Master's Thesis

Computer Engineering Department
Università degli Studi di Padova

Discipline : Sound and Music Computing

Gaussian Framework

for Interference Reduction in Live Recordings

Diego Di Carlo

Advisor Nicola Orio, Università degli Studi di Padova

Members of the jury:

Chair : Maria Elena Valcher, Università degli Studi di Padova

Examiner : Stefano Vitturi, Università degli Studi di Padova

Examiner : Simone Gerardin, Università degli Studi di Padova

Examiner : Andrea Cester, Università degli Studi di Padova

Examiner : Emanuele Menegatti, Università degli Studi di Padova

July 10th, 2017

To Giorgia

Your are the means,
I am the variances,
We define together
the best Gaussian Process ever.

Acknowledgement

This is for all the people who make this work possible.

First, I would like to thank Antoine Liutkus for being more than a supervisor and a *source* of inspiration: thank you so much for giving me so many *fresh* opportunities; I hope you will understand if I do not itemize them all here. In particular for giving me the change to be part of the excellent team of *Multispeech* and introducing me to the French signal processing research grou.

With equal merits, I would like to thank my Italian supervisor, professor Nicola Orio: thank you so much for having always supported me and having gone thought all my absurd paperwork. Thank you, dear professor, for making this thesis possible; actually everything started from you.

Special thanks to Thomas Prätzlich, who provided the basis studies and materials I used for starting my work: seeing you in the audience during the presentation of this work was an honor; I hope I did not ruin something.

I would like to thanks all the the members of Multispeech team which have somehow influenced the development of this thesis: in particular, thank you *fresh* guys, Ken, Matthieu and Aman, for sharing difficult coding session and paper-writing moments. Moreover a special thank goes to Arie: your help in understanding this topic was essential. Theo, Nathan and Iñaki, thank you too for the time spent together.

The unending and honest support, love and encouragement that I was given by Giorgia cannot be acknowledged enough. She was always present and helped me through the good and the bad and made it worth it. This is for you... Finally you can understand what I play with my bass.

A special thank my family, Marta, Franco, Alberto and Candy who always make feel well and safe.

Last but not least, I would like Erasmus+ and CYCI2 project for funding my work.

Abstract

In this study, typical live full-length music recordings are considered. In this scenarios, some instrumental voices are captured by microphones intended to other voices, leading to so-called “interferences”. Reducing this phenomenon is desirable because it opens new possibilities for sound engineers and also it has been proven that it increase performances of music analysis and processing tools (e.g. pitch tracking). Extending state-of-the-art methods, we propose an NMF-based algorithm that iteratively estimate each source contribution, i.e. the power spectral densities (PSDs), and the corresponding strength in each microphone signal, modeled in a interference matrix. Unfortunately our approach suffer of a huge computational load. To address this issue, We show that using random projection method the method is able to process full-length live multi-track recoding in a acceptable time. Experimental results demonstrate the efficiency and the effectiveness of the approach.

Contents

Introduction

Figure 1: Please let me introduce you our mock jazz band, *The Naked Orange Horse*: a saxophone quartet (soprano, alto, tenor, baritone), double bass and piano.

Prologos

Suppose your favorite jazz band, *The Naked Orange Horse*, is recording their last break-avantgarde-jazz-core hit, *The Trans Graph*, as in Figure 1). In order to record their outstanding musical and communicative performance, they need to play all together in the same room. Unfortunately using just a simple microphones, it is not enough to capture every instruments: their sound engineer must carefully close-mike every instruments or part of them. He knows well his band and the way the musicians play, for instance he knows that he is going to increase the gain of the mics of the piano because of the pianist delicate touch.

Now the band is creating his magic. Shapes, colors, notes, chords, glances of intent and creating sweaty hands. *The Naked Orange Horse*'s recording sessions is brilliant as usual.

The raw material, impregnated with magics, is crafted. Now it is time for the sound engineer to give it the right balance, air and space... and make it audible to common people on hi/lo-fi stereo.

It happened that the improvised piano solo is such epic and intense (too divine to be described). Unfortunately the gentle touch of the finger of the keyboard get lost overtaken by the djent riff of the bass and by some soprano sax random noise. No problem, the sound

engineer is going to fix it: he is going to boost piano sound. He solos the piano microphone: together with a close, clear piano sound, the distant but present the sound of all the other instrument can be heard, especially the bass and the soprano sax. When he now mixes the boosted piano recordings with the other instruments: the perfectly matched background accompaniment sound degrades into a overall muddy sound.

The problem is happening because the sound of the other instruments leaked into the piano mic from across the room. It's as if the piano mic has become a "room mic" for the other instruments. The amazing sound waves are slower than their posh light sisters and they suffer of microphone distance even in small room: thus, sound delays occurs because the same sound is recorded from two different point. When these recordings are mixed together muddy reverberation annoys the glutton listeners.

To keep the recorded awesome performance tight and upfront while boosting the divine piano solo, it's important to reduce leakage, in other words, to increase the isolation between microphones.

... The sound engineer has a strong expertise in mixing and find solution to these issues and finally he was able to mix. However the piano solo could not be boosted enough and many notes can still not be heard.

The Naked Orange Horse's pianist comments *never mind, actually it was too soon to reach the undisputed fame*. Such humility!

Parodos

This thesis will introduce you to the problem of the *interference reduction* (IR) which aims to remove the contribution of undesirable signals from the observation of a target one. In other words, the objective of the IR is to enhance, ideally to isolate, the signal of interest when in presence of other sound contributions. The IR finds application in physics, in electronics and in telecommunication etc, where the term *interference* may be used to indicate slightly different phenomenon. In music signal processing, it is also know as *bleeding*, *crosstalk*, or *microphone leakage*. Within this filed, IR is then closely related to two well-know problems: *source separation* (SS) whose objective is to separate a signal *mixture* into its constituent component, called *sources*, and *noise reduction*, also referred as denoising, which aims to remove an noise component which corrupt a signal of interest. Leakage is then the overlap of an instrument's sound into another instrument's microphone. It's an unwanted sound from instruments other than the one at which the mic is aimed. For example, the piano mic also "hears" the bass and the saxophone; the baritone saxophone mic also hears the alto saxophone, and so on.

In this work, a fresh framework to overcome the IR problem will be presented. Such a theoretical framework is a general formalism for SS in which the sources are modeled as the realization of *Gaussian processes* (GP)s. Thanks to that, you will see that complex

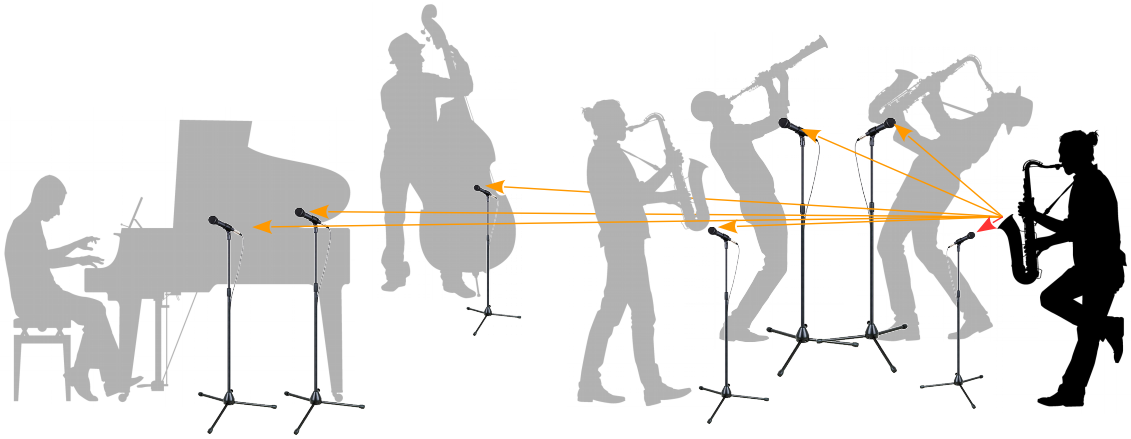

Figure 2: Interference example: from one source to all the microphones. Contribution to the close-mic is highlight with darkest arrow

mixture signals can be separated in a flexible and natural way, taking into account some *a priori knowledge*. Once the separation is obtained, then isolation will be obtained for free of charge.

In the music signal processing context, reducing interferences between various sound sources could be desirable. In fact on one hand, as explained in the Prologue, the interferences often greatly constraint the mixing possibilities for a sound engineer; on the other hand they reduce the accuracy of tools necessary for further audio processing and analysis, such as pitch trackers. Unfortunately, it is difficult today to achieve general optimal quality, often at the price of ad-hoc solutions or a posteriori method demanding important computing resources. Recently the GP framework for separation has been proposed for addressing this issues, and some methods have been proposed in the literature. However these state-of-the-art algorithms contain some ad-hoc choices and heuristic parts.

In this work, the reader will have a digest of the theoretical formulation of the GP framework. Thank to that, a deep study is conducted on how it may be used to yield provably optimal algorithms to learn all the parameters required for good interference reduction. Moreover a typical *full-length multitrack live recording* scenario will be taken presented and the related implementation issues will be discussed. However, the reader will not be left unsatisfied from a practical point of view: an open-source python implementation is provided which he can use for his most brewed applications.

In this introduction, I will first present some motivations which highlight the importance of this topic and will accompany the reader all along this work. Thus, I will draw up a rapid state of the art in order will be drawn to provide a more complete context and try to honor previous researcher and their works. Thus, some notes on *Probability Theory* will be presented providing a simple but important basement for the further concept. Finally, this introduction concludes with the presentation of the plan of the presentation.

