PAM - Projets et Applications Musicales
Sound Recording and Audio Source Separation

Azal LE BAGOUSSE[1], Robin WENDLING[2], and Jiale KANG[3]

[1,2,3]M2 ATIAM, Département Pédagogie, IRCAM, Paris, France, [1]azal.lebagousse@ircam.fr, [2]robin.wendling@ircam.fr, [3]kang@ircam.fr

February 18, 2025

Abstract

This study investigates the impact of controlled recording conditions on audio source separation by integrating sound recording techniques and numerical signal processing methods. By strategically capturing musical performances with optimized microphone setups to obtain both mixtures and isolated instrument tracks, different levels of separation effectiveness can be achieved. Prior information, in the form of a dictionary of notes obtained by recording each instrument individually, can be incorporated to guide the separation process. Both traditional signal processing methods (Gaussian MNMF, FastMNMF2, ILRMA) and deep learning models (Demucs, Spleeter) were evaluated. Subjective (listening task) and objective (numerical metrics) methods were used for the evaluation of all algorithms. The results indicate that separation performance is strongly influenced by instrument playing techniques, recording conditions, and musical genre selection. Additionally, reverberation and leakage present challenges for separation. The study further analyzes the relationship between evaluation methods, the differences between algorithms, and the impact of input configurations on separation performance.

1 Introduction

Audio source separation involves extracting different tracks from a mixture of sound signals. In the musical context, this process isolates individual tracks corresponding to different instruments and vocals, with applications such as post-production, remixing, and karaoke. The recorded signal is the result of a complex process that begins with the vibration of the instrument and involves room response, microphone directivity, and possible mixing operations. This signal may contain multiple channels (stereo or multichannel), making the individual recovery of each source a real challenge.

This challenge becomes more manageable when constraints on algorithm parameters are incorporated, based on prior information of the recording process, such as microphone directivity, played notes, or room response. Documenting this informations enhances the robustness and quality of the separation.

The goal of this project is to understand how controlling the recording process can improve source separation. To achieve this, several steps will be undertaken: a recording session will first be set up and conducted, followed by the selection and implementation of appropriate source separation algorithms. Finally, numerical performance will be evaluated using both subjective and objective criteria.

1.1 Recording Session

First, the music genres were carefully selected : classical music and jazz were chosen as they are respectively one of the easiest and one of the hardest genres for sound source separation. Jazz has dense harmonies, overlapping timbres, and usually requires a greater variety of instruments that frequently blend and interact dynamically. Classical music has structured arrangements, distinct orchestration, and predictable dynamics. To get two different levels of separation effectiveness, both of those genres were selected.

Next, the instrument selection was defined: three instruments for the simpler classical configuration and five for the more challenging jazz configuration. To get as much diversity as possible, it was defined that the classical piece had to include a clarinet, a violin, and a piano, and the jazz arrangement had to include a vocalist, a saxophone, a piano, a bass, and percussion (a conga). The song selection was based on those instruments, chosen from the standard repertoire of both genre : *Fly me to the Moon* by B. Howard for jazz and *Trio No. 2 in E-flat major* by F. Schubert for classical, replacing the cello by the clarinet to be able to also study a wind instrument.

The recordings took place in the auditorium of the Aubervilliers Conservatory (CRR 93). In each configuration, the musicians were positioned in a manner that maximized physical separation, while ensuring that the performers remained within the desired auditory and visual range. The choice of microphones was guided by the musicians' placement and the microphones' directivities.

Further explanation will be given in Section 3.1.

1.2 Separation Algorithms

The selection of algorithms was made in accordance with the recorded material and the desire to achieve different levels of separation effectiveness. As the main methods, 3 signal processing algorithms were selected.

The first one was GaussMNMF [1], an extension of the classic NMF algorithms . This method was the first MNMF method to appear chronogically, which is the reason why it's tested first in this project. With this algorithm, the sources are modeled with non-negative spectrograms while it estimates their spatial properties. It

is effective for complex mixtures with reverberation, but is computationally expensive due to full-rank spatial covariance matrices.

Then, FastMNMF2 [2] was implemented. It is an improvement of the classic MNMF methods as it introduces diagonal approximations of spatial covariance matrices to speed up computation, while still modeling reverberation. FastMNMF2 was chosen instead of FastMNMF1 [2] as it further improves efficiency and separation by factorizing the source power spectrogram into a more structured model.

Finally, ILRMA [3] was implemented. It is a special case of rank-constrained FastMNMF where each source is modeled with a small number of spectral components. It assumes sources have low-rank representations, improving separation for harmonic signals, which is useful when doing musical sound source separation.

Additionally, two deep learning–based methods, Demucs [4] and Spleeter [5], were also evaluated using their open-source implementations. Demucs leverages time-domain convolution and transformer architectures for end-to-end separation, while Spleeter uses a U-Net-based convolutional neural networks (CNNs) for spectrogram masking. Unlike the signal processing algorithms, which depend on clear spatial and spectral models, these deep learning methods learn complex non-linear relationships from large datasets. This can result in more natural separations than other the signal processing algorithms implemented in this project, though their performance may be less high when the test data differs from the training domain.

More information about algorithms will be given in Section 3.2, and all results could be found on https://kjle.github.io/PAM-Music-Source-Separation/.

1.3 Evaluation

The performance evaluation was assessed with subjective methods and objective methods.

For the subjective evaluation, listening tests MUSHRA were conducted to assess the perceptual quality of the separated signals, following feedback from as many people as possible within the given time. For the objective evaluation, standard numerical metrics BSS were used to quantify the quality of separation based on how clear the sound is and how well unwanted noise is removed. This two-way approach gives a complete evaluation, combining human listening opinions with measurable data.

The detail of the evaluation will be given in Section 3.3.

2 State of the Art

2.1 Sound Recording

Sound recording plays a major role in sound source separation by shaping the spatial, tonal, and directional qualities of the captured signals. The clarity of these attributes directly impacts the separation process, as algorithms rely on well-defined spatial and spectral differences to distinguish sources. For example, precise microphone placement can reduce phase issues, improve localization, and isolate sources more effectively, facilitating cleaner separation in the end [6, 7]. Sound recording is not the first step to achieving great separation. It is one of the last steps because it must determine the type of signals to use to properly suit the algorithms and get clear results. But it is still the first concept to think about, as many different techniques and setups exist to get sound mix that can be effectively utilized.

There are three primary categories of microphone techniques relevant to sound source separation: coincident, near-coincident, and spaced configurations. Coincident techniques, like MS (Mid-Side) stereo, are ideal for achieving precise localization and offer excellent mono compatibility. This makes them a good choice for quiet and controlled spaces (like studios) where you can easily adjust how wide or narrow the stereo sound is during editing (easier sound balance) [6] . Near-coincident techniques, such as ORTF and NOS, combine intensity and time differences to create a natural sense of depth and openness while minimizing unwanted phase effects. These techniques work the best in moderately reverberant environments, offering a balance between spatial clarity and natural atmosphere [6, 8]. Spaced techniques, such as spaced bidirectional microphones, provide strong spatial separation and natural reverberation capture but require careful and precise placement to avoid phase irregularities. They are suited for large ensembles in well-controlled spaces, like for classical music orchestras in a theatre [6, 7]. Advanced setups, like the Optimized Cardioid Triangle (OCT) or 3/2-stereo configuration, can improve traditional techniques by integrating psychoacoustic principles [8]. These methods focus on keeping sound positions stable and creating a sense of space, especially for multichannel and surround sound setups. Similarly, ambisonic microphones capture full 3D spatial data, enabling immersive and flexible separation workflows in complex environments [7].

