Sound recording and musical souce separation: acoustic or instantaneous mixtures?

Valentin Bilot, Gabriel Dias Neto, Clément Le Moine Veillon,
Guilhem Marion, Yann Teytaut

Abstract

This project presents a quick survey of the problem of source separation. To this end, we will investigate three state-of-the-art signal processing algorithms and compare them for remixing purpose with two means: objective measures and perceptive experiment. We show that these results are not correlated and that low-separated sources with low artifacts tend to be better perceived than high-separated with artifacts sources.

1 Introduction

1.1 Context of source separation

In signal processing, the goal of source separation is to be able to estimate and recover unobserved signals, corresponding to the so-called *sources*, given several *observations* of some mixtures of the latter. We want to retrieve the signals associated with the sources as if each source was alone and under the same conditions.

Source separation can lead to many applications: isolation of several instruments in a polyphonic music, generation of a karaoke version of a song, remix of a track by transforming the isolated sources, etc.

1.2 Problem statement

Throughout this document, our observations, our mixtures of the different sources, will be denoted $x_1(t), \ldots, x_m(t)$, supposing that m is the number of available observations, and will be concatenated in the vector $\mathbf{x}(t)$. Similarly, our n sources to be isolated will be denoted $s_1(t), \ldots, s_n(t)$ and concatenated in the vector $\mathbf{s}(t)$.

Saying that $\mathbf{x}(t)$ is a mixture of $\mathbf{s}(t)$ can be written mathematically thanks to following general model:

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

where $\mathcal{A}_t$ is referred to as the *mixture function*. By knowing $\mathbf{x}(t)$ and gathering information about $\mathcal{A}_t$, with or without hypothesis on the sources, the objective is to estimate $\mathbf{s}(t)$.

Currently, source separation algorithms mostly rely on two mixture approaches, namely *instantaneous mixture* and *acoustic mixture*. If the former can be interesting to use due to the low number of hypothesis it needs, the latter seems to be more accurate for *real* mixtures as it takes into account acoustic properties (sound reflections, reverberation, etc.).

In this work, we aim to realize some stereo sound recordings on which we will apply several current source separation algorithms (based on a given mixture model) in order to isolate each musical source and finally be able to compare the different algorithms (see example on Figure 1). By intervening in the recording of the signals, we also want to collect several pieces of information (the number of sources, their position, etc.) in order to see if these information can improve the performance of the algorithms.

This paper is organized as follows. In section 2, we propose a description of state-of-the-art techniques used for source separation issues. Then, section 3 introduces the goal of and expectations we have for this project. In section 4, the experimental protocol we have been following is summarized ad the obtained results are presented. Section 5 focuses on interpretation of the results we managed to gather. At last, section 6 allows us to conclude our overall study.

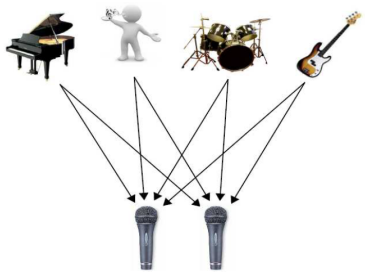

Figure 1: Record stereo sounds and isolate the musical source with several source separation algorithms.

2 State of the art

When it comes to retrieve sources $\mathbf{s}(t)$ from observed mixtures $\mathbf{x}(t)$, it is necessary to get familiar with the notions of instantaneous and convolutive mixtures (see 2.1). Then, depending on the kind of mixture one has to deal with and the context of source separation, two main types of algorithm can be investigated: *non-informed* algorithms like Blind Signal Separation methods (see 2.2) and *informed* algorithms (see 2.3).

2.1 Mixture models

2.1.1 Instantaneous mixtures

The model of *instantaneous mixtures* is one of the most commonly used model which considers that, at a given moment in time t, the observations $\mathbf{x}(t)$ depend on the values of several source signals $\mathbf{s}(t)$ associated with the same time t. In this case, the encountered mixtures are referred to as linear mixtures, that is any observation is a linear combination of the sources at the same time:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

and $\mathbf{A}$ is called the *mixing matrix*.

However, even if instantaneous mixtures constitute a great first approximation of a mixture, they tend to be unadapted to represent real mixtures.

2.1.2 Convolutive mixtures

Contrary to instantaneous mixtures, the *convolutive* ones are aimed at taking into account the acoustic properties involved in the mixtures, hence they are a better representation of real mixtures.

This model is based on a convolutionnal approach to determine the $i^{\text{th}} \in \{1, \dots, m\}$ mixture

$$x_i(t) = \sum_{j=1}^{n} (a_{ij} * s_j)(t). \tag{3}$$

In the frequency domain (STFT[1] filter bench), if we assume that all a_{ij} have slow spectral fluctuations in comparison to the room's acoustic impulse response [1], equation (3) becomes

$$x_i(f,k) = \sum_{j=1}^{n} a_{ij}(f) s_j(f,k) \tag{4}$$

which can finally be written as

$$\mathbf{x}(f,k) = \mathbf{A}(f)\mathbf{s}(f,k). \tag{5}$$

It can thus be seen that, under the given hypotheses, a convolutive mixture corresponds to an instantaneous mixture in each frequency sub-band.