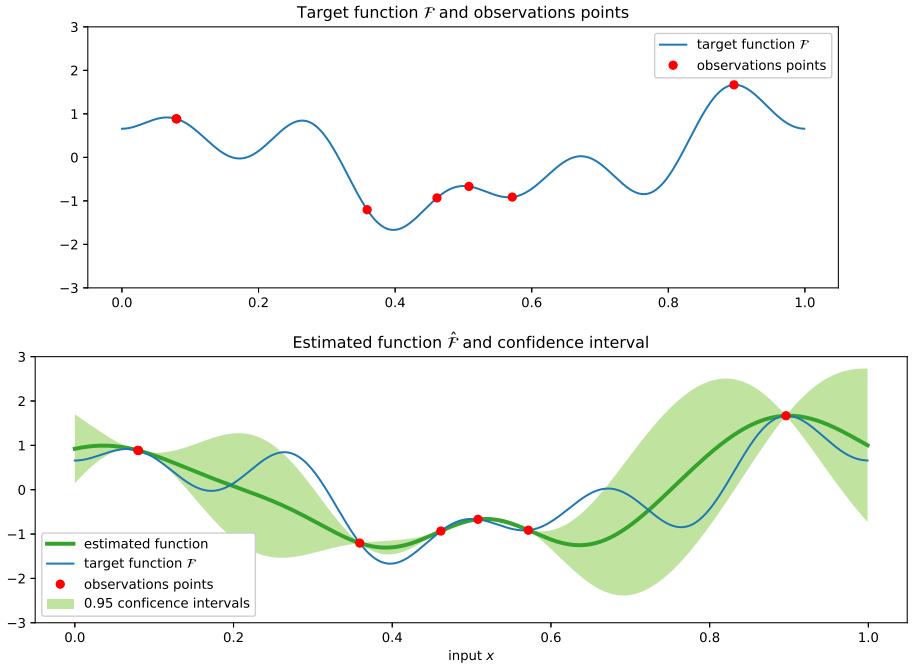

Figure 3: GP for the regression problem: the above plot represent the target function to be estimated and 6 observed points. Form these points GP try to guess the target function: in the lower plot the dark line indicates the posterior mean and the green-shaded area is the posterior variance of the estimated function.

Motivations

This thesis addressed the problem of the interference reduction (IR) in live music recordings as a problem of sound source separation (SSS). Indeed, we will use a probability framework developed for the SSS problem to address this our problem. My interest in SSS pushed my into this framework. Even if it is rare today to achieve sufficient separation quality for straight forward applications (e.g. instrument decomposition form a song), in others, this approach can bring very rewarding results: IR is one of them.

There have been previous attempts to perform interference reduction, but it is not a solved question, and there is still much room for improvement. But first lets analyze the meaning of the thesis title with the following paragraph.

Why Gaussian? Roughly speaking a random *process* is a generalization of a probability distribution (which describes a random variable) to *functions*. They are commonly used in machine learning to model (infer, learn) unknown functions. In simple words: let's have some given learning points, such as in Figure 3-a; our goal is to infer the function that generates them, that is to estimate the values taken by the function at any other point of interest.

That is a typical machine learning task and many approaches can be used: they all differ on which kind of (how to use the) information form the observed point or other

knowledge is used. In the case of *Gaussian Process* (GP), they are common tools to model those functions whose *mean* and *covariance* are known *a priori* [1]. One way to imagine this modeling is that many random functions are drawn from the prior (e.g. mean and covariance), and the ones which do not agree with the observations are rejected.

By focusing on processes which are Gaussian, it turns out that the computations required for inference and learning become relatively easy. Thus, the supervised learning problems in machine learning which can be thought of as learning a function from examples can be cast directly into the GP framework.

As explained in [1], GP theory provide a practical and probabilistic approach to learning in *kernel machines*, an area of machine learning where learning methods are build on kernel functions, such as *support vector machines* (SVM) and *principal component analysis* (PCA). In particular they provide a simple and effective framework for regression and classification as well as an effective tool for optimization. GPs were first proposed in statistics by Tony O'Hagan in the sixties [2]. However they are well-known to the geostatistics community as *kriging* (see figure **??** for an example) and their use can be traced back at least to works by Wiener in 1941 [3] with a different formulation.

In this work, IR will be formulated as a problem involving GP regression. This approach was proposed originally in [4], while the idea of using GP for more general SSS is described in [5]. Starting from these works, a formal model for the IR problem can be formulated and, as we will see, performed even on very large signals, i.e. ordinary multi-track live recordings.

Why Interference? The production of modern music often involves a number of musicians performing together inside the same room with a number of microphones set to capture the sound emitted by their instruments. This typical studio condition promotes spontaneity and musical interaction between the musicians, but also it optimizes studio time usage. This is typical for classical or jazz music recordings, where the interaction is essential. For live musical performances, for obvious reasons, there are no alternatives.
In these situations each musician gets its dedicated microphones, so different voices may be optimized independently and on-demand by sound engineers. Ideally, each microphone should pick up only the sound of the intended instrument, but due to the interaction between the various instruments and room acoustics, each microphone picks up not only the sound of interest but also a mixture of all other instruments (see Figure 5). Indeed sound engineers have a strong expertise in designing specific acoustic setups to minimize them. However, unless the musicians do not play in the same room, which is detrimental to musical spontaneity, interferences are bound to occur in practice. Thus, reducing sound interference may be desirable for several reason.

First, from the sound engineer perspective, this phenomena greatly reduce the mixing possibilities. For instance lets image that tuning up the gain of the piano microphone, the

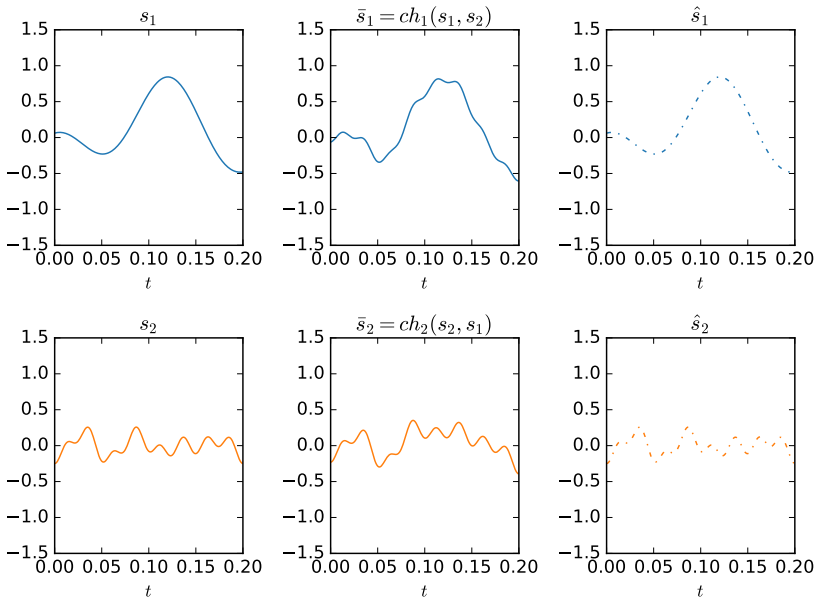

Figure 4: Signal-level example of interference reduction problem. Here two original signal $s_1(t)$ and $s_2(t)$ are transmitted respectively thought two different channels ch_1 and ch_2. Because of the poor isolation of the channels, both the signal are mixed together, in the mixture $\bar{s_1}$ and $\bar{s_2}$. The aim of interference reduction is to estimate the original signal from these mixture, yielding to $\hat{s_1}$ and $\hat{s_2}$

.

bass sound comes up as well. This lead to several problem such as *phase interferences*, *off-axis colorations* or *ghost tracks*. Thus removing interferences, mixing tools, such as compressors, EQs, can be used without limitations, i.e. without any fear of amplifying unwanted sounds nor creating artifacts.

Second, the idea of isolating some source or at least increasing the presence its presence in a mixture yield to better results of many *music information retrieval* (MIR) and *semantic audio signal processing* tasks, such as chord detection, melody extraction, genre classification, instrument identification, pitch tracking and many others. For instance in the melody line estimation some of the difficulties are derived from presence of accompanying components. In this work[6], it has been shown how performance can greatly improve using a SSS method as a preprocessing.

Third, interference prevent the total removal or complete isolation of a voice from the recording. This is because of the so called ghost tracks: even thought the recording of an instruments by its dedicated microphone is removed, its contribution is still present in the recordings of another instrument. Ghost tracks can limit many audio mixing process, such as *overdubbing*, as well as several MIR tasks. Isolating or removing tracks may be useful even for pedagogical reason: can be used by musicology and young musicians as well. In particular, a practical contribution of this work was to develop a tool for Creative Dynamics of Improvised Interaction (DYCI2) project, a collaborative research and

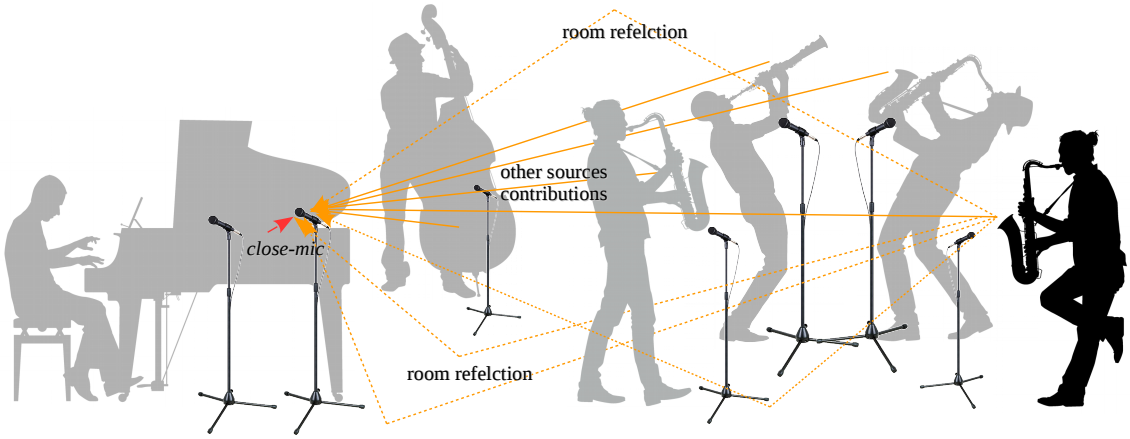

Figure 5: From all the sound sources to the pianist microphone. Close-mic contribution (darkest line) and room reflections (dotted lines) are depicted as well.

development project funded by the French National Research Agency (ANR). In fact, the algorithm developed in this thesis will be used within this project to increase the performance of automatic music improvisation tools. This automatic systems are based on musical features extracted from audio recordings, and the qualities depends on the quality of this extraction processing. Providing good isolation is then desirable to achieve good performances for the overall system.