Ultimately, for the study, the choice of recording method depends on the specific requirements of the separation task. The techniques were chosen based on the objectives of the project, providing both signals for audio production and signals for different tests on separation algorithms.

2.2 Algorithms and Models for Audio Source Separation

Different algorithmic approaches exist for implementing audio source separation. These approaches vary according to the methods employed and are still evolving. The objective of this section is to describe these main approaches, the concepts they rely on, the methods for their implementation, and their primary use cases.

Before describing these approaches, it is essential to recall certain notations and assumptions used in this field.

Generally, suppose:

- $\mathbf{x}(t)$: the vector of M observed mixtures, where each component $x_m(t)$ corresponds to the signal of the m-th microphone;
- $\mathbf{s}(t)$: the vector of N unknown sources $\{s_n(t)\}$;
- $\mathbf{A}$ (or $\mathbf{A}(f)$ in the frequency domain): the mixing matrix that transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$.

The approaches differ according to the type of mixture. There are two main types:

- Instantaneous linear mixtures

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

- Convolutive mixtures

$$x_m(t) = \sum_{n=1}^{N} (a_{mn} \star s_n)(t) \tag{2}$$

2.2.1 Independent Component Analysis

Historically, the first approaches were developed in the framework of instantaneous linear mixtures. One such method is Independent Component Analysis (ICA) [9], which assumes the statistical independence of sources. The objective is to find a separation matrix $\mathbf{B}$ such that the separated signals:

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \tag{3}$$

are as independent as possible. This is achieved by maximizing an independence criterion, using statistical measures such maximum likelihood estimation.

In this framework, ICA methods applied separately to each frequency band suffer from permutation indeterminacies across frequencies. This led to the development of more recent approaches based on local Gaussian modeling [10, 11], where each time-frequency bin is assumed to follow a complex Gaussian distribution. Source estimation is achieved by minimizing the mean squared error between the estimated and true sources, often employing Wiener filtering, which relies on the covariance matrices of the sources.

2.2.2 Non-negative Matrix Factorization

Another approach is Non-negative Matrix Factorization (NMF) [12], which factorizes a spectrogram $\mathbf{V}$ into:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{4}$$

where columns of $\mathbf{W}$ represent spectral bases and rows of $\mathbf{H}$ their activations over time. NMF is widely used in music analysis to separate instrumental components or detect patterns such as piano notes. The problem is formulated as:

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} d(\mathbf{V}\|\mathbf{W}\mathbf{H}) \tag{5}$$

where $d(\cdot\|\cdot)$ is a $\beta-$divergence measure ($\beta = 2$ for Euclidean, $\beta = 1$ for Itakura-Saito, $\beta = 0$ for Kullback-Leibler, etc.).

For multichannel signals, NMF is extended to Multichannel NMF (MNMF) [1]. The spectrogram becomes a tensor, modeling both spectral content and spatial structure.

When microphone configurations and room conditions are documented, these data can be incorporated into MNMF to constrain the spatial covariance matrix and accelerate convergence. Due to the high computational cost of MNMF, Sekiguchi et al. introduced FastMNMF [13], which assumes joint diagonalization of spatial covariance matrices, reducing parameter count and improving convergence speed. Variants such as FastMNMF1 and FastMNMF2 also adapt to partially diffuse sources (whose energy spread across multiple directions rather than originate from a single point, which is often due to reflections and reverberation in a room), improving efficiency and separation quality.

Independent Low-Rank Matrix Analysis (ILRMA) [3] introduced by Ono is a simplified variant of FastMNMF that imposes a rank-1 constraint on each source's spatial covariance. This assumption significantly reduces the number of parameters to estimate, which in turn improves the algorithm's stability and accelerates convergence, particularly in scenarios where the sources' spatial characteristics are relatively simple.

2.2.3 Parameter Estimation

To estimate the matrix parameters, two common methods are Expectation-Maximization (EM) [14] and Multiplicative Updates (MU) rules.

The EM algorithm maximizes the log-likelihood of observations while considering latent variables in the E-step, followed by parameter updates in the M-step. Despite its flexibility, it is computationally expensive and sensitive to initialization.

The MU rules optimize the objective function in NMF by iteratively updating the basis matrix $\mathbf{W}$ and the activation matrix $\mathbf{H}$ through element-wise multiplication and division. Despite their simplicity and guaranteed convergence to a local optimum, the MU rules can be slow to converge and are sensitive to the initial values of $\mathbf{W}$ and $\mathbf{H}$, which may affect the quality of the final solution.

Alternative methods include Majorization-Minimization [15] and auxiliary function approaches [16].

2.2.4 Generalized Framework

Some efforts focus on unifying or modularizing these approaches. Ozerov, Vincent, and Bimbot proposed a general framework incorporating various constraints such as spatial rank, temporal structure, and directivity, releasing the Flexible Audio Source Separation Toolbox (FASST) [17] library.

2.2.5 Deep Learning Models

Recent advances in deep learning have led to the development of highly effective music source separation models. In this project, two state-of-the-art models, Spleeter [5] and Demucs [18], were implemented .

Spleeter [5] is a convolutional neural network (CNN) based method developed by Deezer. It operates in the time-frequency domain by first transforming the input audio into a spectrogram. A U-Net-like encoder-decoder architecture is then applied to predict source-specific masks, which are used to extract individual components such as vocals, drums, bass, and other instruments. Spleeter benefits from training on large-scale datasets, enabling it to generalize effectively across diverse musical genres. Its efficiency and ease-of-use have made it a popular choice for both real-time and offline separation tasks.

Demucs [18], on the other hand, adopts a time-domain approach to source separation. Developed by Facebook AI Research (FAIR), it is a model designed for end-to-end waveform-based separation. Built on an encoder-decoder convolutional architecture with skip connections and, in some variants, recurrent layers, Demucs directly processes the raw audio waveform. This design allows it to capture fine temporal details and transients more accurately, often resulting in outputs with fewer artifacts compared to spectrogram-based methods. By learning to decompose the waveform into its combined sources without relying on clear spectral representations, Demucs shows great performance for separation.

2.3 Evaluation

Evaluating the quality of separated sources is essential for assessing the separation algorithm performance and ensuring perceptual quality. Evaluation methods can be categorized into objective evaluation (quantitative metrics) and subjective evaluation (perception-based methods). This section will focus on an overview of these evaluation techniques, as well as their strengths and limitations.

2.3.1 Objective Evaluation

Objective evaluation metrics provide quantitative measures of the separation quality. They allow for efficient and low-cost performance measurement.

The most commonly used metric is Blind Source Separation Evaluation (BSS Eval) Metrics [19], which includes Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), Sources-to-Artifacts Ratio (SAR) and Sources-to-Noise Ratio (SNR). These four metrics are based on distortion decomposition between estimated source and target source, interference energy, noise energy, artifacts energy. Inspired by SNR, SDR measures the overall quality of the separated source; SIR assesses the suppression of undesired sources in the separated signal; SAR evaluates the amount of additional artifacts introduced by the separation process.

Since the original proposal of SDR, several issues with the metric have been discovered, including an easy way to boost one's scores by changing the amplitude scaling of source estimates. This prompted the proposal of a version of SDR that is not dependent on amplitude scaling, Scale-Invariant SDR (SI-SDR) [20].

