PAM - Projets et Applications Musicales
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 17, 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 knowledge, 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.
+ ADD END OF ABSTRACT ON RESULTS AND DISCUSSION

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 knowledge* 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.

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). For each setup, the musicians were arranged on stage to be as far apart as possible while keeping at hearing distance. The choice of microphones was guided by the musicians' placement and the microphones' directivities. Further explanation is given in Section 3.1.

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 was Gaussian MNMF (Multichannel Nonnegative Matrix Factorization using Gaussian distributions), an extension of the classic NMF algorithms [1]. This method was the first MNMF method to appear chronogically, which is why it's tested first in this project. With this algorithm, the sources are modeled with nonnegative 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. The results obtained

using Gaussian MNMF will not be exploited in this paper due to the generation of erratic and incoherent outputs after an excessively long computation time (ranging from 3 to 13 hours). However, all results are available on the project webpage: **ADD WEBPAGE LINK**. Then, FastMNMF2 was implemented in this study [2]. 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 (Independent Low-Rank Matrix Analysis) was implemented [3]: 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 CNN 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 with the signal processing algorithms we implemented, though their performance may be less high when the test data differs from the training domain. More information about algorithms is given in Section 3.2.

Method evaluation The performance evaluation was assessed with subjective methods and objective methods. The performance of the source separation algorithms was evaluated using both subjective and objective methods. For the subjective evaluation, listening tests based on the MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) methodology 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, including SDR (Signal-to-Distortion Ratio), SIR (Signal-to-Interference Ratio), and SAR (Signal-to-Artifacts Ratio), were used to quantify the quality of separation based on how clear the sound is, how well unwanted noise is removed... This two-way approach gives a complete evaluation, combining human listening opinions with measurable data.

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]. Unlike many think, sound recording is not the first step to achieve great separation but rather one of the last, as we need to know what kind of sounds we want to use to properly suit our 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, the choice of recording method depends on the specific requirements of the separation task. Here, the techniques were chosen based on the available equipment, the instruments, and their positioning in the room.

2.2 Algorithms

Different algorithmic approaches exist for implementing audio source separation. These approaches vary accord-

ing 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, we define:

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

Instantaneous linear mixtures:

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

Convolutive mixtures:

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

Independent Component Analysis (ICA). 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.

Expectation-Maximization. To estimate these parameters, two common methods are Expectation-Maximization (EM) and Multiplicative Updates (MU). The EM algorithm [12] 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. Alternative methods include Majorization-Minimization [13] and auxiliary function approaches [14].

NMF and MNMF. Another approach is **Non-negative Matrix Factorization (NMF)** [15], 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 divergence measure (Euclidean, Kullback-Leibler, Itakura-Saito, etc.).

For multichannel signals, NMF is extended to **Multichannel NMF (MNMF)** [1]. The spectrogram becomes a tensor, modeling both spectral content and spatial structure. There are two main approaches: one based on convolutive modeling with EM, which is more precise but computationally expensive, and another minimizing divergences on power spectrograms using multiplicative updates (MU), which is simpler but relies on stronger assumptions.

When microphone configurations and room conditions are documented, these data can be incorporated into MNMF to constrain the spatial covariance matrix and accelerate convergence [16]. Due to the high computational cost of MNMF, Sekiguchi et al. [16] introduced **FastMNMF**, 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, improving efficiency and separation quality. ILRMA (Independent Low-Rank Matrix Analysis) introduced by Ono [3] is a simplified variant of FastMNMF that imposes a rank-1 constraint on each source's spatial covariance, meaning that each source is modeled as the outer product of a single mixing vector. 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.

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

Deep Learning Methods: Spleeter and Demucs. Recent advances in deep learning have led to the development of highly effective music source separation models. In our study, we focused on two state-of-the-art approaches: **Spleeter** and **Demucs**.

Spleeter 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, 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[18], 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)[19].

