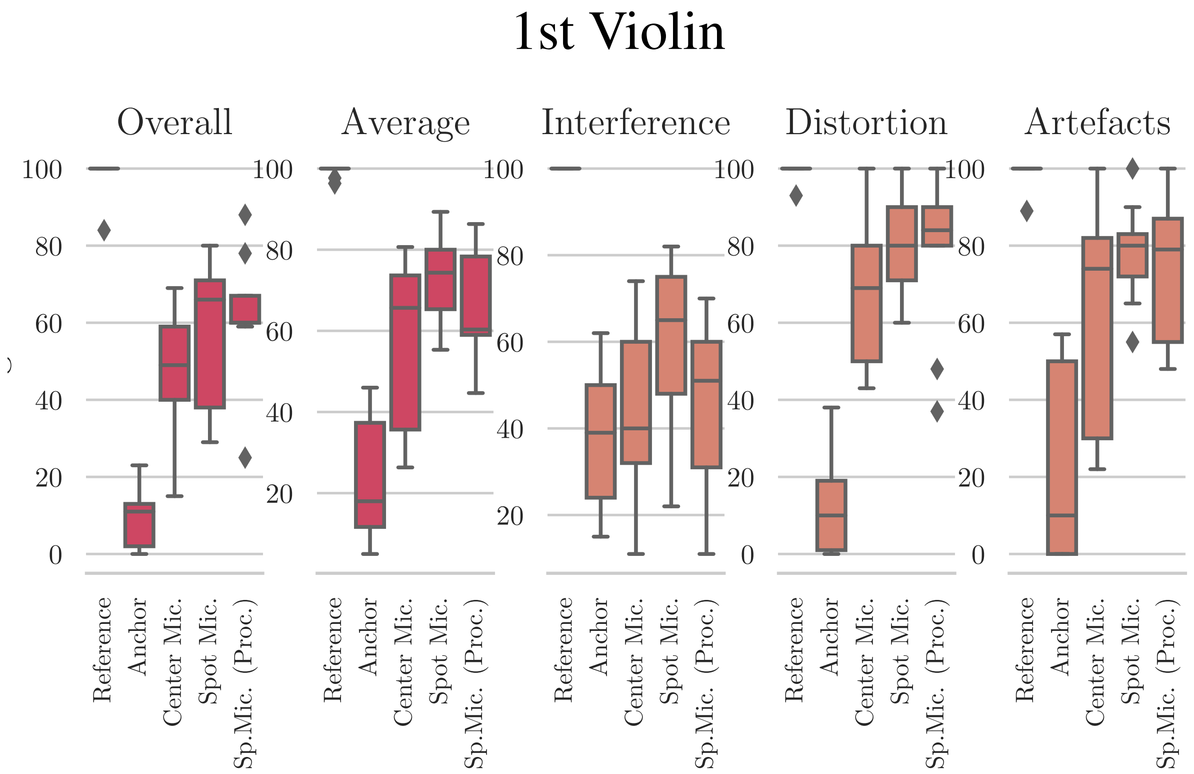


Figure 12: Boxplot of MUSHRA scores for individual instruments

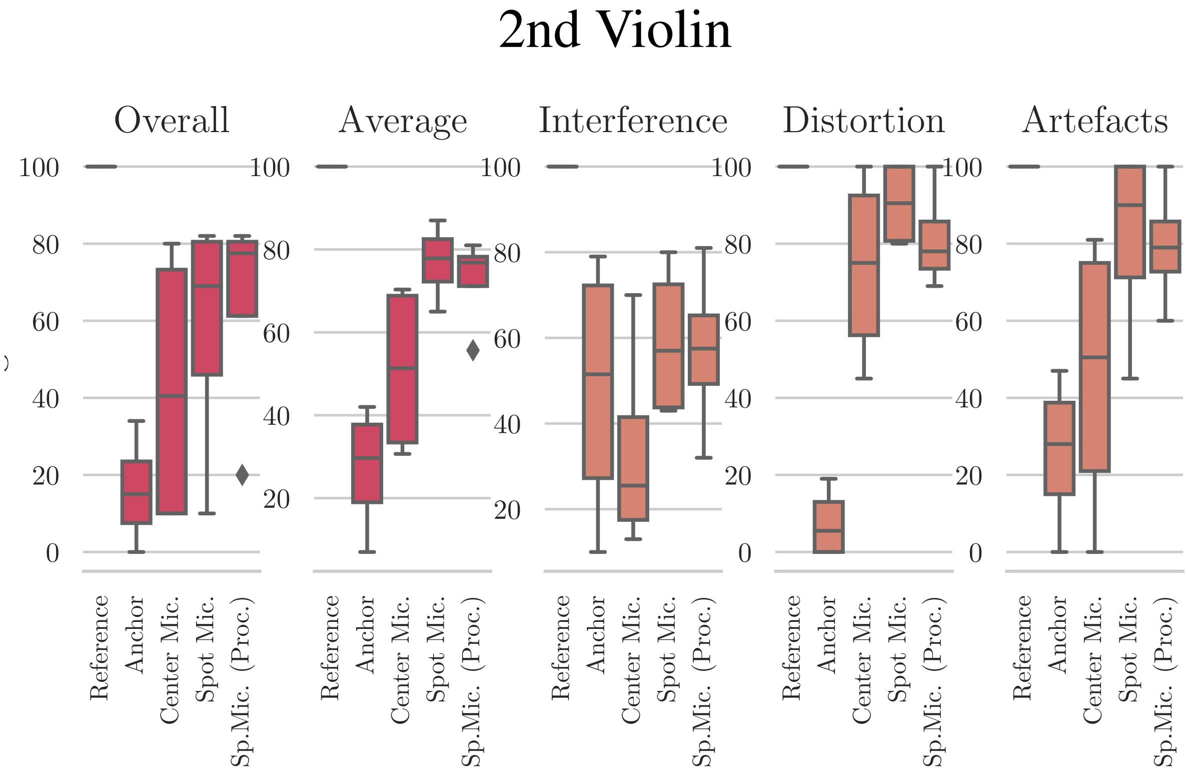