While BSS Eval metrics provide a standard mathematical framework for source separation evaluation, they have limitations in capturing human perceptual quality. To improve this, Perceptual Evaluation methods for Audio Source Separation (PEASS) Metric [21] introduces a perceptually motivated assessment, based on auditory models to provide a more reliable evaluation of separation quality. PEASS metrics consist four perceptual scores: Overall Perceptual Score (OPS) reflects the overall perceptual quality of the separated signal; Target-related Perceptual Score (TPS) measures the extent to which the desired source has been preserved in the separated output; Interference-related Perceptual Score (IPS) quantifies audible influence to human for residual interference from other sources present in the separated signal; Artifact-related Perceptual Score (APS) assesses the presence of artificial distortions introduced during separation.

Since both BSS Eval and PEASS require the target signal and estimated mixture signal as inputs. This is a significant issue when no target audio is available. Inspired by Fréchet Inception Distance (FID), an alternative approach to a reference-less model, Fréchet Audio Distance (FAD) [22], was proposed. FAD compares statistics computed on a set of estimated signals to reference statistics computed on a large set of studio recorded music.

2.3.2 Subjective Evaluation

While objective metrics provide significant and reasonable results, subjective evaluation remains essential for assessing perceptual quality, as human perception is not always well captured by quantitative metrics.

Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [23] is a well-known testing method initially designed for measure the perceptual quality of audio codecs. It can present multiple versions of the same signal, including the original, processed, and degraded versions. Listeners provide scores based on their perceived quality. This method allows a detailed, fine-grained scoring, and enables to compare multiple systems simultaneously. It use a hidden reference and anchor, ensuring an objective calibration.

Another model for multi-stimulus testing, Audio Perceptual Evaluation (APE) was proposed by Brecht De Man et al [24]. The main difference between MUSHRA and APE is that APE encourages participants careful rating by using sliders on a single axis, thus allowing instant visualization of the ratings. Also, the use of reference and anchor is optional as well as the maximum length of the stimuli.

Though subjective methods provide valuable insights, they have limitations:

- Time-consuming: Requires human participants and controlled listening environments;
- Variability: Perception varies among listeners, requires a large test group.

3 Methods

3.1 Recording Session

The recording session took place on Tuesday, February 3, 2025, in the auditorium of the CRR 93. The recording process was supported by two conservatory students managing the control room and computer, along with two interns and one of the professors overseeing this project.

For the setup, auxiliary microphones were used for each instrument (and singer). For the mixture recording, four cardioid microphones were used to replicate the functionality of the ORTF technique. These microphones were mounted on a bar, evenly spaced and oriented as shown in Figure 1, to get a global mixture recording. A soundfield microphone (with a 90-degree angle between each microphone capsule) was also added in order to allow for decoding on different multichannel configurations. Additionally, a room microphone was installed in the fourth row of the audience to experiment with far-field hypothesis: room reflections and reverberation become prominent on the mixture. The selection and placement of the microphones were carefully chosen based on the musicians' positions to maximize source isolation. Recordings were made using Sequoia on the auditorium's control computer.

Figure 1: Front : Bar holding the 4 cardioid microphones for mixture recording / Back : Soundfield microphone

Table 1: Microphone details for instruments and mixture microphones

Instrument	Microphone Model	Directivity
Clarinet (Classical)	AT4050	Cardioid
Violin (Classical)	AT4050	Cardioid
Piano (Classical)	DPA 4007	Omnidirectional
Conga (Jazz)	DPA 4007	Omnidirectional
Bass (Jazz)	DPA 4007	Omnidirectional
Piano (Jazz)	AT4050	Cardioid
Alto Sax (Jazz)	AT4050	Bidirectional
Mixture Microphones		
4 microphones bar	2 Schoeps MK4 (center), 2 DPA 4011 (ext)	
Additionnal microphone	Soundfield	

In the configuration for Schubert music (see Figure 2), the ensemble comprised a clarinet (Yamaha 255), a violin, and a Steinway grand piano. Two takes were recorded playing all together. Then each musician recorded his part individually. These isolated recordings provide a reference dictionary of notes played in the piece that can be used to supervise the source separation algorithms.

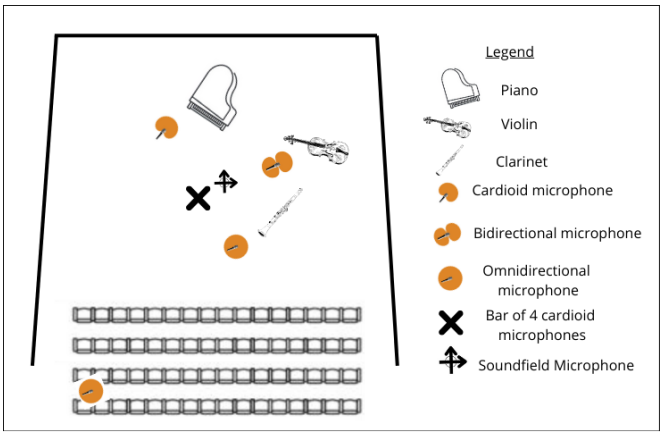

Figure 2: Configuration of the recording of Schubert trio.

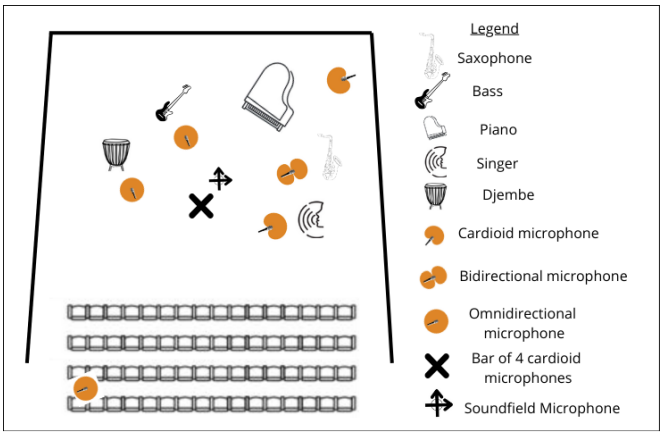

Figure 3: Configuration of the recording of jazz quartet.

For the jazz configuration (see Figure 3), the ensemble played the piece twice. The instruments used were a conga, a Jazz bass (Squier), a grand piano (Steinway), an alto saxophone, and a male baritone voice. Pictures for both configurations could be found in Appendix C.

Room impulse response For a given reverberation time of 1.8 s and a room volume of 2700 m^3, the transition to a diffuse field is estimated at approximately 52 Hz (schroeder frequency) [25], indicating that above this frequency the acoustic field can be considered diffuse. Room impulse responses were measured at two distinct locations in the auditorium using a frequency sweep from 20 Hz to 20 kHz. Once the impulse responses were converted to the frequency domain via FFT, these measurements could have been used as prior information to guide the estimation of spatial covariance matrices. For instance, the first N columns of the frequency-dependent matrices

Q_f (of size $M \times M$) could be assigned to the room impulse responses (represented as an $N \times M \times F$ tensor), while the remaining columns could be filled randomly using, for example, the last columns from the decomposition of the average mixture covariance matrix. Although this approach offers an interesting way to supervise algorithms, it was not implemented in the present work. However, the recorded room impulse responses were analyzed using the REW software, and the reverberation times across different frequencies are provided in the appendix. The measured reverberation time at 500 Hz was found to be 1.3 seconds, which is lower than the initially given value of 1.8 seconds.

Finally, all recorded tracks were organized by configuration and version for subsequent processing by the various source separation algorithms.