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[20] 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)[21], 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)[22] 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[23]. 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 and results

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 was also added for experimentation with spatialization. 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)	AT4040	Cardioid
Violin (Classical)	AT4040	Cardioid
Piano (Classical)	DPA 4007	Omnidirectional
Conga (Jazz)	DPA 4007	Omnidirectional
Bass (Jazz)	DPA 4007	Omnidirectional
Piano (Jazz)	AT4040	Cardioid
Alto Sax (Jazz)	AT4040	Bidirectional
Mixture Microphones		
4 microphones bar	2 Schoeps MK4 (center), 2 DPA 4011 (ext)	
Additionnal microphone	Soundfield	

In the classical configuration (see Figure 2), the ensemble comprised a clarinet (Yamaha 255), a violin, and a Steinway grand piano. They played the piece twice as an ensemble. Then each musician recorded their 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.

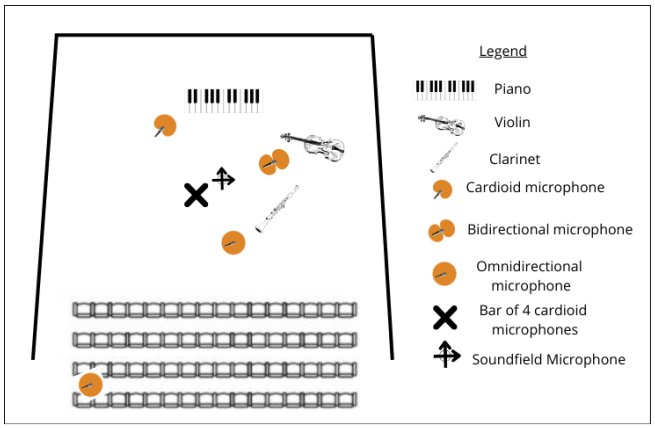

Figure 2: Classical configuration of the recording session.

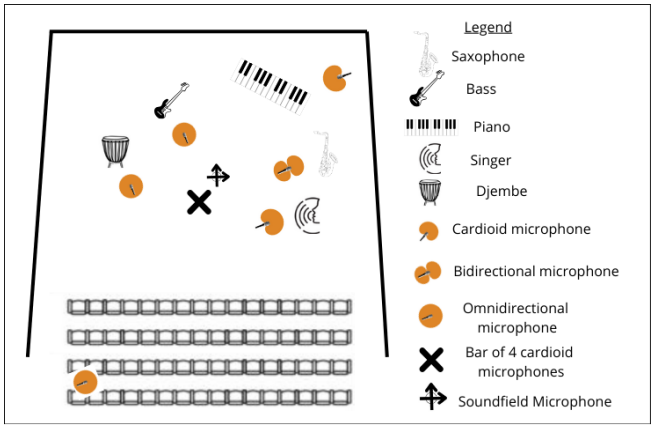

Figure 3: Jazz configuration of the recording session.

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.

Room impulse response For a 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, 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.

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

3.2 Algorithms

The following problem is considered: 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.

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 Rules", parameters could be updated as:

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

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

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

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{V}_{fm})^{-1}e_m. \tag{13}$$

where e_m is a one-hot vector whose m-th element is 1, $\mathbf{V}_{fm} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{ft}\tilde{y}_{ftm}^{-1}$.

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 ILRMA

ILRMA (Independent Low-Rank Matrix Analysis) 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 means that each source is modeled by a single mixing vector, reducing the number of parameters and computational cost while often enhancing robustness in both determined and overdetermined scenarios.

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:

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\Big(0,\, \lambda_{ftn}\, \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}}\Big), \tag{14}$$

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, we propose a method that leverages prior knowledge by fixing the basis matrix $\mathbf{W}$ and only updating the activation matrix $\mathbf{H}$. This approach reduces the number of optimization variables and enhances the stability and efficiency of the model.

To obtain a known $\mathbf{W}$, we 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 construct as

$$w_{nfk_n} = |S_{nfk_n}|^2 \tag{15}$$

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) [24] 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 8 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 [25], 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 our project, we use 5 stems Spleeter.