Types of mixtures (summary) The same source separation algorithms can be used for instantaneous and convolutive mixtures but only in frequency sub-band in the latter case.

2.2 Blind Signal Separation (BSS)

Blind Signal Separation (BSS) is a separation method for instantaneous mixtures [2]. The term *blind* means that we have no *a priori* knowledge on the sources, which is compensated by a strong statistical hypothesis as starting point of the BSS methods: **the sources are supposed to be statistically independent**.

2.2.1 Formalization of BSS

Starting from equation (2), we aim at computing a *separation matrix* $\mathbf{B}$ such as

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \hat{\mathbf{s}}(t) \tag{6}$$

is an estimation of $\mathbf{s}$(t).

In addition to supposing that the sources are mutually independent, we consider only zero-mean sources:

$$\mathbb{E}[\mathbf{s}] = \mathbf{0}. \tag{7}$$

However, it is important to note that the sources can only be retrieved of the form "$\mathbf{C}\mathbf{s}(t)$" since we have

$$\mathbf{x}(t) = (\mathbf{A}\mathbf{C}^{-1})(\mathbf{C}\mathbf{s}(t)) \tag{8}$$

with $\mathbf{C}$ a *non-mixing* matrix[2].

In this case, $\mathbf{y}(t) = \mathbf{C}\mathbf{s}(t)$ is said to be a *copy* of $\mathbf{s}(t)$ and, without any more information on the sources' distribution, we cannot expect a more precise result.

Suppose that the $\mathbf{A}$ matrix is known

By knowing the matrix $\mathbf{A}$, one can determine whether it is (pseudo-)invertible or not.

[1] Short-Time Fourier Transform

[2] A non-mixing matrix only has a unique non-zero coefficient for each row and column that compose it.

If so, then we are dealing with a simple linear source separation: choosing $\mathbf{B} = \mathbf{A}^{-1}$ (invertible case) or $\mathbf{B} = \mathbf{A}^{\dagger}$ [3] (pseudo-invertible case) is sufficient for retrieving the sources.

If $\mathbf{A}$ is not invertible, separation is impossible so far. We will not consider such cases in this project.

Suppose that the $\mathbf{A}$ matrix is unknown

If $\mathbf{A}$ is unknown, an IAC (independent component analysis) is pursued: we look for a separating matrix $\mathbf{B}$ that makes the entries of $\mathbf{y}$ independent, so that

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \mathbf{B}\mathbf{A}\mathbf{s}(t) = \mathbf{C}\mathbf{s}(t) \tag{9}$$

and the problem is then solved if matrix $\mathbf{C} = \mathbf{B}\mathbf{A}$ is non-mixing. A powerful theorem illustrated in [2, 4] assures that as long as **at most one of the sources is a Gaussian**.

In the following of this section, we will assume that matrix $\mathbf{A}$ is unknown. Plus, as it does not affect the general case, we will consider spatially white sources:

$$\mathbf{R}_{\mathbf{ss}}(0) = \mathbb{E}[\mathbf{s}(t)\mathbf{s}(t)^T] = \mathbf{I}. \tag{10}$$

The objective is to determine $\mathbf{B}$. First, we whiten our mixtures $\mathbf{x}(t)$ with a *whitening matrix* $\mathbf{W}$ so that

$$\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}\mathbf{A}\mathbf{s}(t) \tag{11}$$

is spatially white. The matrix $\mathbf{W}$ can be obtained from $\mathbf{R}_{\mathbf{xx}}(\tau) = \mathbb{E}[\mathbf{x}(t)\mathbf{x}(t+\tau)^T]$ [1]. As both $\mathbf{z}(t)$ and $\mathbf{s}(t)$ are spatially white, then

$$\mathbf{U} = \mathbf{W}\mathbf{A}. \tag{12}$$

must be a rotation matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$). Hence, the problem is solved by choosing the separation matrix

$$\mathbf{B} = \mathbf{U}^T\mathbf{W} \tag{13}$$

since $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \mathbf{U}^T\mathbf{W}\mathbf{A}\mathbf{s}(t) = \mathbf{U}^T\mathbf{U}\mathbf{s}(t) = \mathbf{s}(t)$.