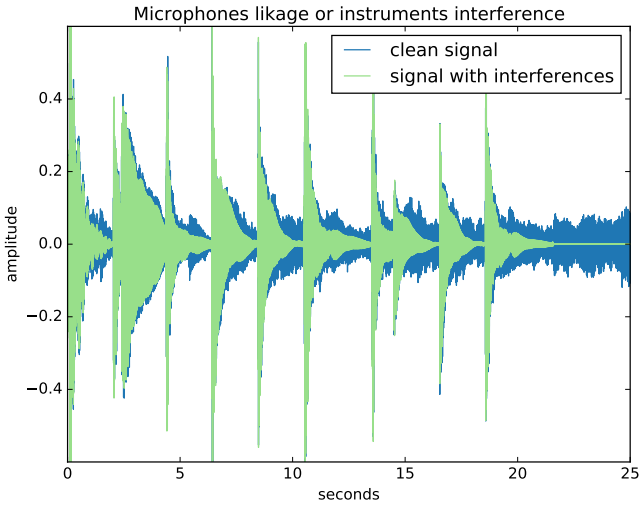

Figure 6: Comparison between clean piano signal against when interferences occours.

Application of IR can be found in many other fields, such as speech enhancement [7] , hearing aid sound processing [8], or for telecommunication [9].

Why separation? Source separation (SS) is a very intense subject of research dealing wit the problem of recovering several unknown *sources* signals underlying a given *mixture*. (see [10] for a review). It is a core problem in several research areas, such as audio signal

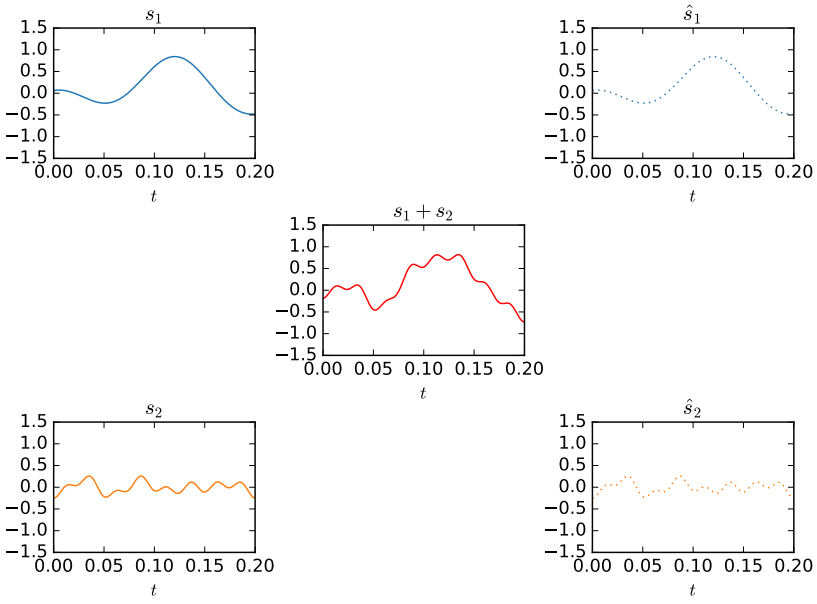

Figure 7: Signal-level example of source separation problem. Here two original signal $s_1(t)$ and $s_2(t)$ are summed. The result is a mixture signal $s_1 + s_2$. The aim of source separation is to estimate the original signal from these mixture, yielding to $\hat{s_1}$ and $\hat{s_2}$

processing, telecommunications, geostatistics or biomedical signal processing. In fact it has a very strong ties with the vast field of *reverse problems*: in this case is the operation to be inverted is that of mixing of source signals, whatever they represents.

In [10] several application of SS are depicted. In telecommunications, the typical scenario is receiving a target signal which has been contaminated by the addition of more parasitic signals. It is then necessary to separate the target signal from this mixture. A similar situation happen in geostatistics, where the measured terrains heights are often captured with uncertainty about the position or value of the measurement. It is then necessary to deduce the desired value from these noisy measurements. Again it is a matter of separating the useful signal from a *noise*. In application such as biological signal processing, the mainly interest is the contribution single sources. For instance when processing an electroencephalograms, what we observe is often modeled as a sum of different contributions from different sources located in the brain. Finding this source can allow for example to eliminate the important influence of blinking Eyes of the subject.

In audio, separation of sources is often introduced by evoking the cocktail party effect [11]. During a party, many simultaneous conversations occur, despite this we are able to focus our attention on a particular one. Doing so, we are able to *tune into* a single source and *tune out* all the others, still having access to the sound of the whole environment. In the same way, I can concentrate on one of the instruments playing in a song, thus mentally isolating it from others. In such case it usually refer as Sound Source Separation (SSS).

Translated into computer words, this ability would means the suppression of any track

of an audio recording. It is therefore natural that a large community of audio signal processing researchers have addressed the problem in several application such as:

- improving hearing aids performances;
- allowing instrument-wise equalization for music post-production
- proved a remixing or sample extraction tool for djs and producers
- stereo-to-surround (or 3D) up-mixing
- karaoke systems
- preprocessing in MIR tasks, e.g. automatic music transcription, melody extraction, etc.

In general it is not necessary that these elementary functions correspond to signals emitted by real independent entities, as is the case of SSS where the various musical instruments are playing in a recording. The objective of SS can simply be to explain at best a complex observation as the sum of several simpler *latent variables*. This early approach has given rise to precursor work in statistics under the name of *generalized additive models* [12], [13].

Contributions

The main contributions stemming from this work can be summarized in the following:

- to show how a rigorous probabilistic Gaussian framework may be used to yield provably optimal algorithms to learn all the parameters required for good interference reduction;
- to provide an open-source Python implementation of the derived algorithms;
- to compare the proposed approach with state-of-the-art approach in a perceptual study led on real legacy multitrack recordings from the Montreux Jazz Festival[1], one of the most important musical events in Europe for more than 50 years.

This work will hopefully be of relevance to the research community, contribute to the body of knowledge and the state-of-the-art in the field, and eventually to improvements in the application of source separation techniques in MIR, and vice versa. Additionally, it can be useful for the industry, in terms of the previously introduced applications

[1] `www.montreuxjazzfestival.com`

State of the Art

Gaussian Process for Audio

The machine learning developments of the last decade have made GP a serious competitor for real supervised learning applications. They have been successfully employed to solve nonlinear estimation problems in machine learning, but that are rarely used in signal processing [14]. Only in the last 5 years, some studies have been conducted on this topic, especially on audio signal processing [5]. In this works, GP regression problem is presented as a natural nonlinear Bayesian extension to the linear *minimum mean square error* (MMSE) and *Wiener filtering*. It results an effective and elegant framework for performing SS. Here the GP of the mixture signal is modeled as a linear combination of independent convolved versions of latent sources, defined as GPs. See chapter **??** for a review.

Only recently GPs have gained momentum in the audio signal processing community. In fact they have been used also for numerous problems such as estimating spectral envelope and fundamental frequency of a speech signal, for music genre classification and emotion estimation, pitch estimation and inferring missing segments in a polyphonic audio recording (see [15] and references therein).

Interference Reduction

The central question surrounding this topic is: is it possible to remove interferences to get clean, isolated source signals? In the last 10 years, several studies has been conducted on this problem. In each of these studies, the main assumption is that for each source there is at least one primary microphone, that is the number of sources and their corresponding microphones are known. This assumption is know as the *close-microphone* (or close-mic) technique and it is reasonable in almost all cases. In fact it is common for sound engineer to place of the microphone in close proximity to the sound source it is intended to capture (see Figure **??**). In these works, different approaches have been studied:

Time domain methods have been investigated by the authors of [16] and of [17]. These approaches attempt to perform a time domain *blind* source separation (BSS), which is the general case of source separation where there is not prior knowledge about the sources (number nor type). They overtake the complex BSS formulation and its high computational load, using echo cancellation strategies and adaptive *infinite impulse response* (IIR) filtering. The core work is the estimation of the delays (see as delay line, that is a propagation filters) between sources and microphones which has been made exploiting inter-microphones phase dependencies.

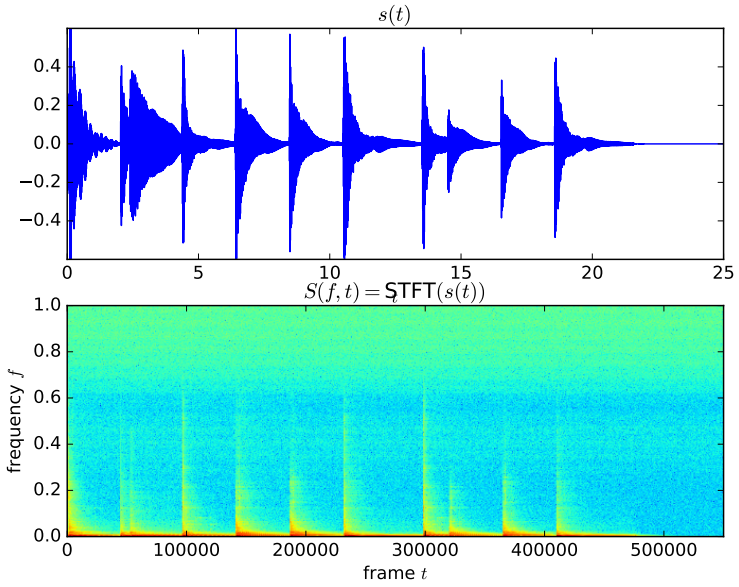

Figure 8: Short Time Fourier Transform representation of a piano sound signal $s(t)$

Time-Frequency (TF) domain methods use the *Short Time Fourier Transform* (STFT) representation for the recordings as in Figure 8. The main approach, firstly proposed by Kokkinis in [18], [19], was to perform interference reduction thought Wiener filter. This work was a breakthrough on this topic and the author made it clear that neglecting these dependencies and rather concentrating on energy redundancies over channels brings robustness and computational effectiveness. Thanks to the close-mic approximations, in[18] the author simply assume that the STFT of each recordings is already a good estimate for its dedicated sound sources. In this way after identifying the *Power Spectral Densities* (PSD) of the sources, a simple Wiener filter is applied in each channel to recover the desired signals [20] at small computational cost. In his more recent works [19], [21], Kokkinis introduce further temporal constraints on the sources so as to better identify them form the mixture recordings. In particular, in [21] the author have focused on real-time alternatives for ad hoc situations, i.e. applied to drums sound signals, leading to the development of some dedicated commercial products[2].