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 [26] with 79 hours music of Pop and Rock.

3.2.7 Experimental parameters

In our experiments, several key parameters were evaluated to assess their impact on source separation performance. For the MNMF-based methods (Gaussian MNMF, FastMNMF2, and ILRMA), the number of bases was fixed to $K = 30$. We tested three types of input configurations: 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, we compared supervised and unsupervised approaches. 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 **W** matrix (Section 3.2.4). For the deep learning models (Demucs and Spleeter), no explicit parameter tuning was performed as the models were used with their default configurations.

3.3 Subjective evaluation

Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)[22] 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. The interface used for MUSHRA testing is shown in the Figure 9 in Appendix B. All tracks shown in the MUSHRA test were obtained using the main mixture that was recorded with the bar of 4 cardioid microphones as the algorithms input.

3.4 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 [18], 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. In this project, we only consider SDR, SIR and SAR.

4 Results

4.1 Subjective evaluation

Over the course of 3 days, 53 participants completed the 10-minute MUSHRA subjective evaluation test available online at https://perso.telecom-paristech.fr/jkang-23/Evaluation/. The participants, both male and female, ranged in age from 19 to 62 years. Their responses were automatically recorded in a .csv file after each trial, which was then processed using the Python library Pandas to extract participant information:

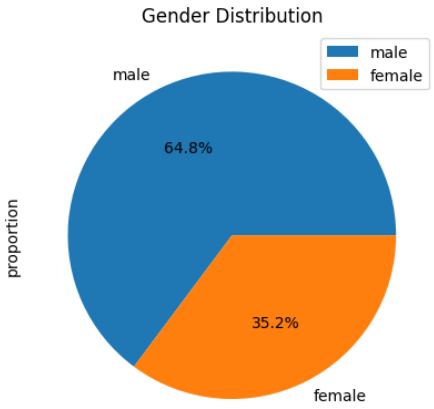

Figure 4: Gender distribution

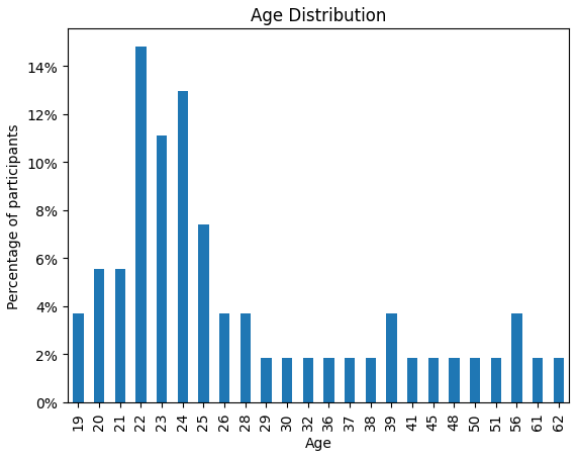

Figure 5: Age distribution

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

Table 2: Mean MUSHRA scores /100

Config/Trial ID	Mixture	Reference	FastMNMF2	ILRMA	Spleeter	Demucs
classical reconstruction	N/A	88.57	66.57	42.76	N/A	77.80
classical clarinette	33.67	84.30	37.87	25.50	N/A	43.81
classical piano	34.48	81.98	20.02	48.19	N/A	81.52
classical violin	61.70	77.22	21.37	27.87	N/A	35.04
jazz reconstruction	N/A	90.00	71.26	60.48	84.63	36.13
jazz bass	38.57	49.65	22.39	19.19	85.65	85.67
jazz drum	32.87	55.04	47.54	49.76	43.11	87.70
jazz piano	28.57	73.96	53.07	25.93	86.56	N/A
jazz saxophone	35.70	85.94	27.00	37.46	44.00	N/A
jazz voice	41.13	68.83	53.87	51.00	89.83	78.83