Summary

- Our goal is to estimate: $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \hat{\mathbf{s}}(t)$;
- We search $\mathbf{B} = \mathbf{U}^T\mathbf{W}$ where $\mathbf{U}$ is rotation matrix and $\mathbf{W}$ a whitening matrix for our mixtures $\mathbf{x}(t)$;
- Our mixtures $\mathbf{x}(t)$ allow to determine $\mathbf{W}$ through calculation of $\mathbf{R}_{\mathbf{xx}}(\tau)$ but $\mathbf{U}$ remains unknown ;
- According to the powerful theorem introduced in [2, 4], if more than one of the sources are Gaussian, $\mathbf{U}$ cannot be found. In order to determine $\mathbf{U}$, the non-gaussianity of the sources must be explicitly exploited through higher order cumulants. This is the purpose of the next section, which introduces contrast functions for determining $\mathbf{U}$ (then $\mathbf{B}$).

[3] Notation † refers to pseudo-inverse: $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$.

2.2.2 Maximum likelihood principles applied to contrast functions

By definition, ϕ is said to be a *contrast function* if

$$\begin{cases} \forall \mathbf{C}, \phi[\mathbf{C}\mathbf{s}(t)] \geqslant \phi[\mathbf{s}(t)] \\ \phi[\mathbf{C}\mathbf{s}(t)] = \phi[\mathbf{s}(t)] \iff \mathbf{C} \text{ is non-mixing} \end{cases} \tag{14}$$

[4]

Such functions constitute objective functions for source separation since they reached their minimum as soon as the audio separation is completed. The source separation is done by minimizing $\phi[\mathbf{y}(t) = \mathbf{C}\mathbf{s}(t)]$ with respect to $\mathbf{U}$ (or $\mathbf{B}$). In this section, we will present the maximum likelihood principles to contrast functions.

Canonical contrasts: mutual information

The maximum likelihood principles rely on contrasts measured as Kullback divergence[5].

By denoting $\tilde{\mathbf{y}}$ a random vector with independent entries whose each entry is distributed as the corresponding entry of $\mathbf{y}$, then it can be shown that the maximum likelihood principle leads to minimize the contrast function referred to as *mutual information* and defined as follows

$$\phi_{MI}[\mathbf{y}] = \mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]. \tag{15}$$

We have $\phi_{MI}[\mathbf{y}] = 0 \iff \mathbf{y}$ is distributed as $\tilde{\mathbf{y}}$, that is the entries of $\mathbf{y}$ are independent. In short, $\phi_{MI}[\mathbf{y}]$ appears as a measure of the the independence between the entries of $\mathbf{y}$.

Should one consider to add the whiteness constraint

$$\mathbb{E}[\mathbf{y}\mathbf{y}^T] = \mathbf{I} \tag{16}$$

then one would use the *orthogonal* contrast function instead, which is given by

$$\phi^o_{MI} = \sum_i \mathbf{H}[y_i] \tag{17}$$

[4] Notation $\phi[\mathbf{s}]$ means that ϕ is a function of the probability distribution of $\mathbf{s}$ like, for example, the expectation function $\mathbb{E}[.]$.

[5] The kullback divergence of two vectors $\mathbf{a}$ and $\mathbf{b}$ measures the "proximity" between their probability distributions and is denoted $\mathbf{K}[\mathbf{a}|\mathbf{b}]$.

where $\mathbf{H}[y_i] = -\int p_i(y_i) \log p_i(y_i) dy_i$ corresponds to Shannon's entropy of the distribution p_i of y_i.

This is what is qualified as *canonical* contrast for source separation, i.e. the mutual information, as it only expresses source independence property without including any other assumption about the distributions of the sources.

High order approximations: cumulants

It is relevant to use high order statistics when it comes to approximate the contrast functions from the mutual information approach.

They are expressed thanks to 2nd and 4th order *cumulants* which are for a given vector $\mathbf{y}$

$$\begin{array}{rcl} \mathcal{C}_{ij}[\mathbf{y}] & = & \mathbb{E}[y_i y_j] \\ \mathcal{C}_{ijkl}[\mathbf{y}] & = & \mathbb{E}[y_i y_j y_k y_l] - \mathbb{E}[y_i y_j]\mathbb{E}[y_k y_l] \\ & & - \mathbb{E}[y_i y_l]\mathbb{E}[y_j y_k]. \end{array} \quad (18)$$

With their introduction, it is possible to determine an orthogonal fourth-order approximation of $\phi_{MI}[\mathbf{y}]$:

$$\phi^o_{MI}[\mathbf{y}] \simeq \phi^o_{ICA}[\mathbf{y}] = \sum_{ijkl \neq iiii} \mathcal{C}^2_{ijkl}[\mathbf{y}]. \quad (19)$$

This will be the function on which we will use descent algorithm with respect to $\mathbf{U}$ (or $\mathbf{B}$).

Summary

- We search the separation matrix $\mathbf{B} = \mathbf{U}^T\mathbf{W}$;
- We calculate $\mathbf{W}$ knowing the mixtures $\mathbf{x}(t)$;
- We estimate $\mathbf{U}$ by minimization of the contrast function ϕ^o, and $\mathbf{B}$ is now fully characterized ;
- sources can be retrieved: $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \hat{\mathbf{s}}(t)$.