3.2 Separation Algorithms and Models

The following problem is extracted: the observed multichannel mixture signal in the time-frequency domain is assumed to be a linear combination of multiple source signals. Let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ denote the STFT spectrogram of the observed multichannel mixture signal, $\mathbf{X}_n = \{\mathbf{x}_{ftn}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ denote the image of source n, where M is the number of microphones, N is the number of sources. The mixture signal can be modeled as:

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{ftn} \in \mathbb{C}^{M} \tag{6}$$

Given $\mathbf{X}$ as observed data, the objective is to estimate the latent source images $\{\mathbf{X}_n\}_{n=1}^{N}$.

3.2.1 GaussMNMF

Sawada et al. [1] proposed a multichannel extension of Non-negative Matrix Factorization (NMF), GaussMNMF, which extends conventional NMF to handle complex-valued data by modeling the spatial characteristics of audio signals using a complex Gaussian distribution, which forms the theoretical foundation for the subsequent methods. In this method, each source image $\mathbf{x}_{ftn}$ is assumed to be generated by a complex Gaussian distribution:

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\Big(0,\, \lambda_{ftn}\, \mathbf{G}_{nf}\Big), \tag{7}$$

where λ_{ftn} denotes the power spectral density (PSD) of source n at frequency f and time t, and $\mathbf{G}_{nf} \in \mathbb{C}^{M\times M}$ is the spatial covariance matrix for source n at frequency f. These covariance matrices are constrained to be Hermitian and positive semidefinite, ensuring a valid statistical model of the spatial characteristics.

To capture the spectral structure of each source, GaussMNMF employs a low-rank NMF model for the PSD:

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf}\, h_{nkt}, \tag{8}$$

with $w_{nkf} \geq 0$ representing the spectral basis for source n at frequency f and $h_{nkt} \geq 0$ its temporal activation for basis k. Note that k can depends on the source n for music source separation as each instruments does not have the same number of notes for instance etc.

The quality of the approximation is measured using the multichannel Itakura–Saito divergence.

The spatial matrices are estimated by solving an algebraic Riccati equation through eigenvalue decomposition. Once the parameters $\{t_{ik}, v_{kj}, H_{ik}\}$ are estimated, source separation is performed using a multichannel Wiener filter, which minimizes the mean squared error (MSE). This filter reconstructs the separated time-domain signals after an inverse transformation (e.g., iSTFT).

This full model can be progressively simplified: first by imposing joint diagonalization to obtain FastMNMF2, and then by enforcing a rank-1 spatial constraint to derive ILRMA. This two methods are described in following sections.

3.2.2 FastMNMF2

Building upon the complete GaussMNMF model, FastMNMF2 introduces a key simplification to reduce the computational complexity of the spatial modeling. In this approach, instead of estimating each full-rank spatial covariance matrix $\mathbf{G}_{nf}$ independently as has been done in FastMNM1, it is assumed that for each frequency bin f there exists a common diagonalizer $\mathbf{Q}_f \in \mathbb{C}^{M\times M}$ such that

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\, \mathrm{Diag}(\tilde{\mathbf{g}}_n)\, \mathbf{Q}_f^{-H}, \quad \forall\, n, \tag{9}$$

where $\tilde{\mathbf{g}}_n \in \mathbb{R}_+^M$ is a frequency-invariant non-negative vector. This joint diagonalization assumption drastically reduces the number of free parameters compared to the full GaussMNMF model, while still preserving a full-rank spatial representation.

Therefore, by using "Multiplicative Update" rules, a method to iteratively update the parameters by computing the partial derivative along all parameters in the negative log-likelihood and zeroing the gradient to make sure the updated parameters remains non-negative, parameters could be updated as:

$$w_{nkf} \leftarrow w_{nkf} \sqrt{\frac{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \tag{10}$$

$$h_{nkt} \leftarrow h_{nkt} \sqrt{\frac{\sum_{f,m=1}^{F,M} \omega_{nkf}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M} \omega_{nkf}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \tag{11}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm} \sqrt{\frac{\sum_{f,t,k=1}^{F,T,K} \omega_{nkf}h_{nkt}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,t,k=1}^{F,T,K} \omega_{nkf}h_{nkt}\tilde{y}_{ftm}^{-1}}}. \tag{12}$$

where w_{nkf} and h_{nkt} have the same definition with Equation (8), which is basis matrix and activation matrix respectively. $\mathbf{Q}_f$ is given by iterative projection (IP) [3,

26]:

$$\mathbf{V}_{fm} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{ft}\tilde{y}_{ftm}^{-1} \tag{13}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{V}_{fm})^{-1}e_m \tag{14}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{H}\mathbf{V}_{fm}\mathbf{q}_{fm})^{-1/2}\mathbf{q}_{fm} \tag{15}$$

e_m is a one-hot vector whose m-th element is 1.

Compared to FastMNMF1, FastMNMF2 introduces two major improvements:

- **Frequency Invariant**: FastMNM2 shares the directional feature of each source ouver all frequency bins;
- **Computational Efficiency:** FastMNMF2 incorporates advanced optimization techniques to reduce computational complexity while maintaining high separation accuracy.

And these improvements make it particularly suitable for complex scenarios such as:

- Separating overlapping speech signals in reverberant environments;
- Extracting individual instruments from multi-track music recordings;
- Enhancing audio quality in multi-microphone setups.

3.2.3 Independent Low-Rank Matrix Analysis

Independent Low-Rank Matrix Analysis (ILRMA) is essentially a variant of the FastMNMF2 method with stricter spatial modeling assumptions. While FastMNMF2 employs a full-rank spatial model, using jointly diagonalizable covariance matrices with frequency-invariant eigenvalues to capture complex acoustic effects, ILRMA enforces a rank-1 constraint on each source's spatial covariance matrix. This is reducing the number of parameters and computational cost.

As in FastMNMF2, the observed multichannel mixture is represented in the time-frequency domain by the STFT spectrogram $\mathbf{x}_{ft}$ and is modeled as a sum of source images. In ILRMA, each source image $\mathbf{x}_{ftn}$ is assumed to follow a complex Gaussian distribution with a rank-1 covariance structure:

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\Big(0,\, \lambda_{ftn}\,\mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}}\Big), \tag{16}$$

where λ_{ftn} denotes the power spectral density (PSD) of source n at frequency f and time t, and $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the mixing vector associated with that source.

To model the temporal and spectral evolution of each source, ILRMA adopts a low-rank NMF framework to factorize the PSD as shown in Eq. (8).

Parameter estimation proceeds iteratively in two main steps. First, similar to FastMNMF2, a set of demixing matrices $\mathbf{W}_f$ is updated to extract the separated signals from the mixture. Second, multiplicative update rules, typically derived via Itakura–Saito divergence minimization, are employed to optimize the NMF parameters w_{nkf} and h_{nkt}. The estimation scheme closely follows that of FastMNMF2, with the key distinction being the enforced rank-1 constraint on the spatial model.

3.2.4 Prior Information Trick

To improve the performance of FastMNMF2 and ILRMA, a method that leverages prior knowledge by fixing the basis matrix $\mathbf{W}$ and only updating the activation matrix $\mathbf{H}$ was proposed. This approach reduces the number of optimization variables and enhances the stability and efficiency of the model.

To obtain a known $\mathbf{W}$, it should calculate the Power Spectral Density (PSD) of audio recordings of musical notes (e.g., C1 to C5) played by the same instrument as the target audio. These PSDs are then mapped to $\mathbf{W}$ to form a dictionary.