Non-Negative Matrix Factorization (NMF) approach follow the famous algorithm commonly used in source separation[22] (see Chapter **??** for a more details). It is used as a global model for the source spectrograms [23] exploiting the knowledge about the close-mics proprieties and number and type of instruments making up the observed mixture. To this end, a set of instrument models are learned from a training database and incorporated into a multichannel extension of the NMF algorithm (see Figure 9). Here the result mixture is assumed to be the product of a constant (or instantaneous) *mixing matrix* and the signal components. In this work it has been shown that the instantaneous mixing assumption

[2]See, e.g `http://accusonus.com/products/drumatom`.

provide similar performance to other state-of-the-art approaches.

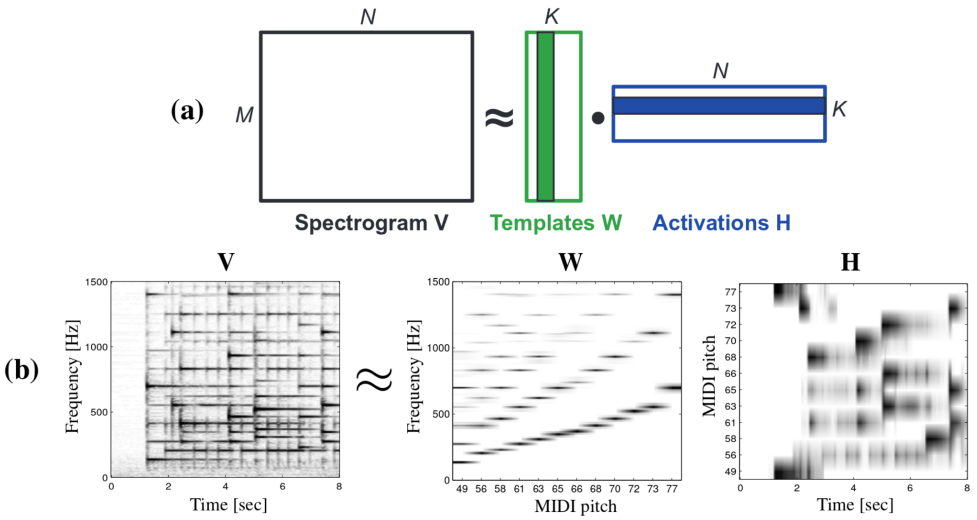

Figure 9: (a) A given non-negative matrix V representing a spectrogram is approximated as a product of two non-negative matrices W and H, called respectively *templates* and *activation* matrix. This two matrices typically have a much smaller rank the V. (b) Example factorization of a magnitude spectrogram for a piano signal. Courtesy of Ewert, Sebastian and Müller, Maynard

Despite the better performance of the TF approach with respect to the ones working in time, the main challenge to face is the estimation of the PSD of the sources which is determinant to achieve good performance. In [19] the author introduce the idea of the temporal constraint: a weighting coefficients model to quantify how much each voice is present in each track. Extending this idea as well as the mixing matrix of [23], Prätzlich in [4] introduce the *interference matrix*. The concept is illustrated in figure **??**. while [4] then concentrates on a grounded way to learn this interference matrix automatically, the spectral models, i.e. PSDs, are updated in a somewhat ad hoc fashion, leading to clear sub-optimality of the estimation algorithm.

This thesis is a clear extension to [4] works. In particular an estimation procedures for all parameters (interference matrix and PSDs) of the model have been derived, leading to provably optimal methods for leakage reduction.

Road Map

This work has a natural split into two part: the introduction which cover the motivation and the state of the art; an applicative part dealing with the IR problem applied to the music signal processing domain follows. The latter one is presented as two paper: one published in the proceeding of the *Audio Engineering Society (AES) Semantic Audio 2017* conference in Erlangen (DE) on July 2017, the other (still in draft mode) will be submitted to *IEEE International Conference on Acoustic Signal, Speech and Signal Processing 2018*

in Seul (KR).

The first of the two paper cover Chapter 1. Here the problem of interference reduction in live recording is formulated. Here the state-of-the-art framework based on GP and generalized multichannel Wiener Filters s is extended by fixing some heuristic parts of their algorithms. Finally in a perceptual evaluation on real-world multitrack live recordings shows that the resulting principled techniques yield improved quality.

The second paper follows in Chapter 2, where we focus on how to reduce the computational load of the proposed algorithms. The original approach require huge computational resources both in time and space, which it makes infeasible the processing of ordinary multi-channel (30 tracks) audio longer than 20 seconds. We show how random projection can address this issue providing full length track recordings processing in acceptable user time.

Part I

Interference Reduction

Chapter 1

Interference Reduction for Multitrack Live Recordings

The thesis is the result of an 6-months internship in the *Multispeech* team in INRIA GRAND EST[1]. It was supervised by Antoine Liutkus and founded by Erasmus+ grant and by Creative Dynamics of Improvised Interaction (CYCI2) project fundings[2].

The scope of this internship was firstly to develop an open-source Python3 portable implementation of the `MIRA` algorithm [24]. Secondly my attentions would have moved on to an online implementation with a graphical user interface. In fact this tool would have used by researcher of CYCI2 project as a pre-processing for MIR tasks, mainly automatic computer music improvisation. However it turns out that a Python3 command-line implementation an implementation would have been sufficient. Thus, in agreement with my supervisor, we decided to focus on the research side of the project, investigating optimal and fast strategies to perform IR extending `MIRA`.

In the following pages, my work on interference reduction for live recording will be reported. It has been redacted during February 2017, after 3 months of work and it has been presented in at the Conference on Semantic Audio 2017, the 22^{nd} of June in Erlangen, Germany. This paper is the result of some consideration on and improvements to the `MIRA` method made during the first 3 months of the work. Please, help your self and freshly enjoy it.

[1] Villers-les-Nancy, F-54600, France

[2] `http://repmus.ircam.fr/dyci2/home`

Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Gaussian framework for interference reduction in live recordings

Diego Di Carlo[1], Ken Déguernel[1,2], and Antoine Liutkus[1]

[1]*Inria, Multispeech team, Villers-les-Nancy, F-54600, France*
[2]*IRCAM STMS Lab (CNRS, UPMC, Sorbonne Universités), Paris, France*

Correspondence should be addressed to Diego Di Carlo (diego.dicarlo89@gmail.com)

ABSTRACT

In live multitrack recordings, each voice is usually captured by dedicated close microphones. Unfortunately, it is also captured in practice by other microphones intended for other sources, leading to so-called "interferences". Reducing this interference is desirable because it opens new perspectives for the engineering of live recordings. Hence, it has been the topic of recent research in audio processing. In this paper, we show how a Gaussian probabilistic framework may be set up for obtaining good isolation of the target sources. Doing so, we extend several state-of-the art methods by fixing some heuristic parts of their algorithms. As we show in a perceptual evaluation on real-world multitrack live recordings, the resulting principled techniques yield improved quality.

1 Introduction

In typical studio conditions, instrumental voices are often recorded simultaneously because this promotes spontaneity and musical interaction between the musicians, but also because it optimizes studio time usage. For live musical performances, each musician from a band gets its dedicated microphones, so that the different voices may be optimized independently and on-demand by sound engineers.

In all these situations, having clean isolated recordings for all instrumental voices is desirable because it allows much flexibility for further processing, remixing and exploitation. However, it is inevitable that *interferences* will occur, so that some voices are captured by microphones intended to other voices. This classical fact is also called *leakage* or *bleeding* by sound engineers, who have a strong expertise in designing specific acoustic setups to minimize them. However, unless the musicians do not play in the same room, which is detrimental to musical spontaneity, interferences are bound to occur in practice.

In the last 10 years, research has been conducted on the topic of interference reduction [1, 2, 3, 4]. Its goal is to propose signal processing algorithms that may be used by sound engineers to reduce the amount of leakage in live multitrack recordings. Most of the time, these methods are applicable a posteriori and require important computing resources. However, some studies have focused on real-time alternatives for ad hoc situations [5] leading to the development of some dedicated commercial products[1]. We shortly review this line of research now.

[1]See, e.g http://accusonus.com/products/drumatom.

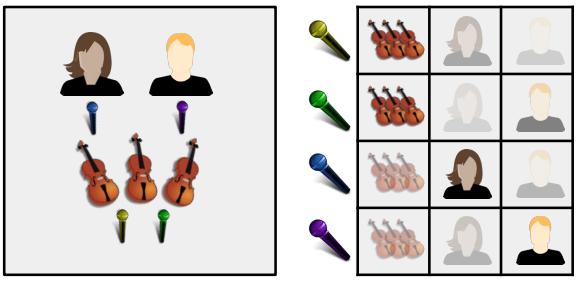

Fig. 1: Illustration of typical interferences found in multitrack live recordings. In the setup considered here: violin section, male singer, female singer, each voice gets its own dedicated microphones. However, the resulting signals all get leakage from all voices. The amount of interference is quantified in our model by the interference matrix, as proposed in [7] (courtesy of R. Bittner).