2.2.3 JADE and SOBI algorithms

A lot of concepts were introduced in this section. Let us review all of them here in order to highlight the two main BSS algorithms, namely JADE and SOBI.

Overall summary

Blind Source Separation models are mostly considered in the case of instantaneous mixtures. Convolutive ones can also be dealt with similar techniques. They exploit the independence/decorrelation of the sources $\mathbf{s}(t)$ in a whitening step (matrix $\mathbf{W}$), performed prior to the proper separation. However, imposing the decorrelation is insufficient to achieve the separation since a unitary rotation (matrix $\mathbf{U}$) remains to be identified.

JADE algorithm

The very first approach to determine $\mathbf{U}$ is to consider contrast functions ϕ^o as objectives functions that are to be minimized. In practise, this can be done by considering some higher-order statistics such as cumulants from (18) and (19). This constitues JADE algorithm.

JADE algorithm allows to estimate a rotation matrix $\mathbf{U}$ so that $\mathbf{y}(t) = \mathbf{U}^T\mathbf{W}\mathbf{x}(t) = \mathbf{C}\mathbf{s}(t)$ is a copy of the sources $\mathbf{s}(t)$ ($\mathbf{C}$ is non-mixing). There is no guarantee that the determined rotation will lead to the exact sources instead of a copy.

SOBI algorithm

The other approach consists in finding the rotation matrix $\mathbf{U}$ in accordance with the properties of sound recordings and thus allowing to retrieve a much relevant source estimation. This is the SOBI algorithm.

In order to so, several covariance matrices $\mathbf{R}_{\mathbf{xx}}(\tau)$ calculated at several τ follow a joint diagonalization based on Givens rotations and relaxation algorithms [1, 3], which finally leads to identify the relevant $\mathbf{U}$.

2.3 Techniques with prior knowledge on the sources

Recently, plenty source separation techniques were proposed aiming at tackling different particular sources, channels and mixing configurations.

However, the main difficulty is that audio source separation problems are usually mathematically ill-posed [3], including the BSS models introduced earlier. One solution is to incorporate additional knowledge about the mixing process and/or the source signals and espacially regarding. In this section, we will try to give a concise overview of how to deal with separation issue when adding prior information-based methods.

2.3.1 Spectral power models

Several models were proposed to fit with particular source configurations based on prior information regarding spectral power.

Among these ones, we find gaussian mixture models (GMM) or gaussian scaled mixture models (GSMM) which seem suitable for monophonic sources or speech whereas Non-negative Matrix Factorization (NMF) [6] seemed for polyphonic musical instruments.

Unconstrained models [4], hidden Markov models (HMM) [5], harmonic NMF or temporal activation constrained NMF, and source-filter models [7] are also referenced in the literature. A few of these methods have already been combined, however there are still many combinations to try to improve separation methods.

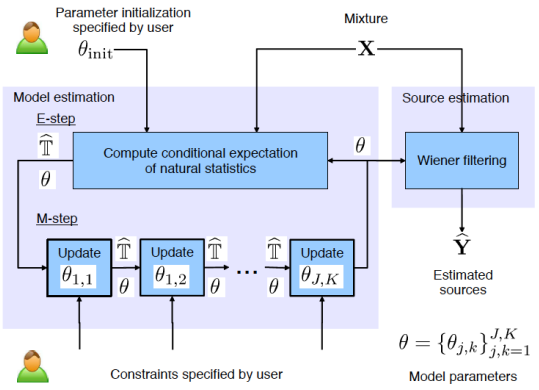

Figure 2: Overview of the proposed in [3] general algorithm for parameter estimation and source separation.

2.3.2 Prior related parameters

Using one or another model, at this step the structure of each source can be encoded via Non-negative Matrix Decomposition. In [3], eight parameters subsets are added via matrix decomposition in respect of a *prior distribution*.

Among those parameters we find mixing parameters, narrowband spectral patterns, spectral pattern weights, time pattern weights, time-localized patterns both for the excitation and filter part of the source model.

2.3.3 Constraints

Prior information about each source can be used by fixing constraints on the parameters. Each parameter can be fixed (unchanged during estimation), adaptive (fully fitted to the mixture) or even partially adaptive (only some parameters within the subset are adaptive).

2.3.4 Estimation of the parameters

In order to process source separation, all parameters must be estimated *via* an estimation algorithm. In [3], an expectation–maximization algorithm is performed, it consists in an iterative method to find maximum likelihood (ML) or maximum *a posteriori* (MAP) of the whole parameters set.

First, given initial parameter values, the model parameters θ are estimated from the mixture $\mathbf{x}$ using the algorithm where the E-step consists in computing a *conditional expectation of the natural statistics* $\mathbb{T}$, and the M-step consists in updating the parameters θ given $\mathbb{T}$ by alternating optimization of each of the parameter subsets (see Figure 2).