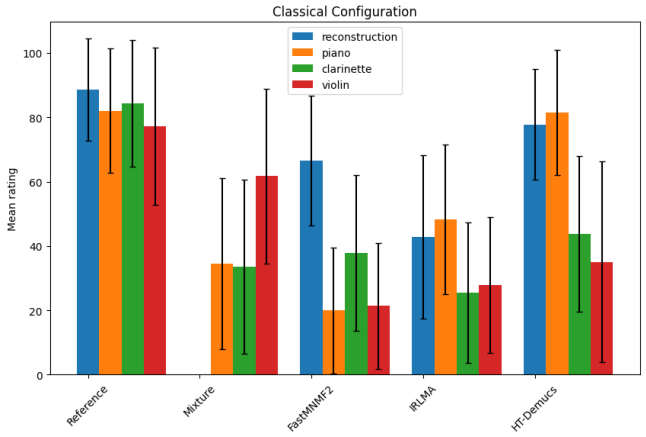

Figure 6: Histo 1

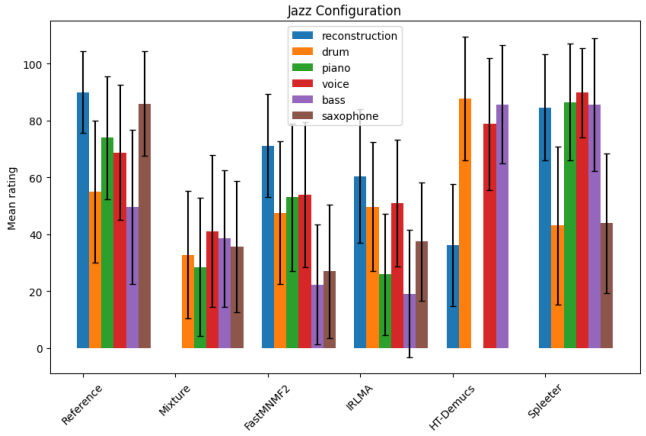

Figure 7: Histo 2

COMPARAISON (genre, algos)
Additional info on gender and age ? only if we have time

4.2 Objective evaluation

PETITE PHRASE DINTRO

Table 3: Objective metric results for the jazz dataset

Algorithm	SDR (dB)	SIR (dB)	SAR (dB)
Demucs	0.4	30.8	0.4
FastMNMF2	-4.6	26.0	-4.6
ILRMA	-15.0	16.3	-14.9
Spleeter	11.9	41.3	11.9

Table 4: Objective metric results for the classical dataset

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

5 Discussion

5.1 Signal processing algorithms

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) and Emiya et al. (2011) 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 discrepancy arises because objective metrics may not account for perceptual nuances that significantly impact listener experience. Consequently, some algorithms may achieve favorable objective scores by minimizing measurable distortions but still produce audio that listeners perceive as unnatural or less pleasing. Conversely, algorithms that introduce subtle distortions might receive lower objective scores yet preserve perceptual integrity, making them subjectively preferred. This highlights the importance of incorporating both objective and subjective evaluations to comprehensively assess the performance of sound source separation algorithms.

-> comparaison entre les 2-3 algos ts • explication probable : hypothèses du modèle (rang de la matrice), salle

For the signal processing algorithms, only FastMNMF2 and ILRMA will be compared as GaussianMNMF was highly unsuccessful in separating the sources.
Objectively, FastMNMF2 has better results than ILRMA

5.2 Comparaison with Demucs/Spleeter

For both configurations, the quality of the separation was consistently better with modern deep learn-

ing algorithms 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.

5.3 Influence of the input

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. 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 our website //////.

5.4 Influence of the supervision

Prior information was introduced in the process for the classical configuration with FastMNMF2 and ILRMA. This prior information was the notes dictionary, obtained with NMF on solo instrument tracks. It improves FastMNMF2 results, but not that much ILRMA results. This could be because FastMNMF2 jointly models spectro-temporal patterns and spatial properties, allowing it to better integrate the dictionary constraints while still adapting the separation spatially. ILRMA, on the other hand, enforces a strict rank-1 spatial model, which may limit its ability to fully exploit additional spectral constraints.