Figure 13: Boxplot of MUSHRA scores for individual instruments

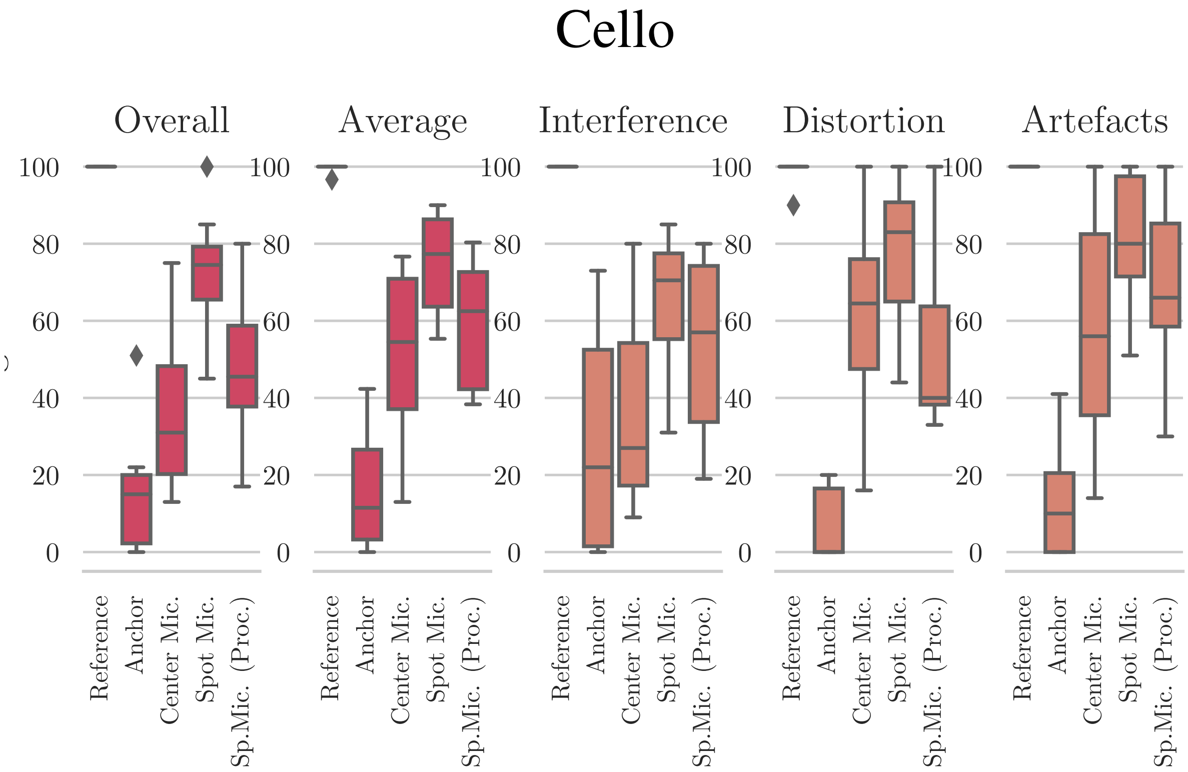