Finally, given the mixture $\mathbf{x}$ and the estimated model parameters θ, source estimations $\mathbf{y}$ are computed using Wiener filtering [3].

Prior knowledge and FASST library (summary)

All in all, [3] aims at providing a powerful library, FASST (Flexible Audio Source Separation Toolbox), which is based on the above-mentioned concepts of prior information regarding the mixtures or the sources. One of the key information to possess is the number of sources to be retrieved.

JADE and SOBI algorithms (section 2) are general algorithms with few hypotheses while FASST library provides more specific algorithms with more infomation, like the NMF algorithm we will rely on.

Hence, naturally, better separation results are to be expected with the use of FASST library but its prior-informed algorithms are less general than JADE and SOBI algorithms which can be applied to any mixtures without knowing any information.

3 Project presentation

3.1 Goals and expectations

Facing the extent of the source separation problems and the variety of configurations to examine, we will focus on one classical situation: starting from a full acoustic recording configuration and applying different algorithms to perform the best source separation.

Then comes an interesting question: what is a good separation? The most we can assume is that an evaluation necessarily depends on what we aim to do with the separated tracks. On the one hand, we chose to perform separation for remixing purpose so that our evaluation is based on a perceptual evaluation of new mixes made with extracted tracks. On the other hand we evaluate separations with the standard evaluation process *BSS eval* (see subsection 4.2 for more details).

3.2 Recording session

All recordings took place at IRCAM, in the studio 6 which is totally sound proof with very few reverberation. This place is perfectly adapted to our purpose as it allows evaluation of source separation algorithms on both dry sources and artificial-reverberated sources.

We chose to record a jazz classic (*Black Orpheus*, Louis Bonfi) with a pared-down trio including electric guitar, cello and clarinet using Neumann KM 184 stereo ORTF couples on each source directly plugged into the RME Fireface 802 preamps. It is to be noticed that the set up we chose was not properly a traditional recording configuration as we wanted to catch every source in each observation channel (microphone) in order to be able to perform separation with algorithms such as JADE or SOBI.

This led to keep distance between the microphones and the sources, enough to catch every instruments while keeping a predominant one.

4 Methods and results

4.1 Convolutive approach

4.1.1 Spectral smoothness method

As exposed in 2.1.2, under the adequate hypotheses, dealing with the convolutive approach on algorithms such as JADE or SOBI is equivalent to apply them instantaneously to each frequency sub-band.

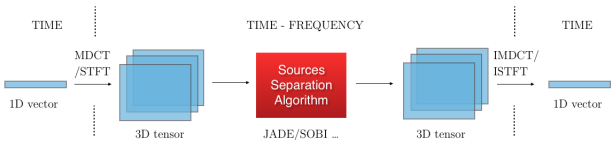

Figure 3: Source separation algorithms performed on each frequency bin.

The main issue remains the absence of guarantee regarding the order of the sources after being processed through the algorithm. Facing this problem, we have tried to develop a program *Spectral Smoothness* that sorts sources according to audio features (centroïd) continuity over time.

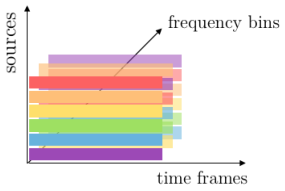

Figure 4: Tensor Y of sources randomly stored in each frequency bin 2D layers.

The idea is to observe the evolution of the centroïd over time, the problem can be solved via tree exploration. Let us assume that we have to sort 3 sources, to do so we explore bands that are groups of d successive frequency bins. We denote p the tree exploration matrix so that $p[br, k] \in \{0, 1, 2\}$.

The centroid for a branch br at a time t is then computed according to the following formula:

$$c_b(br, t) = \frac{\sum_{k=1}^{d} Y[p[br, k], t, k] f_k}{\sum_{k=1}^{d} Y[p[br, k], t, k]}. \tag{20}$$

Then we compute an average of differences between centroids at successive times on l time duration starting from t_0 for each branch br of the tree:

$$\Delta_b(br) = \frac{1}{l} \sum_{t=t_0}^{t_0+l-1} |c(br, t+1) - c(br, t)|. \tag{21}$$

Therefore, finding the right path through the tree means minimizing Δ_b. At this step, we have sorted

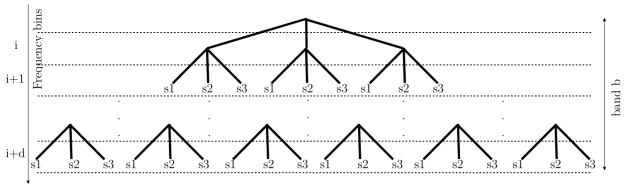

Figure 5: Tree representing a band b of depth d in the case of 3 sources.

all sources for the band b. Afterwards, it remains to sum over bands taking already computed centroids into account when computing new ones.