Suppose each instrument plays $\{k_n\}_{n=1}^{N}$ notes individually. The dictionary $\mathbf{W}$ could be constructed as

$$w_{nfk_n} = |S_{nfk_n}|^2 \tag{17}$$

where $S_{nfk} \in \mathbb{C}$ is the STFT spectrogram of $n-$th source instrument individually played audio at f frequency bin for k_n-th note.

3.2.5 Hybrid Transformer Demucs

Hybrid Transformer Demucs (HT Demucs) [18] is an advanced deep learning model for music source separation, combining the strengths of convolutional neural networks (CNNs) and transformer architectures. It is designed to separate mixed audio signals into individual sources, such as vocals, drums, bass, and other instruments, with high precision.

HT Demucs integrates a hybrid architecture, as Figure 13 shows in Appendix A, that operates in both the time and frequency domains:

- **Encoder:** A series of convolutional layers processes the raw waveform to extract hierarchical features.
- **Transformer Blocks:** Transformer layers are used to model long-range dependencies in the frequency domain, enhancing the model's ability to separate overlapping sources.
- **Decoder:** Transposed convolutional layers reconstruct the separated sources from the encoded features.

HT Demucs has been trained on a substantial dataset, MUSDB18 [27], comprising both mixed and isolated audio tracks, with a total duration of approximately 10 hours of music.

3.2.6 Spleeter

Developed by Deezer, Spleeter [5] is a popular deep learning-based tool for music source separation. It is designed to separate mixed audio tracks into 2, 4 or 5 stems. In this project, 5 stems were chosen for Spleeter configuration.

Spleeter utilizes a 12-layer U-Net architecture, with an 6-layer encoder and 6-layer decoder of CNN units. The models were trained on Bean datasets [28] with 79 hours music of Pop and Rock.

3.3 Evaluation

3.3.1 Subjective Evaluation

The MUSHRA webpage, as well as a screen capture used for subjective evaluation is shown in the Figure 10 in Appendix B. All tracks shown in the MUSHRA evaluation were obtained using the mixture that was recorded with the bar of 4 cardioid microphones as the algorithms input. The results from GaussMNMF weren't included in the subjective test as the algorithm produced erratic and incoherent outputs, even after very long computation times (ranging from 3 to 13 hours on a standard computer). The audios were unpleasant to listen to, and it was expected that participants would rate them very poorly, therefore, they were excluded from the subjective evaluation.

3.3.2 Objective Evaluation

Blind Source Separation Evaluation (BSS Eval) Metrics [19] is used for objective evaluation, including Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and Sources-to-Artifacts Ratio (SAR). These three metrics are based on the decomposition of the estimated signal $\hat{s}_i(t)$ into four separate components:

$$\hat{s}_i(t) = s_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t) \quad (18)$$

where:

- $s_{\text{target}}(t)$ is an acceptable deformation of the true source $s_i(t)$;
- $e_{\text{interf}}(t)$ represents interference from other undesired sources;
- $e_{\text{noise}}(t)$ corresponds to perturbation noise not originating from the sources;
- $e_{\text{artif}}(t)$ accounts for artifacts introduced by the separation algorithm, such as distortions or artificial sounds.

From this decomposition, the following metrics are computed as

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2} \quad (19)$$

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \quad (20)$$

$$\text{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2} \quad (21)$$

4 Results and Analysis

4.1 Choice of parameters

In experiments, several key parameters were evaluated to assess their impact on source separation performance. For the MNMF-based methods (GaussMNMF, FastMNMF2, and ILRMA), the number of bases was fixed to $K = 30$. Three different types of input configurations were under evaluation: a stack of all auxiliary microphones, a stack of the four cardioid microphones on the bar (mixture recording), and recordings obtained with the soundfield microphone.

In addition, a comparison was made between two approaches: supervised and unsupervised. In the supervised setting, a dictionary of spectral bases, derived from individual note recordings, was used to initialize the model, thereby setting the number of bases according to the MNMF factorization. This was used to fix the $\mathbf{W}$ matrix (Section 3.2.4). For the deep learning models (Demucs and Spleeter), no fine-tuning was performed as the models were used with their default configurations.

It should be mentioned that classical recordings were not processed with GaussMNMF or Spleeter, which is the reason why these algorithms do not appear in the classical categories of the evaluations later in this report. For Spleeter, the model was not trained with the appropriate number of stems for the Schubert trio, nor was it trained on the right type of instruments for this track. For GaussMNMF, most trials were conducted on jazz recordings. Once it became evident that the outputs were of poor quality and the focus shifted to testing new algorithms, supervised trials for classical music were not pursued.

4.2 Subjective Evaluation

Over the course of 3 days, 53 participants completed the 10-minute MUSHRA subjective evaluation evaluation online. Figure 4 and 5 show the distribution of participants in gender and age, respectively.

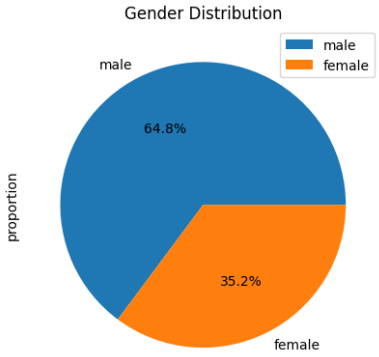

Figure 4: Gender distribution

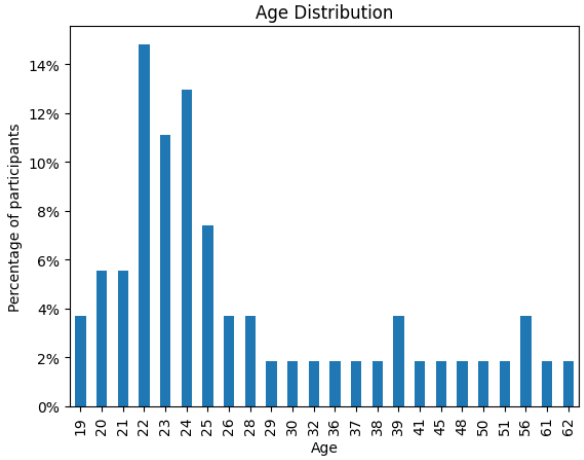

Figure 5: Age distribution

The participant distribution was approximately 2/3 male and 1/3 female, with a bell curve centered around the age of 22. This shows a diversity of participant profiles, leading to a variety of results across the two musical pieces.

Table 2: Mean MUSHRA scores, normalized from 0 to 100, higher is better

Config/Trial ID	Reference	FastMNMF2	ILRMA	Spleeter	Demucs
clas. reconstruction	88.6	66.6	42.8	N/A	77.8
clas. clarinette	84.3	37.9	25.5	N/A	43.8
clas. piano	82.0	20.0	48.2	N/A	81.5
clas. violin	77.2	21.4	27.9	N/A	35.0
jazz reconstruction	90.0	71.3	60.5	84.6	36.1
jazz bass	49.7	22.4	19.2	85.7	85.7
jazz drum	55.0	47.5	49.8	43.1	87.7
jazz piano	74.0	53.1	25.9	86.6	N/A
jazz saxophone	85.9	27.0	37.5	44.0	N/A
jazz voice	68.8	53.9	51.0	89.8	78.8

Figure 6, 7 and Table 2 present the mean MUSHRA scores for different source separation algorithms across various configurations and trials. The scores are normalized to a scale of 0 to 100, where higher scores indicate better separation quality. The mixture refers to the audio segment captured from main 4 microphones bar. For the "reconstruction" trials, the reference is the audio segment captured from the main 4 microphones bar. For other trials, the reference is the audio segment captured from a auxiliary-mic configuration specific to the target instrument.