-> comparer les résultats objectifs et subjectifs en apportant des éléments d'explication • demucs au top • spleeter mieux pour objectif que pour objective • SIR SDR SAR ? à quoi cela correspond ? -> âge/genre /////

expliquer pq spleeter pas sur classique

Remerciements

References

[1] Hiroshi Sawada et al. "Multichannel Extensions of Nonnegative Matrix Factorization". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 21.4 (2013), pp. 750–762.

[2] Kouhei Sekiguchi et al. "Fast Multichannel Nonnegative Matrix Factorization With Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2610–2625. ISSN: 2329-9304. DOI: 10.1109/TASLP.2020.3019181. (Visited on 02/08/2025).

[3] Naonori Ono. "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique". In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2011, pp. 189–192.

[4] A. Défossez et al. *Demucs: Deep Extractor for Music Sources*. 2019. arXiv: 1911.13254 [cs.SD]. URL: https://arxiv.org/abs/1911.13254.

[5] Romain Hennequin et al. "Spleeter: a fast and efficient music source separation tool with pre-trained models". In: *Journal of Open Source Software* 5.50 (2020), p. 2154. DOI: 10.21105/joss.02154. URL: https://doi.org/10.21105/joss.02154.

[6] Streicher Ron and Dooley Wes. "basic stereo microphone perspectives-a review". In: *journal of the audio engineering society* 33 (7/8 Aug. 1985), pp. 548–556.

[7] François Salmon et al. "A Comparative Study of Multichannel Microphone Arrays Used in Classical Music Recording". In: *Journal of the Audio Engineering Society* (2023). URL: https://api.semanticscholar.org/CorpusID:259868549.

[8] Günther Theile. "Multichannel Natural Music Recording Based on Psychoacoustic Principles / 1". In: 2001. URL: https://api.semanticscholar.org/CorpusID:16585354.

[9] P. Comon. "Independent Component Analysis". In: *Signal Processing* 36.3 (1994), pp. 287–314.

[10] L. Benaroya, N. McDonald, and N. Dehak. "Single channel separation of voice and music using spectral modeling". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2006), pp. 1382–1395.

[11] A. Ozerov and C. Févotte. "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), pp. 550–563.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B* 39.1 (1977), pp. 1–38.

[13] K. Lange. *MM Optimization Algorithms*. SIAM, 2016. ISBN: 978-1-611974-39-3.

[14] D. Lee and H. Seung. "Algorithms for non-negative matrix factorization". In: *Advances in Neural Information Processing Systems (NIPS)*. 2000.

[15] D. Lee and H. Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[16] T. Sekiguchi, H. Sawada, and S. Araki. "Fast multichannel NMF for high-dimensional source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2618–2633.

[17] A. Ozerov, E. Vincent, and F. Bimbot. "A general flexible framework for the handling of prior information in audio source separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2012), pp. 1118–1133.

[18] E. Vincent, R. Gribonval, and C. Fevotte. "Performance measurement in blind audio source separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469. DOI: 10.1109/TSA.2005.858005.

[19] Jonathan Le Roux et al. *SDR - half-baked or well done?* 2018. arXiv: 1811.02508 [cs.SD]. URL: https://arxiv.org/abs/1811.02508.

[20] Valentin Emiya et al. "Subjective and Objective Quality Assessment of Audio Source Separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2046–2057. DOI: 10.1109/TASL.2011.2109381.

[21] Kevin Kilgour et al. *Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms*. 2019. arXiv: 1812.08466 [eess.AS]. URL: https://arxiv.org/abs/1812.08466.

[22] Michael Schoeffler et al. "webMUSHRA — A Comprehensive Framework for Web-based Listening Tests". In: *Journal of Open Research Software* (Feb. 2018). DOI: 10.5334/jors.187.

[23] Brecht De Man and Joshua Reiss. "APE: Audio Perceptual Evaluation toolbox for MATLAB". In: Apr. 2014.