4.1.2 Convolutive or instantaneous?

We missed time to properly test different settings of this *Spectral Smoothness* method. Indeed, for now we are still not able to find the right order of sources over frequency bins. Hence, we chose to restrain ourselves to the instantaneous mixtures approach.

4.2 Outcome measures of source separation algorithms

Several source separation algorithms have been used but how to say that one is better than the others? In order to be able to highlight relevant conclusions, two outcome measures have been pursued. On the one hand, we used the *BSS eval* as a quantitative approach. On the other hand, we cannot ignore the fact that we are reaching a tight issue implying perception capabilities, hence perceptive tests should also be discussed.

For both evaluations we chose to focus on three different separation methods : JADE, SOBI and FASST NMF. JADE and SOBI are standard *BSS* methods fed with as much observations as sources to separate. FASST NMF is rather a more elaborated technique fed with a stereo mix. Hence, the tasks we try to evaluate are not entirely the same as FASST NMF is given very less information than the two others.

4.2.1 Objective quantitative test

BSS eval is a standard Matlab toolbox dealing with the performance estimation of (blind) source separation algorithms within an evaluation framework where the original source signals are available as ground truth. The measures are based on the decomposition of each estimated source signal into a number of contributions corresponding to the desired source, interference from unwanted sources, and artifacts (noise, clicks, etc.).

In our situation, we do not use proper source signals but observations with a predominant source (although there is no absolute acoustic source signal). As a consequence, we are not able to get significant measures for a given separation method, although the comparison between methods is still relevant and absolutely useful for us to draw conclusions.

4.2.2 Subjective perceptive test

To complete the objective quantitative evaluation, we chose to evaluate source separation algorithms through their remixing application using a perceptive test which can be phrased with the following question: *when remixing separated sources, which source separation technique is the best ?*

The test consists in listening to 3 15-second remixes for 4 separation cases (JADE, SOBI, NMF, and the ground truth) in 2 acoustical contexts, namely anechoic (without reverberation) and room-like (with reverberation). Our recordings are given to all of these algorithms which return the separation of the 3 instruments. We used these files to reconstruct 3 new mixes, with the same mixing methodology for every algorithm. The ground truth refers to remixes made out of the original recordings, and is used as control experiment.

This idea of using reconstructed mixes for the subjective test is to compare algorithm depending on the perceptive quality in a musical context.

For each mix the subjects listen to, they have to note in a given range two overall subjective factors which are *quality* and *defaults*.

- The relative criterion *quality* highlights if the audio file is more or less pleasant to listen than the firstly listened original/reference mix. Its range is $[-5, +5]$. A -5 rating implies that the subject found the remix awful while a $+5$ rating implies that the subject found the remix much better than the original. A 0 grade means that no differences have been detected.

- The absolute criterion *defaults* highlights if the audio file contains unpleasant effects such as artifacts, noise or unusual filtering. Its range is $[0, 5]$. A 0 rating implies that no defaults have been detected while a 5 rating implies that lots of defaults

have been heard.

The test follows a double-blind methodology that implies that the data are shuffled and neither the subject and the experimenter know the order of the mixes, in order to avoid biases related to their influence.

4.3 Results

4.3.1 Objective evaluation results

Figures (6), (7) and (8) show the quantitative results obtained when applying the BSS evaluation to separated sources compared to the original recordings.

As mentioned before, absolute values are not relevant since we do not use the real sources as references in the algorithm. Instead we use the closest approximation that we have, which is the raw audio recording with the microphones just in front of the instruments for the anechoic case, and the same audio supplemented with their reverberation sends for the reverberated case.

Thus, comparisons between the reverberated and anechoic values are not relevant because the reference signals used are not ground truth signals and are not the same for these two cases.

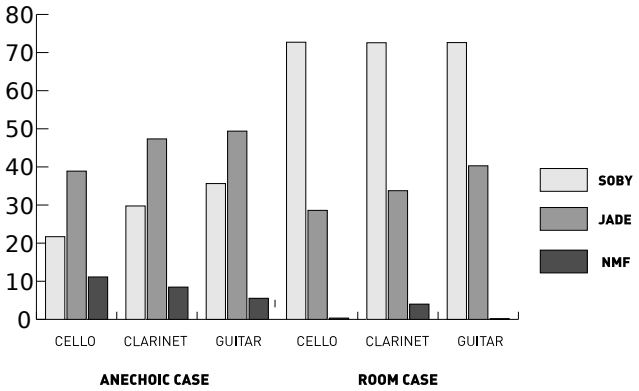

Figure 6: BSS Eval: Signal to Artifacts Ratio (SAR) for each source (left to right: clarinet, cello and guitar).