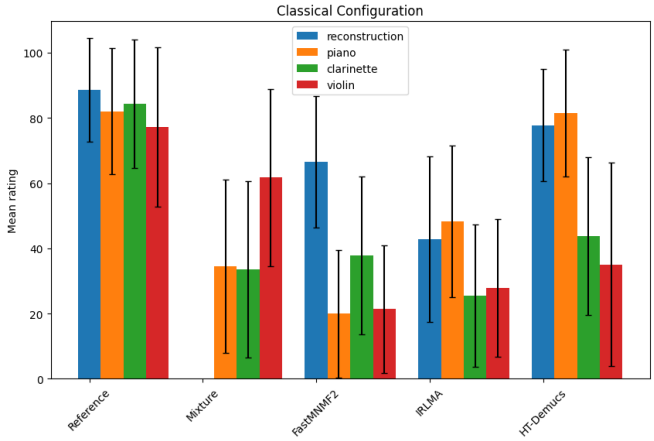

Figure 6: Statistics for Classical Configuration in MUSHRA evaluation, normalized from 0 to 100, higher is better

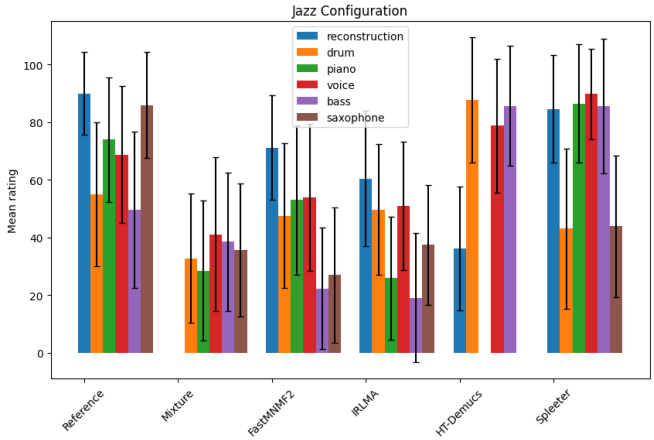

Figure 7: Statistics for Jazz Configuration in MUSHRA evaluation, normalized from 0 to 100, higher is better

4.3 Objective Evaluation

Objective evaluation was conducted on the reconstructed signals produced by the different algorithms. The ground truth used for comparison was the mixture signal recorded with the 4-cardioid bar. The results are presented in Tables 3 and 4. Higher results mean better separation according to the specified metrics.

Table 3: Objective metric results for the jazz reconstruction

Algorithm/Metric	SDR (dB)	SIR (dB)	SAR (dB)
GaussMNMF	-9.6	22.1	-9.5
FastMNMF2	-4.6	26.0	-4.6
ILRMA	-15.0	16.3	-14.9
Demucs	0.4	30.8	0.4
Spleeter	11.9	41.3	11.9

Table 4: Objective metric results for the classical reconstruction, higher is better

Algorithm/Metric	SDR (dB)	SIR (dB)	SAR (dB)
FastMNMF2	-7.3	23.0	-7.3
ILRMA	-8.2	20.6	-8.1
Demucs	13.3	43.7	13.3

This objective evaluation was also conducted on the separated instrument tracks. The objective metrics were computed for each source extracted by each algorithm, which allows for analysis of the dependency of separation performance on both the source type and recording configuration. In particular, the results for the piano in the jazz and classical configurations are shown in Figure 8.

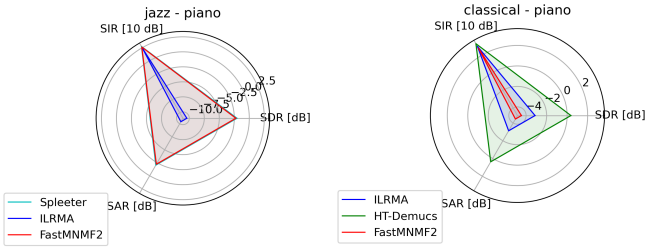

Figure 8: Objective evaluation metrics for the piano tracks in both jazz (left) and classical (right) configurations across different algorithms.

5 Discussion

5.1 Relation between Subjective and Objective Evaluation

Subjective evaluations of audio source separation are based on perceptual factors such as timbral fidelity, spatial attributes, and the overall listening experience. These aspects are not always captured by objective metrics like SDR, SIR, and SAR: those focus on quantifiable elements such as signal distortion, interference, and artifacts. Studies by Rumbold et al. (2024) [29] and Emiya et al. (2011) [30] have demonstrated a weak correlation between these objective metrics and human perception. The best mean correlation coefficient across different tasks and evaluation setups was only 0.246 in the 2024 study (for SDR and other metrics unspecified here) and 0.5 in the 2011 study (for SDR, SIR and SAR and other metrics unspecified here). This happens because objective metrics don't fully capture the way people actually hear and judge sound. Some algorithms can get high scores by reducing measurable distortions but still produce audio that sounds unnatural to the human ear. On the other hand, some algorithms might not score well objectively but still sound more natural to listeners. This shows why it's important to use both objective and subjective tests to properly evaluate sound separation quality. To find the most reliable metric in this study, we refer to the 2011 study [30], which compares SDR, SIR, and SAR. SIR has the highest accuracy (0.72) for individual subjective scores, while SDR and SAR are much lower (0.37 and 0.31) in the same case. SIR is the best match with human perception because it measures how well interference is removed, which is the main factor in perceived separation quality. On the other hand, distortions and artifacts do not always negatively impact perceived quality for listeners as long as the source remains intelligible. While SIR is the most reliable predictor, its correlation with human evaluation is still only moderate. The limitations of objective metrics can be seen here: while GaussMNMF shows better objective results than ILRMA, its perceptual quality is actually worse when listening to the separated tracks. However, the metrics were more consistent in distinguishing between ILRMA and FastMNMF2, where the differences aligned better with subjective impressions.

5.2 Comparison for Signal Processing Algorithms

In this section, the three main signal processing algorithms (GaussMNMF, FastMNMF2, and ILRMA) will be compared. Based on objective evaluation, FastMNMF2 achieves the highest performance levels in dB, consistently outperforming the other algorithms across all metrics. GaussMNMF follows, with ILRMA ranking last. This aligns with the different assumptions underlying each algorithm, with ILRMA relying on a rank-1 covariance matrix. This constraint limits its ability to accurately separate sources in reverberant environments, where more complex spatial modeling is required. As of subjective evaluation, only FastMNMF2 and ILRMA were ranked (cf. Section 4.1). FastMNMF2 scored better for the reconstruction for both classical and jazz tracks : this is also coherent with the rank of the matrices induced by each algorithm. ILRMA is also mostly evaluated less effective by the listerners in the jazz configuration because its performance was limited by this property, and because there were more sources. In the jazz configuration, the saxophone track processed with ILRMA was the only one rated by listeners as better separated than the tracks obtained with FastMNMF2. This better performance can be due to the source's strong directionality, enhanced by the placement of the saxophone, close to the main 4 microphones bar. On the other hand, for the classical trio, ILRMA is ranked better than FastMNMF2 for the piano separation. This can be due to the fact that there are only three sources, which means less overlapping sound energy and fewer room reflections. Finally, the violin in the classical piece and the bass in the jazz piece received the lowest separation ratings. The violin was difficult to isolate due to its homorhythmic interplay with the clarinet, while the bass blended with the drums and, being a non-acoustic instrument, was more challenging to control in the separation process.