Although early research in interference removal has been focused in exploiting inter-microphone phase dependencies [1], the breakthrough brought in by [2, 4] made it clear that neglecting these dependencies and rather concentrating on energy redundancies over channels brings robustness and computational effectiveness. After identifying the Power Spectral Densities (PSD) of the sources, a simple Wiener filter is applied in each channel to recover the desired signals [6]. Therefore, the main challenge these methods face is the estimation of the PSD of the sources to achieve good performance [4]. Their main working hypothesis is that the close-microphones for a given voice already present good isolation properties and may be used as the PSD to use for Wiener filtering. This idea can be further improved by enforcing some prior information about what each voice should sound like in terms of spectral characteristics. This led to products specialized in the reduction of interferences for drum signals [5], as well as to recent developments able to concentrate on orchestral leakage reduction [7, 8].

While early methods based on Wiener filter are straightforward to implement [2], they suffer from one important drawback: the voice models are initialized using their close-mic recordings and are assumed to have the same energy within all tracks. Extending the weighting coefficients model [4] as a way to quantify how much each voice is present in each track, Prätzlich in [7] introduce the *interference matrix*. The concept is illustrated in figure 1. While [7] then concentrates on a grounded way to learn this interference matrix automatically, the spectral models are updated in a somewhat ad hoc fashion, leading to clear sub-optimality of the estimation algorithm.

In this study, we show how a rigorous probabilistic Gaussian framework [6, 9, 10] may be used to yield provably optimal algorithms to learn all the parameters required for good interference reduction. We present and detail four alternative algorithms to this end and provide an open-source Python implementation. The discussed methods are compared with state of the art in a perceptual study led on real legacy multitrack recordings from the Montreux Jazz Festival[2], one of the most important musical events in Europe for more than 50 years.

2 Model and Methods

2.1 Notation and probabilistic model

First, we detail our notations for referring to the signals. Let J be the number of voices and I be the number of microphones. For $i = 1 \dots I$, x_i is the signal recorded by the i^{th} microphone, called a *mixture*. In full generality and because of interferences, this i^{th} mixture captures sound from all the voices. Hence, for $j = 1 \dots J$, we define the *image* y_{ij} as the contribution of voice j in mixture i, so that we have $x_i = \sum_{j=1}^{J} y_{ij}$.
Let $X_i(f,t)$ be the STFT of mixture x_i and similarly for Y_{ij} with y_{ij}. They are all complex matrices of dimension $F \times T$, where F is the number of frequency bands and T the number of frames. We have:

$$X_i(f,t) = \sum_{j=1}^{J} Y_{ij}(f,t). \tag{1}$$

An entry (f,t) of any such matrix is referred to as a Time-Frequency (TF) bin. Now, let finally the power *spectrogram* of x_i be the $F \times T$ matrix V_i with nonnegative entries defined as:

$$V_i(f,t) \triangleq |X_i(f,t)|^2. \tag{2}$$

where $\triangleq$ denotes a definition. The goal of interference reduction is to compute an estimate $\hat{Y}_{ij}$ of the images Y_{ij}, for all i and j.

[2] www.montreuxjazzfestival.com

Second, we now briefly present our probabilistic model. To begin with, we assume that the signals originating from different voices $j = 1 \dots J$ are independent. Then, for each voice j, we assume that its contributions Y_{ij} in the different mixtures i are independent. This means we do not take the phase dependencies between the different channels into account. That arguable assumption proves important in practice for both robustness to real-world scenarios and computational complexity. Finally, for a given Y_{ij}, we model it through the Local Gaussian Model (LGM, [11, 9]), a popular model accounting for the local stationarity of audio. All the entries of Y_{ij} are taken independent and distributed with respect to a complex isotropic Gaussian distribution:

$$Y_{ij}(f,t) \sim \mathcal{N}_c(0,\, P_{ij}(f,t)), \tag{3}$$

where $P_{ij}(f,t) \geq 0$ is the *Power Spectral Density* (PSD) of y_{ij} and stands for its time-frequency energy.

Third, we detail the core idea we use for interference reduction, presented in [7]. Although phase dependencies between channels are neglected, the PSDs P_{ij} of a voice image in all channels are assumed to be the same up to channel-dependent scaling factors $\lambda_{ij}(f)$:

$$P_{ij}(f,t) = \lambda_{ij}(f) P_j(f,t), \tag{4}$$

where $P_j(f,t) \geq 0$ is called the latent PSD of voice j and is independent of the channel i. The scalar $\lambda_{ij}(f) \geq 0$ specifies the amount of interference of voice j into microphone i at frequency band f. They are gathered into $I \times J$ matrices $\Lambda(f)$ called *interference matrices*.

As a consequence of our assumptions (1) and (4), the observations $X_i(f,t)$ also follow the LGM as in (3) but with PSDs written $P_i(f,t)$. We have:

$$X_i(f,t) \sim \mathcal{N}_c(0, P_i(f,t)), \text{with } P_i(f,t) = \sum_{j=1}^{J} P_{ij}(f,t). \tag{5}$$

The free parameters of our model are written

$$\Theta = \left\{ \Lambda(f), \left\{ P_j(f,t) \right\}_j \right\}. \tag{6}$$

Then, if the parameters are known, the model readily permits effective filtering to recover the voice images. Indeed, according to the Gaussian theory, it is easy to compute the posterior distribution of a voice image Y_{ij} given X_i and the parameters Θ [9]:

$$Y_{ij} \mid X_i, \Theta \sim \mathcal{N}_c\left(\frac{P_{ij}}{P_i} X_i,\, \left(1 - \frac{P_{ij}}{P_i}\right) P_{ij} \right), \tag{7}$$

where we drop the dependence in (f,t) of all quantities for readability. From a Bayesian perspective, this distribution encapsulates everything we know about Y_{ij} once the mixtures and the parameters are known.
Following (7), the maximum a posteriori (MAP) estimate of Y_{ij} is given by:

$$\hat{Y}_{ij} \triangleq \mathbb{E}[Y_{ij} \mid X_i, \Theta] = W_{ij} X_i \triangleq \frac{P_{ij}}{P_i} X_i. \tag{8}$$

In the Gaussian case, this estimate also happens to be the Minimum Mean Squared Error (MMSE) and the Best Linear Unbiased Estimate (BLUE). In any case, the coefficient $W_{ij}(f,t)$ is usually called the *Wiener gain*. The time-domain signals of the estimated images can be obtain from (8) via inverse STFT.

For a given voice j, we are usually not interested in estimating Y_{ij} for all recordings i, but rather only for some, that we call the *close-mics* for voice j, as in [7]. They are given by the channel selection function for voice j, $\varphi(j) \subseteq \{1, \dots, I\}$. It indicates which microphones were positioned to capture voice j and is assumed known.

2.2 Parameter estimation

As discussed in the previous section, if the parameters are known, excellent separation performance can be obtained using the simple Wiener filter (8). The challenge to be overcome is hence to estimate those parameters from the observation of the mixture signals X_i only.
In this section we describe two procedures to perform parameter estimation. They both take as input the STFTs X_i of the recorded signals and the channel selection function φ. Then, they return estimates $\hat{\Theta}$ for the parameters, to be used for separation. A summary can be found in the Algorithm 1 box.

2.2.1 Marginal Modeling

According to [9], a way to estimate our parameters is to maximize the likelihood of the observations, that is find the Θ such that $\mathbb{P}[X \mid \Theta]$ is maximum.
According to our probabilistic framework, all entries $\{X_i(f,t)\}_{i,f,t}$ of the STFTs of the observed microphone signals are independent and distributed according to (5). It follows that we can compute the negative

log-likelihood $\mathscr{L}(\Theta)$ of the parameters Θ as:

$$\begin{aligned}\mathscr{L}(\Theta) &= -\log \mathbb{P}\left[\{X_i(f,t)\}_{i,f,t} \mid \Theta\right] \\ &= -\sum_{f,t,i} \log \mathbb{P}\left[X_i(f,t) \mid \Theta\right].\end{aligned} \tag{9}$$

Maximum Likelihood Estimation (MLE) of the parameters Θ then simply amounts to minimize (9):

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \mathscr{L}(\Theta). \tag{10}$$

It can be shown equivalent to:

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \sum_{f,t,i} d_0\left(V_i(f,t) \| \sum_j \lambda_{ij}(f) P_j(f,t)\right) \tag{11}$$

where d_0 is the Itakura-Saito divergence [3], presented as "a measure of the goodness of fit between two spectra"[12].

Whereas [7] used the cost function (11) only for optimizing over Λ, we use it now for all Θ. This is done using the classical Non-negative Matrix Factorization (NMF) methodology, where both $\Lambda(f)$ and $\{P_j\}_j$ are updated alternatively, in a multiplicative fashion. As can be seen, the procedure can simply be understood as fitting the power spectrograms V_i of the recordings to their model P_i. This is done by exploiting the marginal distribution of the mixtures.

Using classical NMF derivations, we can show that optimizing (11) over both Λ and P_j amounts in alternating between the two following updates:

$$P_j(f,t) \leftarrow P_j(f,t) \cdot \frac{\sum_{i=1}^{I} P_i(f,t)^{-2} V_i(f,t) \lambda_{ij}(f)}{\sum_{i=1}^{I} P_i(f,t)^{-1} \lambda_{ij}(f)} \tag{12}$$

$$\lambda_{ij}(f) \leftarrow \lambda_{ij}(f) \cdot \frac{\sum_{t=1}^{T} P_i(f,t)^{-2} V_i(f,t) P_j(f,t)}{\sum_{t=1}^{T} P_i(f,t)^{-1} P_j(f,t)} \tag{13}$$

2.2.2 Expectation Maximization

The second strategy involves the *Expectation-Maximization* iterative algorithm (EM, [13]). Instead of fitting the model directly using the marginal distribution of the observations, the EM methodology introduces the images Y_{ij} as latent variables and each EM iteration alternates between separation and re-estimation of the parameters [11].