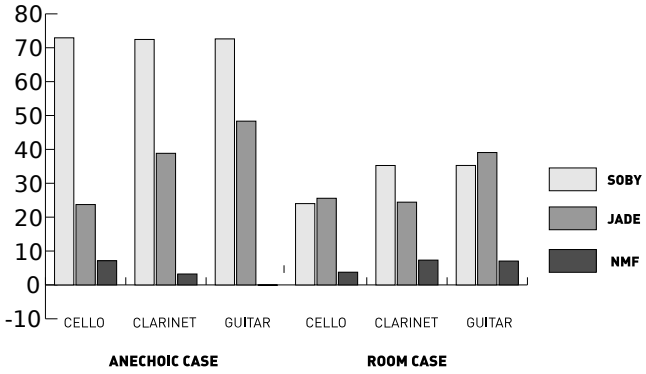

Figure 7: BSS Eval: Signal Distortion Ratio (SDR) for each source (left to right: clarinet, cello and guitar).

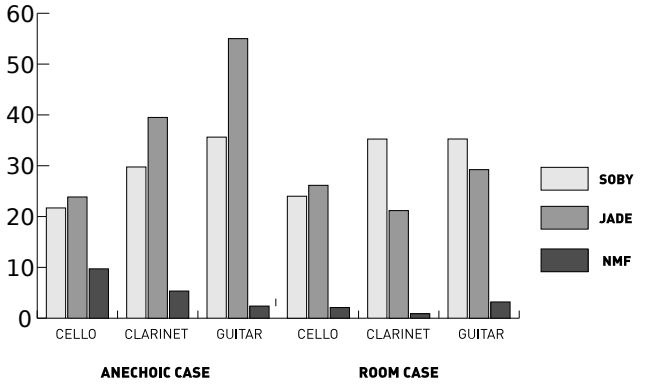

Figure 8: BSS Eval: Signal to Interference Ratio (SIR) for each source (left to right: clarinet, cello and guitar).

4.3.2 Perceptive tests results

A number of 16 people went through the test. All of them are working at IRCAM and are highly related to music, so this sample is not so representative but we can expect to get relevant data due to the knowledge of music.

In order to process the data, we use define a score for each listening

$$\text{score} = \frac{\text{quality} + 5 + (10 - (2 \times \text{defaults}))}{2} \tag{22}$$

which gives an idea of the global quality of the listening. We want to be able to compare the global quality of every algorithm depending on the acoustic parameter. For so, we claim that our data are following 8 random variables (one for each algorithm and acoustic parameter) denoted

$$X_{J,A}, X_{G,A}, X_{N,A}, X_{S,A}, X_{J,R}, X_{G,R}, X_{N,R}, X_{S,R}$$

Our experiment gives us a sample of each of these random variables that we assume to follow an unknown distribution of mean μ and of finite variance σ^2.

The Central Limit Theorem assumes that the empiric mean $\tilde{\mu}$ computed on samples is converging in law to a Normal distribution $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ with n the size of the the sample. In order to compare our results we need to define an confidence interval with Θ such as

$$P(\tilde{\mu} - \Theta \geq \mu \geq \tilde{\mu} + \Theta) \approx 0.95 \tag{23}$$

Using pivotal quantity, Central Limit Theorem, and the Student's t-distribution, we find that the following Θ satisfy our request:

$$\Theta = 1.96 \cdot \frac{\sigma}{\sqrt{n}} \tag{24}$$

We can so compare our results using empirical mean $\tilde{\mu}$ using the confidence interval $[\tilde{\mu} - \Theta, \tilde{\mu} + \Theta]$.

The figure (9) shows the previously explained global quality for every algorithm.

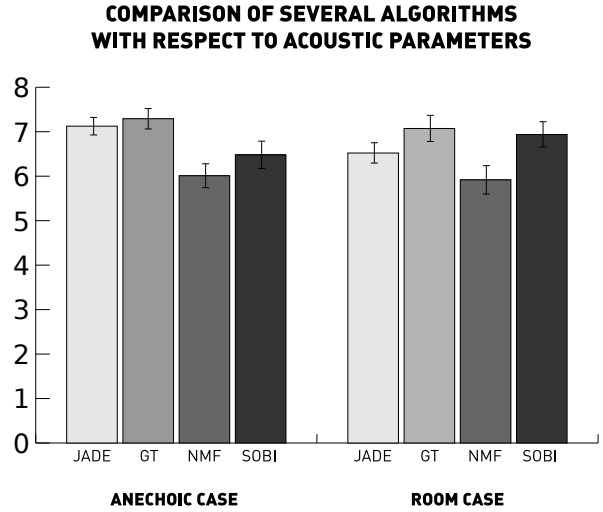

Figure 9: Perceptive tests results evaluating global quality.