5.3 Comparaison between Signal Processing Algorithms and Deep Learning Models

For both configurations, the quality of the separation was consistently better with modern deep learning models compared to FastMNMF2, ILRMA, and GaussMNMF. It is the case for objective and subjective results. This can be attributed to the training of deep learning models on large and diverse musical datasets, allowing them to better capture harmonic structures and the temporal characteristics of both instruments and vocals. In particular, Demucs, which operates in the time domain, may leverage a broader set of features, leading to a more natural reconstruction of the separated sources. Additionally, deep learning models inherently learn to handle reverberation as it is present in their training data, whereas traditional signal processing methods struggle more with reverberant environments, making source separation more challenging.

5.4 Influence for Different Inputs and Configurations

For each algorithm and configuration, three different input setups were tested. The first experiment used auxiliary microphones, where the input signals were already partially separated. This was expected to be the easiest case, and indeed, it yielded the best results across all algorithms, overall on ILRMA which quality was really better than with the two other inputs. The second experiment used the mixture signals recorded with the 4-cardioid microphone bar. To ensure a determined case for the jazz configuration (five sources), we also incorporated the room microphone. The performance of the signal processing algorithms remained relatively similar to the first experiment, except for the vocal track, which was noticeably degraded.

The third experiment used soundfield recordings as input. While ILRMA produced similar results to the previous configurations, FastMNMF2 and Demucs performed worse. This could be due to the lower spatial precision provided by the soundfield recordings compared to the 4-cardioid bar, which provides more distinct directivity information. Some of these observations are not reflected in the histograms, which focus on the second input experiment, but all corresponding audio files are available on the project webpage.

The separation quality is generally better in the classical configuration than in the jazz configuration for signal processing algorithms. This is mainly because jazz has more sources to separate, making the task harder. However, deep learning models perform less on some instrument sets, like clarinet, piano, and violin. This is because they were not trained on similar data, showing their limits when faced with unfamiliar instrument combinations. This highlights the need for a dataset that matches the separation task.

Therefore, quality of the recording setup strongly influences the results: preliminary separation using auxiliary microphones facilitates the algorithms' task, whereas diffuse-field recordings present additional challenges.

6 Conclusion

Understanding the factors that influence source separation difficulty is essential for improving the effectiveness of algorithms. The way instruments are played significantly impacts separation difficulty. Avoiding homorhythmic passages can help, as overlapping note onsets make it harder to distinguish individual sources. The disposition of instruments during recording is also crucial: placing them too close together increases leakage, complicating separation. For deep learning methods, using instruments that match standard stem categories (vocals, bass, piano, percussion) improves performance, as these models are typically trained on such configurations. The use of auxiliary microphones can enhance separation, and their directivity plays a key role in pre-isolating sources and should be carefully chosen. Finally, room reverberation is also an important factor: highly reverberant spaces introduce longer decay times, making separation more challenging for signal processing algorithms.

Combining optimized recording techniques with separation algorithms can contribute to the development of more robust models. A structured approach that minimizes interference between instruments can provide cleaner training data, making deep learning models more adaptable to real-world scenarios. Additionally, studying how reverberation and directivity affect separation can help refine algorithmic design for better performance in complex acoustic environments.

Future improvements could include integrating additional prior knowledge into the separation process. For instance, reverberation models based on the room's impulse response could help algorithms better handle realistic acoustic conditions. Similarly, incorporating spectral-temporal information, such as a note dictionary for each instrument, has already proven beneficial in improving FastMNMF2 results. Finally, training deep learning models on specific music styles and recording conditions would make them more reliable in these contexts.

Several aspects of the project were challenging. The choice of musical genres was an important factor. Jazz and classical music were selected because they have distinct characteristics. Choosing the right algorithm for each recording condition was also difficult. The assumptions behind each method must be well understood, as their effectiveness depends on how well these assumptions match the recording conditions. The way the recording is done, like microphone placement, room acoustics, and instrument positioning, directly impacts separation performance. Some algorithms, like GaussMNMF, did not perform well in this study, highlighting how hard it is to predict results in real-world conditions. Evaluating the algorithms added another layer of complexity. Selecting the right objective metrics required defining what makes a "good" separation, which depends on the context. The

challenge was to balance objective measures with perceptual evaluations while considering the impact of different recording conditions.

Acknowledgement

We would like to express our gratitude to Mr. Benoît Fabre and Mr. Mathieu Fontaine for their guidance throughout this project.

We also extend our thanks to the Aubervilliers Conservatory for providing us with the recording space.

We are also grateful to the conservatory students for their assistance during the recording session.

Finally, we sincerely appreciate each of the musicians who contributed their time and talent to the recordings.