[24] Simon Rouard, Francisco Massa, and Alexandre Défossez. *Hybrid Transformers for Music Source Separation*. 2022. arXiv: 2211.08553 [eess.AS]. URL: https://arxiv.org/abs/2211.08553.

[25] Zafar Rafii et al. *MUSDB18-HQ - an uncompressed version of MUSDB18*. Aug. 2019. DOI: 10.5281/zenodo.3338373. URL: https://doi.org/10.5281/zenodo.3338373.

[26] Laure Prétet et al. "Singing Voice Separation: A Study on Training Data". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 506–510. DOI: 10.1109/ICASSP.2019.8683555.

Appendix A Architecture of HT Demucs

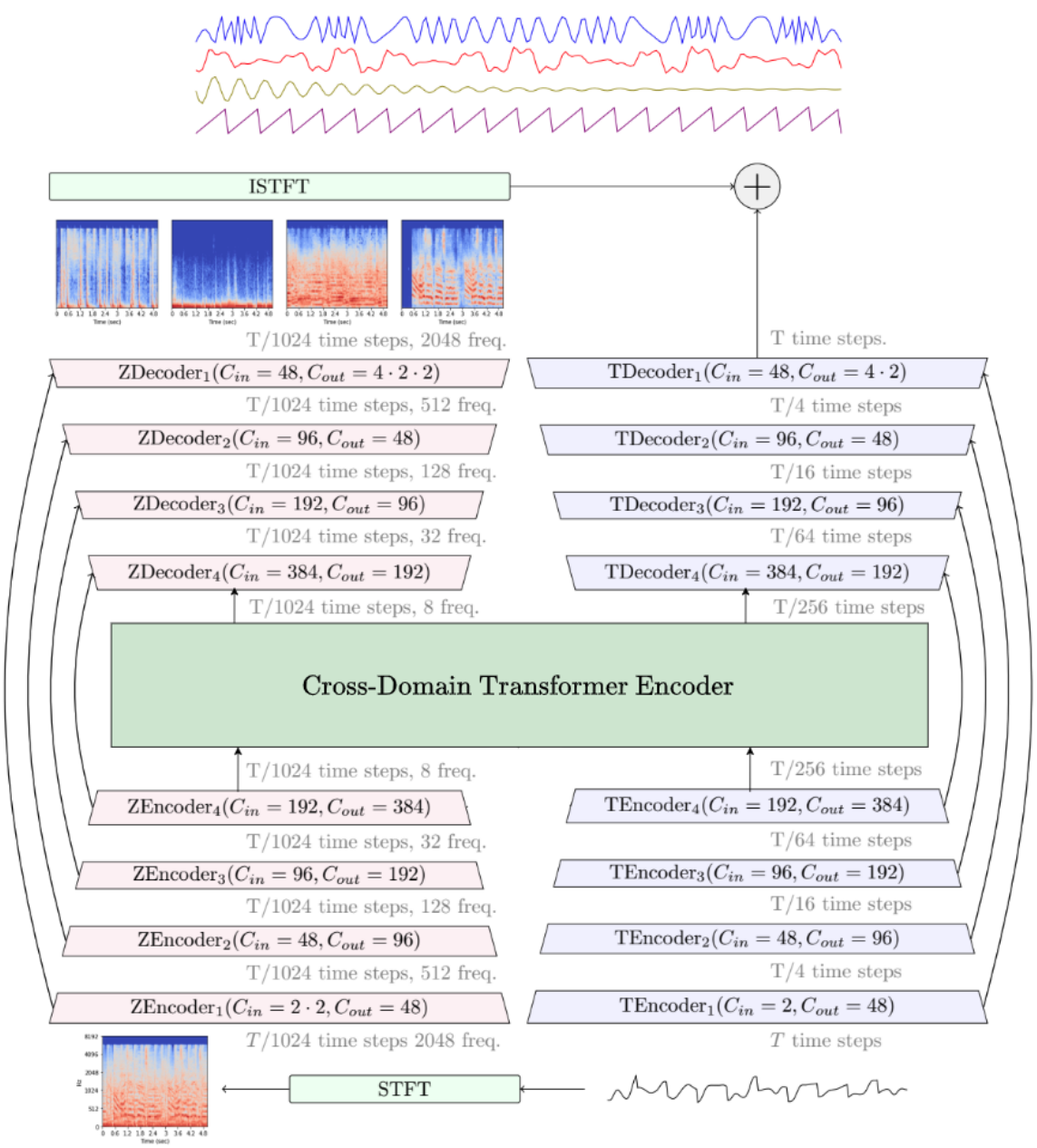

Figure 8: Architecture of Hybrid Transformer Demucs

Appendix B MUSHRA Subjective Evaluation

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

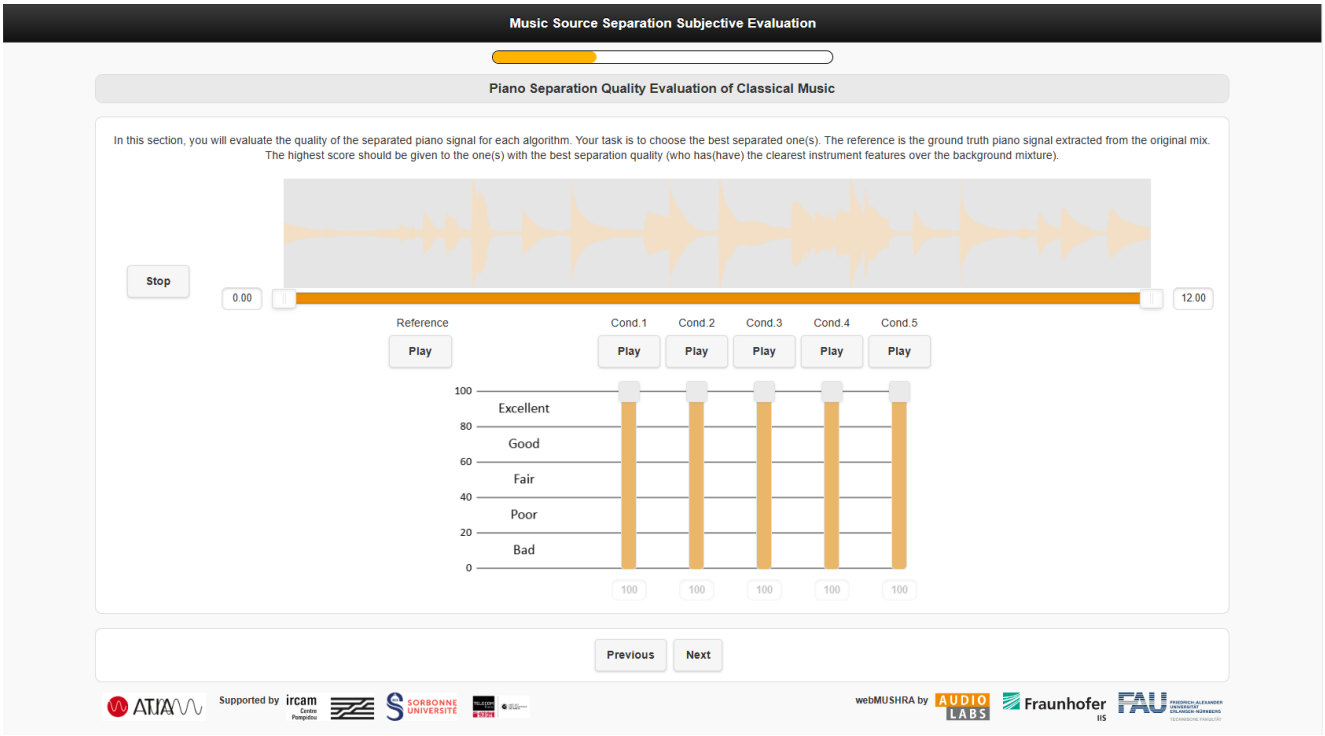

Figure 9: MUSHRA Evaluation Interface