5 Discussion

5.1 Qualitative approach

According to the objective BSS eval, NMF algorithm based on the FASST library is the one computing the best separation regarding all criteria (artifacts, distortion and interference). Under each condition (with and without reverberation), FASST-NMF has the lowest SAR, SDR and SIR for each source by far. JADE and SOBI algorithms seem to pursue a close separation, quality-speaking, regarding the same criteria although SOBI algorithm does introduce more distortion than JADE and seems to struggle with the cello. JADE and SOBI have more difficulties to deal with the guitar according to the criteria we chose, resulting in a higher SAR, SDR and SIR for the guitar source.

Moreover, comparing the anechoic and reverberated cases, we can see that for the SAR and SIR values, SOBI gives better results than JADE in the anechoic case, and JADE gives better results than SOBI in the reverberated case.

We must be careful drawing conclusions in this case since we cannot rely on absolute values. There are different possible scenarios that can explain those results. It is possible that one of the algorithms has the same performance for both reverberated case and anechoic case, and that the second one has significantly different performance depending on the reverberation. Maybe SOBI is really bad at dealing with reverberated signals, maybe JADE is really better at dealing with reverberated signals than with anechoic ones, or maybe the truth is between those two possibilities. Hence, our results does not allow us to draw more conclusions.

5.2 Perceptive approach

As we only proceed the test on a few sample of persons (16), our confidence interval are a little bit small in order to conclude but, in any case, we can assume some statistical trends.

First, we can see that NMF is definitely a little bit worst than the other ones. We can interpret it from the artifacts added with this algorithm, however, this algorithm is the one seems to separate sources the best (according the objective measures).

We also see that the mix with no changes added (no algorithm used) seems to be better perceived than the ones made from sources separated with an algorithm.

We can investigate that for musical purpose, it is better to do mix with low-separated sources without artifacts, than well-separated sources with artifacts, and that in some case (as ours) it is not relevant to use a source separation algorithm in order to perform another mix.

5.3 Comparison between both approaches

The first thing to mention is the opposition between both evaluations results. As FASST NMF seems to objectively perform the best separation among the three techniques, the perceptive test tends to show that the two others, JADE and SOBI are a bit more suitable for remixing purpose. This assumption seems to corroborate our first intuition: evaluating a separation definitely depends on what we aim to do with the extracted tracks. Actually, a good separation does not guarantee that the extracted tracks are adapted to a remixing context.

Furthermore, the assumption concerning JADE being better than SOBI in the reverberated case, shown in the objective test, is in contradiction with the perceptive test results that tend to show that JADE is evaluated better than SOBI for the anechoic case.

We must keep in mind that there are only a few applications of musical source separation methods among which we find remixing and karaoke. Consequently, the true evaluation of such methods must be perceptive ones even if they go the opposite way than objective quantitative measures.

6 Conclusion

During this work, we implemented several state-of-the-art source-separation algorithms from the literature in the interest of making a quick survey. In order to control the parameters of the original sources, we recorded them in a proper studio using professional studio recording gears. We compared these algorithms with respect to acoustic parameter (anechoic and room) with two means: objective measures coming from signal processing techniques and a subjective experiment involving music-related subjects.

With this study we showed that objective measures does not correlate so much the perceived quality in a musical context, and that people tend to prefer a mix made with low-separated sources with small artifacts, than high-separated sources with artifacts. This tend to understand that having separated sources is not a highly important point for mixing. We also showed that the algorithms SOBI and JADE are preferred by people rather than Fasst with NMF parameter.

Some improvements would be to investigate some higher level algorithms that ask more data or computational power, such as Neural Networks based ones, in order to have a more complete survey. In addition, our experiment provides us with a nice trend but a more complete test involving more people would give better insurance for conclusion.

References

[1] *Séparation de source audio.* R. Badeau, Master ATIAM course, 2019.

[2] *Blind signal separation: statistical principles*, J.-F. Cardoso, Proceedings of the IEEE, vol. 86, no. 10, Oct. 1998

[3] *A general flexible framework for the handling of prior information in audio source separation*, A. Ozerov, E. Vincent, and F. Bimbot, IEEE

[4] *Underdetermined instantaneous audio source separation via local Gaussian modeling* E. Vincent, S. Arberet, and R. Gribonval, in Proc. Int.Conf. on Independent Component Analysis and Blind Source Separation (ICA'09), 2009, pp. 775 – 782.

[5] *New EM algorithms for source separation and deconvolution* H. Attias in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), 2003, pp. 297–300.

[6] *Polyphonic transcription by nonnegative sparse coding of power spectra*, S. A. Abdallah and M. D. Plumbley in Proc. 5th International Symposium Music Information Retrieval (ISMIR'04), Oct. 2004, pp. 318–325

[7] *Source/filter model for unsupervised main melody extraction from polyphonic audio signal*, J. L. Durrieu, G. Richard, B. David, and C. Fevotte, IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 3, pp. 564–575, 2010.

[8] *Basic stereo microphone perspectives – a review.* R. Streicher W. Dooley, JAES vol 33, no7/8 pp548-556, 1985.