Appendix A Architecture of HT Demucs

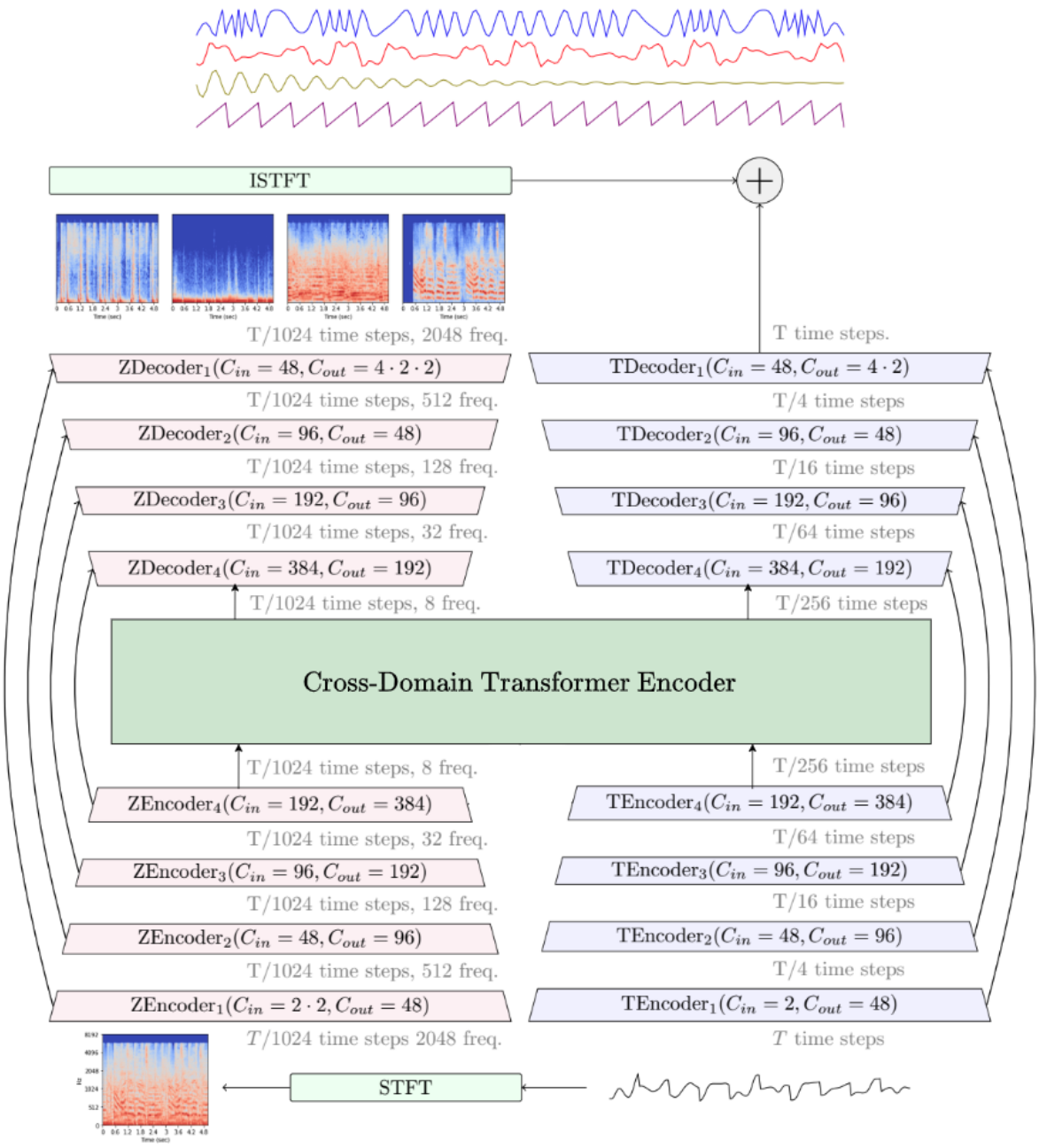

Figure 9: Architecture of Hybrid Transformer Demucs

Appendix B MUSHRA Subjective Evaluation

The MUSHRA Subjective Evaluation test is available on https://perso.telecom-paristech.fr/jkang-23/Evaluation/.

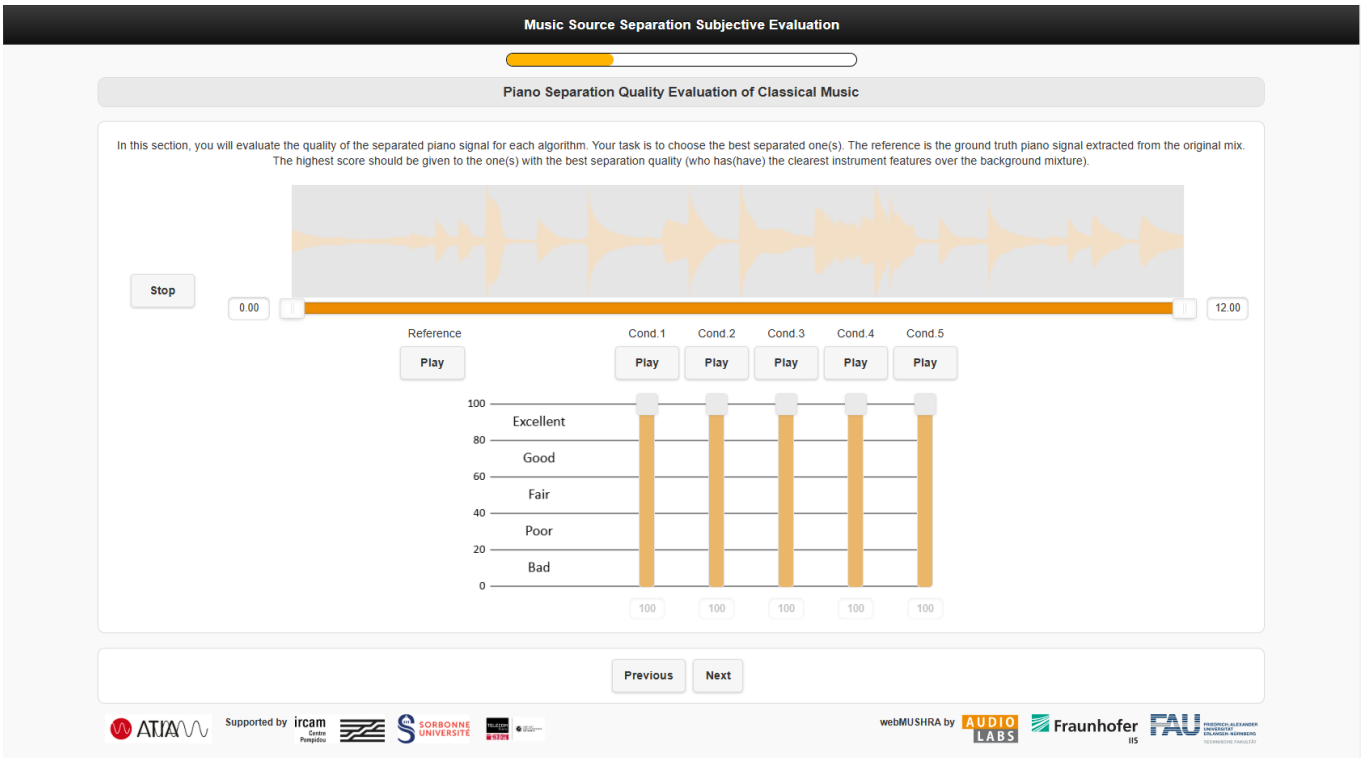

Figure 10: MUSHRA Evaluation Interface

Appendix C Pictures

Pictures from the recording day (04/02/2025, at the CRR 93 - Aubervilliers):

Front picture

Above picture

Figure 11: **Classical configuration**
Piano: Aurélien J / Violin: Simon J / Clarinet: Damien F

Front picture

Above picture

Figure 12: **Jazz configuration**
Conga: Mathys D / Bass: Marco M / Piano: Alice P / Sax: Adrien ? / Voice: Simon J

Appendix D Reverbertion time

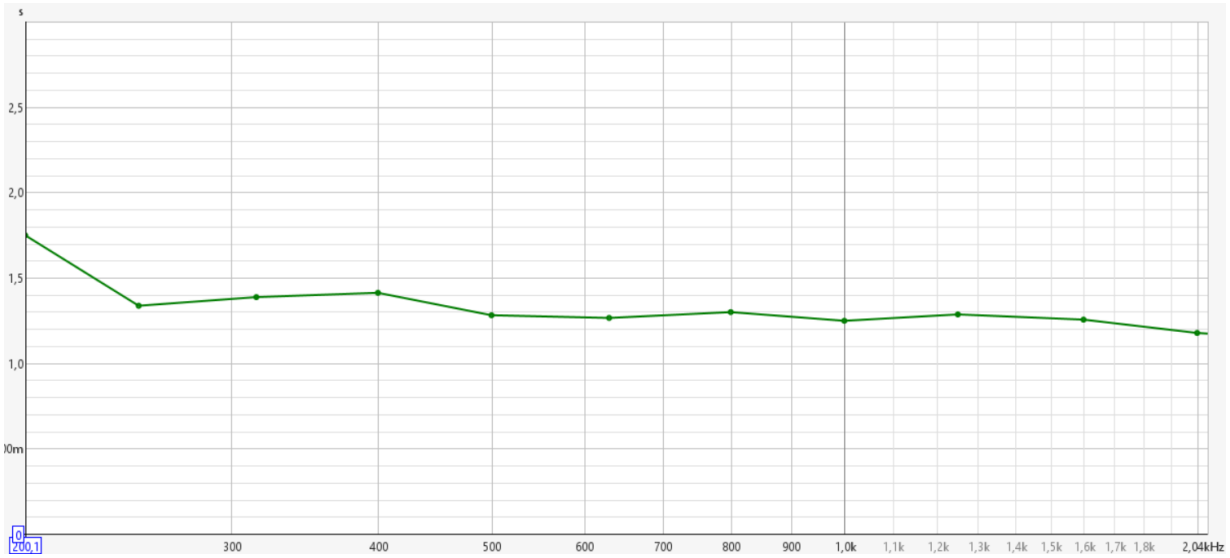

Figure 13: Reverberation times across frequency measured in CRR93 Auditorium, Aubervilliers