Figure 14: Boxplot of MUSHRA scores for individual instruments

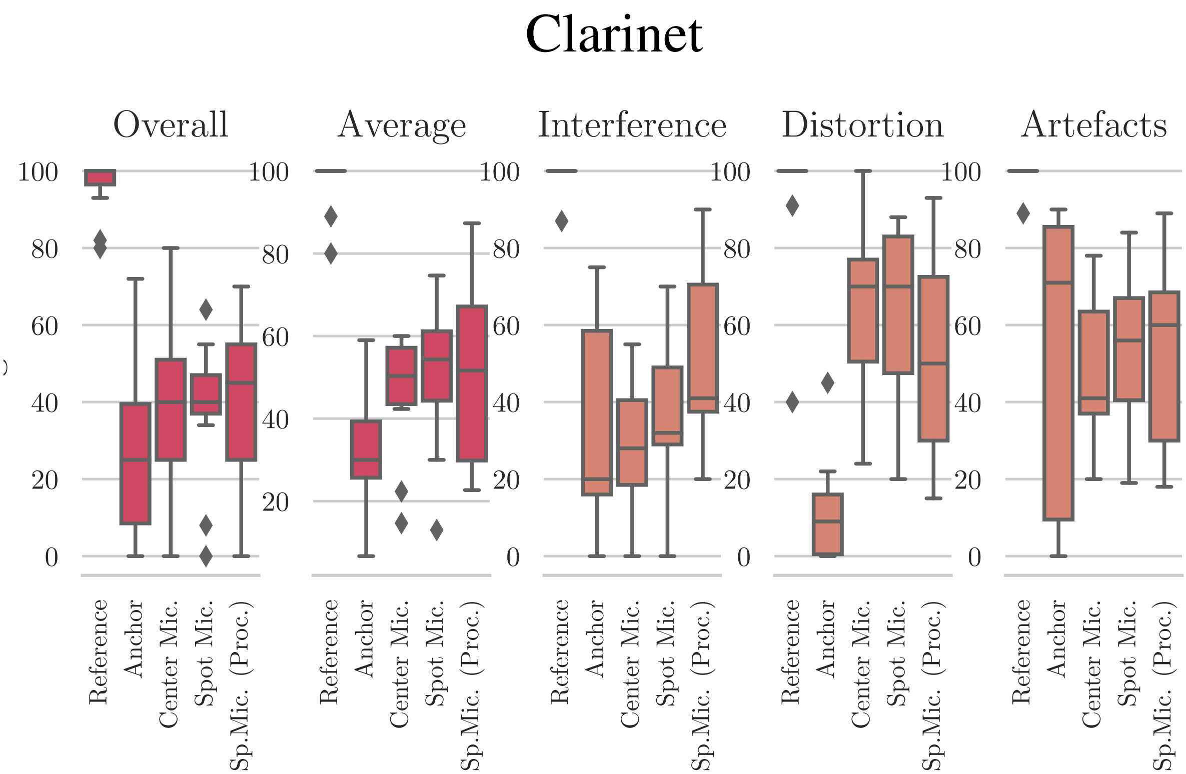


Figure 15: Boxplot of MUSHRA scores