[3] a particular case of β-divergence, d_β, with $\beta = 0$

In the so-called *E-step*, exploiting the posterior distribution $\mathbb{P}[Y_{ij} \mid X_i, \Theta]$ of the images, we can compute the posterior total variance $Z_{ij}(f,t)$ as:

$$Z_{ij} \leftarrow \mathbb{E}\left[\left|Y_{ij}\right|^2 \mid X_i, \Theta\right] = W_{ij}^2 V_i + \left(1 - \frac{P_{ij}}{P_i}\right) P_{ij}. \tag{14}$$

In the *M-step*, the parameters are re-estimated so that the image PSDs P_{ij} fit the posterior total variances (14):

$$\Theta \leftarrow \arg\min_{\Theta} \sum_{f,t,i,j} d_0\left(Z_{ij}(f,t) \| P_{ij}(f,t)\right). \tag{15}$$

As in the section 2.2.1, we derive the corresponding updating rule for P_j and $\lambda_{ij}(f)$:

$$P_j(f,t) \leftarrow P_j(f,t) \cdot \frac{\sum_{i=1}^{I} P_{ij}(f,t)^{-2} Z_{ij}(f,t) \lambda_{ij}(f)}{\sum_{i=1}^{I} P_{ij}(f,t)^{-1} \lambda_{ij}(f)} \tag{16}$$

$$\lambda_{ij}(f) \leftarrow \lambda_{ij}(f) \cdot \frac{\sum_{t=1}^{T} P_{ij}(f,t)^{-2} Z_{ij}(f,t) P_j(f,t)}{\sum_{t=1}^{T} P_{ij}(f,t)^{-1} P_j(f,t)} \tag{17}$$

It should be emphasized that the computation of P_{ij} always involves the latest version available of the parameters P_j and Λ. It can be shown that iterating over this EM procedure is guaranteed to lead the parameters to a local optimum for the optimization problem (10) [13].

2.3 Enforcing W-disjoint orthogonality

In the previous section, we presented two alternative methods to estimate our parameters under a maximum likelihood criterion. In both cases, the parameters are refined iteratively so as to best match the observations. We highlight here that the overall optimization problem (10) is non-convex, so that both optimization methods we proposed are sensitive to initialization.

As already advocated in [7], initializing the voice PSD P_j using $\varphi(j)$ already provides a very good efficiency for the algorithm. The rationale of this procedure is that close-mics should already provide a good guess of what each voice should sound like, taking us close to the desired solution. Pioneering work in the field [2] can actually be understood as directly separating the mixtures with this initialization and $\lambda_{ij}(f) = 1$, through the Wiener filter (8).

In this study, we go further than just hoping our initialization will be close enough for the algorithms to obtain good results. On top of our datafit criterion embodied

Algorithm 1: Gaussian Interference Reduction

1. **Input:**
   - $X_i(f,t)$ for each channel x_i;
   - Channel selection function $\varphi(j)$ for each voice j;
   - Minimal interference ρ;
   - Number N_{iter} of iterations;
   - Number N'_{iter} of *inner* iterations (*only for EM*).
   - Sparsity coefficient γ;
2. **Initialization:**
   (a) For each f,i,j, $\lambda_{ij}(f) = \begin{cases} 1 & : i \in \varphi(j) \\ \rho & : \text{otherwise} \end{cases}$
   (b) $P_j(f,t) \leftarrow \frac{1}{|\varphi(j)|} \sum_{i \in \varphi(j)} \frac{1}{\lambda_{ij}(f)} V_i(f,t)$
3. **Parameter Fitting:**
   Marginal Modeling algorithm (*MM*):
   (a) Update all $P_j(f,t)$ with (12), including (21) and (22) to numerator and denominator, respectively
   (b) Update all $\lambda_{ij}(f)$ as in (13)
   Expectation-Maximization algorithm (*EM*):
   (a) Compute Z_{ij} as in (14)
   (b) Update all $P_j(f,t)$ with (16), including (21) and (22) to numerator and denominator, respectively
   (c) Update all $\lambda_{ij}(f)$ as in (17)
   (d) For another *inner* iteration, return to step 3b
4. For another iteration, return to step 3
5. **Separation and output:**
   $\forall j, \forall i \in \varphi(j)$: compute $\hat{Y}_{ij}(f,t)$ as in (8)

by the negative log-likelihood in (9), we propose to also enforce *W-disjoint orthogonality* of the different sources PSDs, as formalized in [14].
W-disjoint orthogonality means that the voices will mostly have energy in different TF bins. Equivalently, it says that for any TF bin, only a few voices should have a significant energy. This phenomenon is often observed in practice and has been exploited for the separation of audio. One contribution of this study is to notice that W-disjoint orthogonality can be understood in terms of sparsity of the vectors $P(f,t)$, defined as the concatenation of the voice PSDs:

$$P(f,t) \triangleq [P_1(f,t), \ldots, P_J(f,t)]. \tag{18}$$

We propose to estimate the parameters by using a new regularized criterion, as:

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \mathscr{L}(\Theta) + \gamma \sum_{f,t} \Psi(P(f,t)), \tag{19}$$

where $\gamma \geq 0$ indicates the strength of the regularization, while Ψ is a *regularizing function* or *sparsity criterion* that is small whenever its argument is sparse (see [15] for a review). In this study, we considered the Wiener Entropy as a sparsity regularization. For a vector p of length J, it is given by:

$$\Psi(P(f,t)) = \frac{\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{\frac{1}{J}\left(\sum_{j=1}^{J} P_j(f,t)\right)}. \tag{20}$$

Since Ψ is independent of Λ, the updates (13) and (17) for Λ are unchanged. Concerning the updates of P_j, as in [15] the formulas (12) and (16) are modified adding the quantities $\nabla^{-}_{\Psi,j}(f,t)$ to their numerator and $\nabla^{+}_{\Psi,j}(f,t)$ to their denominator, as defined by:

$$\nabla^{-}_{\Psi,j}(f,t) = \gamma \frac{J\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{\left(\sum_{j=1}^{J} P_j(f,t)\right)^{2}}. \tag{21}$$

$$\nabla^{+}_{\Psi,j}(f,t) = \gamma \frac{\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{P_j(f,t)\left(\sum_{j=1}^{J} P_j(f,t)\right)} \tag{22}$$

3 Evaluation

In order to evaluate the proposed algorithms, we conducted an online listening test. The algorithms were applied on a whole pop rock live recording session of 'Huey Lewis and the News' *Hip to Be Square* at the Montreux Jazz Festival 2000 (length: 4'40"). This recording features 23 microphones recording 20 voices. It has a sample-rate of 48 kHZ and a depth of 16 bits/sample. The multitrack recording was provided by the Montreux Jazz Digital Project and EPFL. From this full-length processed recording, a set of two 10 seconds excerpts was extracted for perceptual evaluation.
Because of the live setup, all the microphone signals contain interferences, so that the standard evaluation metrics for blind source separation [16] were not applicable, since they require a clean reference signal against which to compare the results. Instead, we performed a perceptual audio evaluation inspired by the ITU-BS.1534-2 protocol, a.k.a. MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA, [17]), with some modifications and simplifications based on [18].

MUSHRA is a standard methodology for subjective evaluation of audio with "intermediate impairments" (i.e. significant degradation noticeable in most listening environment), such as in source separation and in interference reduction.
However in our context MUSHRA protocols can not be strictly applied: the reference sound is not hidden and not able to be evaluated and there are not any anchors, that are very bad sounds. We therefore adapted it by following the guidelines found in [18].

3.1 Listeners, data and procedure

There were 28 participants (24 men and 4 women), including the authors, aged between 23 and 57 yr (mean=32.9 yr). Web listening evaluations must take hearing abilities and listening environments of the participants into account. Thus, some preliminary questions about gear and musical background were asked.
The participants were asked 9 questions on the two different 10 seconds excerpts. Each question corresponded to a couple comprising one particular voice instrument and one quality scale. Each question was formulated as a MURSHA-like trial: given a question, it was asked to rate different stimuli on a 100-based quality scale in comparison to a reference. There were 6 sounds to evaluate per question, corresponding to the different algorithms. The instrument selected were the voice of Huey Lewis, the bass guitar and the drums.
The presented scales are a modification of the ones presented in [18] to fit the interference reduction problem:

1. *Acoustic quality of the target sound*: how does the target sound.
   Here is the exact wording of its explanation: "only pay attention to the target sound and do not consider the background, such as other instruments. Provide bad ratings if the target sound is highly distorted, highly unnatural, badly equalized, or misses some parts."
2. *Suppression of background sounds*: how much the background sounds have been suppressed from the recording.
   "Only pay attention to the background (e.g. other instruments or the audience) and do not consider the target sounds. Provide good ratings if background is silent and bad ratings for loud artificial or loud original background sound."
3. *Acoustic quality of background sounds*: how does the background sound.
   "Only pay attention to the background sounds and do not consider the target one. Provide bad ratings if the backgrounds sounds (e.g. other instruments or the audience) are highly distorted, badly equalized, present loud bleeps, rumbles, pops that are not included in the mixture."

3.2 Considered algorithms

With this perceptual evaluation we want to compare the performance of the proposed 4 alternative algorithms and the KAMIR algorithm and its fast approximation presented in [7]. Methods in [2, 4] are not taken into account in this work because they have been already compared to KAMIR in [7]. So that the comparison considered methods are the followings:

K:	KAMIR algorithm
$\tilde{\mathbf{K}}$**:**	Approximation to KAMIR
EM:	Expectation Maximization
$\mathbf{EM}+\mathbf{S}$**:**	Expectation Maximization with sparsity
MM:	Marginal Modeling
$\mathbf{MM}+\mathbf{S}$**:**	Marginal Modeling with sparsity

For all the tests, we chose an FFT size of 4096 samples with 75% overlap, an initial floor interference parameter $\rho = 0.1$, $N_{\text{iter}} = 5$ iterations for the algorithm and $N'_{\text{iter}} = 5$ inner iterations for the EM variants. For the sparse variants, we picked a sparsity weight $\gamma = 1000$.