for individual instruments

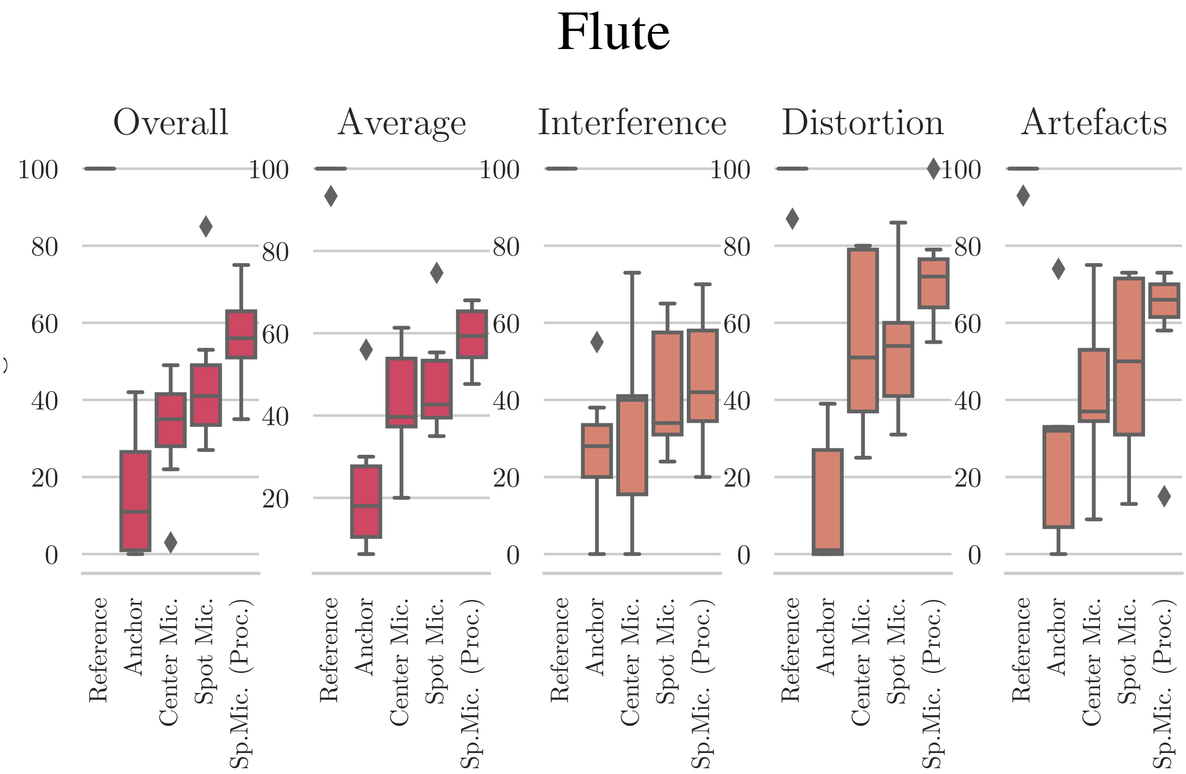


Figure 16: Boxplot of MUSHRA scores for individual instruments