3.3 Results

In order to conduct a statistical analysis on the collected subjective data, the assessments for each participant are converted linearly to the range 0 to 100. Using some data-visualization tool, we could detect outliers: 3 incomplete and 1 totally-inconsistent evaluations have been legitimately removed. Moreover, dividing the participants according to a self-declared musical expertise significantly changed the results. For instance, background quality ratings are significantly different between non-experts and experts: p-value($\mathbf{EM}+\mathbf{S}$) = 0.0084, p-value($\mathbf{MM}+\mathbf{S}$) = 0.009). Moreover, the outliers mentioned before all belong to the non-experts group. We believe that non-expert participants introduced a big bias in the evaluation and were discarded for analysis, leaving 24 sets of results in total.

As a first analysis, we performed a non-parametric *Friedman test* to compare the results of each pair of algorithms along the three proposed scales. These results indicate that the **MM** and **EM** algorithms performs significantly better than $\tilde{\mathbf{K}}$ and **K** in terms of quality for

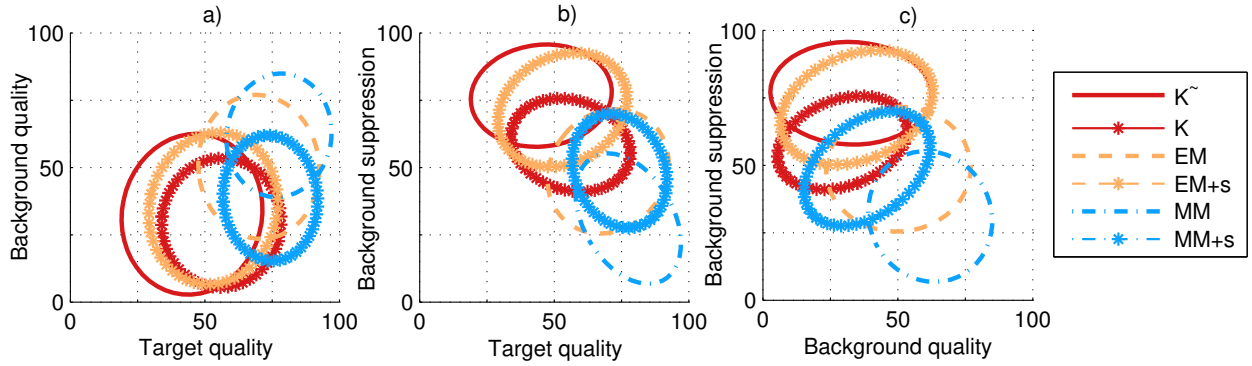

Fig. 2: Listening test result as confidence ellipse

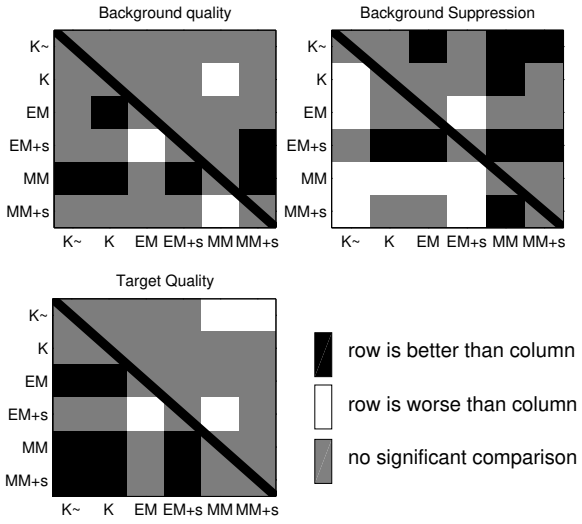

Fig. 3: Pair-wise test for each scale. Lower triangles are for all instruments, upper triangles for vocals only.

both background sounds and target sounds, but worse in terms of suppression. This indicates that these proposed modifications lead to better acoustic quality at the expense of less isolation. However, including a sparsity penalty term to both **EM** and **MM**, improves the suppression capability of the algorithms, suggesting that γ acts as a trade-off between isolation and target quality. Considering now the upper parts of the matrices, we see that the results for vocals only are slightly different, in any case in favor of the proposed modifications.

Figure 2 shows the confidence ellipse of the scores obtained by each algorithm on each pair of scales. It shows how the **EM** and **MM** perform slightly better than KAMIR in both of its fashions. As in Figure 3, we see the benefits of the sparsity penalty as improving background suppression at the cost of introducing some artifacts. An interesting observation is that $\mathbf{EM}+\mathbf{S}$ and $\mathbf{MM}+\mathbf{S}$ appear closer to $\mathbf{K}$ and $\tilde{\mathbf{K}}$ than **EM** and **MM**.

Regardless of the amount of noise that may affect the evaluation results, the EM method presented in this paper leads to slightly better results than state of the art. Close investigation reveals that its main difference with KAMIR lies in handling the uncertainty of the model through the posterior variance in (7). Then, the W-disjoint orthogonality penalty γ in (19) is seen as controlling the trade-off between isolation and distortion. The MM approach does not seem to perform significantly better than KAMIR algorithms, especially for the suppression of background. Still, adding a penalty γ brings it closer to EM, while having a significantly smaller computational complexity.

4 Conclusion

In this paper, we showed how a Gaussian probabilistic model for multitrack signals is useful in designing effective interference reduction algorithms. The core ideas of the model are twofold: neglecting the overly-complex phase dependencies between channels and rather focusing on energy relationships. In contrast to previous studies, we derived estimation procedures for all parameters of the model, leading to provably optimal methods for leakage reduction with this model. In a perceptual evaluation on real-world live recordings

from the Montreux Jazz Festival, we showed that the proposed method behave well when compared with state-of-the-art.

Acknowledgment

This work is made with the support of the French National Research Agency, in the framework of the project DYCI2 "Creative Dynamics of Improvised Interaction" (ANR-14-CE24-0002-01). Access to the Montreux Jazz Festival Database provided by EPFL in the context of this project.

References

[1] Uhle, C. and Reiss, J., "Determined Source Separation for Microphone Recordings using IIR Filters," in *Proceedings of the Audio Engineering Society Convention(AES)*, 2010.

[2] Kokkinis, E. K. and Mourjopoulos, J., "Unmixing Acoustic Sources in Real Reverberant Environments for Close-Microphone Applications," *Journal of the Audio Engineering Society*, 58(11), pp. 907–922, 2010.

[3] Clifford, A. and Reiss, J., "Microphone interference reduction in live sound," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2011.

[4] Kokkinis, E. K., Reiss, J. D., and Mourjopoulos, J., "A Wiener Filter Approach to Microphone Leakage Reduction in Close-Microphone Applications," *IEEE Transactions on Audio, Speech & Language Processing*, 20(3), pp. 767–779, 2012.

[5] Kokkinis, E., Tsilfidis, A., Kostis, T., and Karamitas, K., "A New DSP Tool for Drum Leakage Suppression," in A. E. S. (AES), editor, *Audio Engineering Society Convention*, 2013.

[6] Benaroya, L., Bimbot, F., and Gribonval, R., "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 191–199, 2006.

[7] Prätzlich, T., Bittner, R. M., Liutkus, A., and Müller, M., "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 584–588, IEEE, 2015.

[8] Prätzlich, T., Müller, M., Bohl, B. W., Veit, J., and Seminar, M., "Freischütz Digital: Demos of Audio-related Contributions," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain*, 2015.

[9] Liutkus, A., Badeau, R., and Richard, G., "Gaussian Processes for Underdetermined Source Separation," *IEEE Transactions on Signal Processing*, 59(7), pp. 3155 –3167, 2011, ISSN 1053-587X, doi:10.1109/TSP.2011.2119315.

[10] Souviraà-Labastie, N., Olivero, A., Vincent, E., and Bimbot, F., "Multi-channel audio source separation using multiple deformed references," *IEEE Transactions on Audio, Speech, and Language Processing*, 23(11), pp. 1775–1787, 2015.

[11] Duong, N., Vincent, E., and Gribonval, R., "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7), pp. 1830 –1840, 2010, ISSN 1558-7916, doi:10.1109/TASL.2010.2050716.

[12] Févotte, C. and Idier, J., "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, 23(9), pp. 2421–2456, 2011.

[13] Feder, M. and Weinstein, E., "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, 36(4), pp. 477–489, 1988.

[14] Jourjine, A., Rickard, S., and Yilmaz, O., "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 5, pp. 2985–2988, IEEE, 2000.

[15] Joder, C., Weninger, F., Virette, D., and Schuller, B., "A comparative study on sparsity penalties for NMF-based speech separation: Beyond LP-norms," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 858–862, IEEE, 2013.

[16] Vincent, E., Gribonval, R., and Févotte, C., "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, 14(4), pp. 1462–1469, 2006.

[17] Series, B., "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[18] Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M., "Fast and easy crowdsourced perceptual audio evaluation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 619–623, IEEE, 2016.

Chapter 2

Interference Reduction for Full-Length Live Recordings

As well as in the Chapter 1, in the following pages you will be delighted with a paper. Now the main focus is on how to apply the previously presented algorithm on full-length live recordings. Professional music recordings have really big size. This because high sampling frequency and bit depth are used: typical scenarios use respectively 48kHz and 32bits/sample.

On one hand this high resolution allows a better estimation of the parameters and verifies the condition for the Local Gaussian Model [25] (see Chapter **??**). On the other hand they adversely affect computational performances. For instance for 5 minutes multitrack recording with 40 mics and 30 voices, the time consumption was more then 30 minutes. Having the idea of developing an end-user (sound engineer or researcher) tool for IR, it was unthinkable to propose.

A solution was found in the random projection technique. This idea is presented in the following draft paper which will be submitted to IEEE Internation Conference on Acustic, Speech and Signal Processing (ICASSP) 2018 in Seul, South Korea. Please, help yourself and go with the (tensor)flow.

INTERFERENCE REDUCTION ON FULL-LENGTH LIVE RECORDINGS BY RANDOM PROJECTION AND GAUSSIAN MODELLING

Diego Di Carlo, Antoine Liutkus[*]

Inria, *Multispeech team*
Villers-les-Nancy, F-54600, France School

ABSTRACT

In this study, typical live full-length music recordings are considered. In this scenarios, some instrumental voices are captured by microphones intended to other voices, leading to so-called interferences. Reducing this phenomenon is desirable because it opens new possibilities for sound engineers and also it has been proven that it increase performances of music analysis and processing tools (e.g. pitch tracking). Extending state-of-the-art methods, we recently used an NMF-based algorithm that iteratively estimate each source contribution, i.e. the power spectral densities (PSDs), and the corresponding strength in each microphone signal, modeled in a interference matrix. Unfortunately our approach suffer of a huge computational load. As a contribution of this work, we show how random projection method suit the original problem formulation, yielding a good approximation of the parameters. This allow the algorithm to process full-length live multi-track recoding in a acceptable time. Moreover an online implementation is proposed. Experimental results demonstrate the efficiency of the approach.

Index Terms— interference reduction, microphone leakage, bleeding, cross-talk, source separation, random projection, compressive sensing

1. INTRODUCTION

When recordings a performance in typical studio conditions, instrumental voices are often recorded simultaneously because this promotes spontaneity and musical interaction between the musicians, but also because it optimizes studio time usage. In live musical performances, each musician gets its dedicated microphones, so that expert sound engineer may optimized each the different instruments independently and on-demand, see Figure 2 for an illustration.

In all these situations, having clean isolated recordings for all instrumental voices is desirable because it allows much flexibility for further processing, remixing and exploitation. However, complete isolation is impossible and *interferences* are bound to occur, i.e. some voices are captured by microphones intended to other voices. Sound engineers refer to this classical fact as microphone *leakage* or *bleeding*. They have a strong expertise in designing specific acoustic setups to isolate all the voices as much as possible. However, unless the musicians do not play in the same room, which is detrimental to musical spontaneity, interferences are inevitable in practice.

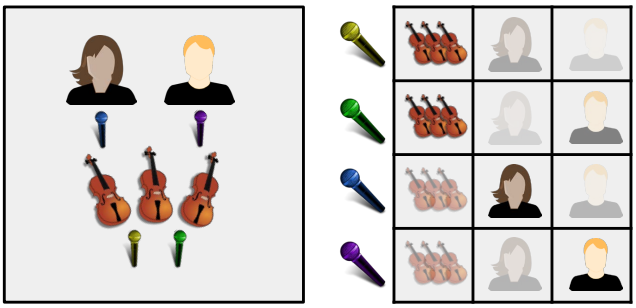

Fig. 1. Illustration of typical interferences found in multitrack live recordings. In the setup considered here: violin section, male singer, female singer, each voice gets its own dedicated microphones. However, the resulting signals all get leakage from all voices. The amount of interference is quantified in our model by the interference matrix, as proposed in [5] (courtesy of R. Bittner).

In the last 10 years, many studies have been conducted on the topic of interference reduction [1, 2, 3, 4, 5]. State-of-the-art approach consist in identifying the Power Spectral Densities (PSDs) of the sources and then applying a generalized Wiener filter to each channel to recover the desired signals [6, 2, 4]. In these works it has been shown that neglecting these dependencies and rather concentrating on energy redundancies over channels brings robustness and computational effectiveness. However the main challenge of these methods is the estimation of the PSD of each sound source. Thanks to the Gaussian framework [7] and to the Nonnegative-Matrix-Factorization method [8, 9], the author of [5, 10] proposed an elegant algorithm for performing interference reduction.

In our previous work [11] we fix some ad-hoc solution used used in [10]: we show how a rigorous probabilistic Gaussian framework [6, 7, 12] may be used to yield prov-

[*]Thanks to XYZ agency for funding.

ably optimal algorithms to learn all the parameters required for good interference reduction. Unfortunately this method is applicable a posteriori and require important computing resources. A solution has been found in random projection technique.

A statistical optimal way of dimensionality reduction is to project the data onto a lower-dimensional orthogonal subspace that captures as much of the variation of the data as possible [13]. Some methods, such as the well-known principal component analysis (PCA), leads to the best (in mean-square sense) way to do it; unfortunately it is really expensive to compute. Random projection has been found to be computationally efficient, yet sufficiently accurate method for this problem [14]. Here the core idea is to project high-dimensional data onto a lower-dimensional subspace using a random matrix, usually following a normal distribution.

In this work, we show how it is possible to achieve similar state-of-the-art performances using an approximation of one parameter, the interference matrix. Thus, the main contribution of this paper is the application of a random projection technique to speed up the algorithm. In particular we show that, thank to the Gaussian assumption, the original problem formalization still holds in a smaller-dimension random subspace. So the interference matrix can be estimated faster in this subspace and finally it can be used as a prior knowledge in the original problem. This permits our algorithm to be computationally efficient while achieving similar performance.

The remainder of the paper is organized as follows: In Section 2.1 we present our probabilistic model and the corresponding algorithms. In Section 3 the proposed modification to our previous work are detailed. Finally in Section 4 we evaluate the proposed approach with respect to the reference algorithm [11].

2. MODEL AND METHODS

2.1. Notation and probabilistic model

First, we introduce the notation for the involved signals. Let be J voices and I microphones. The *mixture* x_i is the signal recorded by the i^{th} microphone for $i \in \{1 \dots I\}$. In full generality and because of interferences, every voice $j \in \{1 \dots J\}$ is present in all mixtures x_i. Hence, defining the *image* y_{ij} as the contribution of voice j in mixture i, we can write $x_i = \sum_{j=1}^{J} y_{ij}$.
Let X_i be the STFT of mixture x_i. This yields I complex matrices of dimension $F \times T$, where F is the number of frequency bands and T the number of frames. Hence, for every Time-Frequency (TF) bin, we have:

$$X_i(f,t) = \sum_{j=1}^{J} Y_{ij}(f,t), \tag{1}$$

where Y_{ij} denotes the complex $F \times T$ STFT of y_{ij}.
Now, we define the power *spectrogram* of x_i as the $F \times T$ matrix V_i with nonnegative entries:

$$V_i(f,t) \triangleq |X_i(f,t)|^2. \tag{2}$$

where $\triangleq$ denotes a definition.
With our notation, the objective of interference reduction is to compute a good estimate $\hat{Y}_{ij}$ of the images Y_{ij}, for all i and j.

Second, we now briefly present the assumptions for our probabilistic model. First of all, we make the assumption that the images $\{Y_ij\}$ are independent for every i and j. The independence among j is the common assumption for considering voices produced by different physical instrument and acoustic process. Independence along i derive from neglecting phase dependences between different channel, which are assumed to be related thought their energy. This strong and arguable assumption proves important in practice for both robustness to real-world scenarios and computational complexity. Finally, we model our signal following the Local Gaussian Model (LGM, [15, 7]) for local stationary audio signal. It assumes that all the entries of Y_{ij} are taken independent and distributed with respect to a complex isotropic Gaussian distribution:

$$Y_{ij}(f,t) \sim \mathcal{N}_c(0,\ P_{ij}(f,t)), \tag{3}$$

where $P_{ij}(f,t) \geq 0$ is the *Power Spectral Density* (PSD) of y_{ij} and stands for its time-frequency energy.
Note that these hypothesis state the independence along all the four dimension: f,t,i,j.
Finally as presented in [5], we assume that the PSDs P_{ij} of each voice in all channels are the same up to channel-dependent scaling factors $\lambda_{ij}(f)$:

$$P_{ij}(f,t) = \lambda_{ij}(f)\, P_j(f,t), \tag{4}$$

where $P_j(f,t) \geq 0$ is the latent PSD of voice j and is independent of the channel i. The scalar $\lambda_{ij}(f) \geq 0$ gives the amount of interference of voice j into channel i at frequency band f. So we can define the $I \times J$ matrix $\Lambda(f) = (\lambda_{ij}(f))$ as *interference matrix*.

As a consequence of (1) and (4), the observations $X_i(f,t)$ also follow the LGM as in (3):

$$X_i(f,t) \sim \mathcal{N}_c\left(0,\ \sum_{j=1}^{J} \lambda_{ij}(f) P_j(f,t)\right). \tag{5}$$

The free parameters of our model are written

$$\Theta = \left\{\Lambda(f), \{P_j(f,t)\}_j\right\}. \tag{6}$$

Then, if these parameters are know, the model readily permits effective filtering to recover the voice images. Indeed, according to the Gaussian theory, the maximum a posterior

(MAP) estimate of the voice image Y_{ij} given X_i and the parameters Θ [7] is given by:

$$\hat{Y}_{ij} \triangleq \mathbb{E}\left[Y_{ij} \mid X_i, \Theta\right] = W_{ij} X_i \triangleq \frac{P_{ij}}{\sum_{j'=1}^{J} P_{ij}} X_i, \qquad (7)$$

where the coefficient $W_{ij}\,(f,t)$ is the so-called *Wiener gain*. In the Gaussian case, this estimate also happens to be the Minimum Mean Squared Error (MMSE). Finally, the time-domain signals y_{ij} of the estimated images can be obtain from (7) via inverse STFT.

2.2. Previously proposed method

In this section we briefly summarize our Marginal Modeling (MM) method which as been already proposed in [11]. It takes the STFTs X_i of the recorded signals and return estimates $\hat{\Theta}$ for the parameters, to be used for separation.

In this section we describe the main idea behind the proposed MM approach for Interference Reduction. According to [7], a way to estimate our parameters is to maximize the likelihood of the observations, that is find the Θ such that $\mathbb{P}\left[X \mid \Theta\right]$ is maximum.
The MM approach to learn the parameters is indeed to optimize it so as to enforce (5), by minimizing the discrepancies between the left and right-hand sides of (5):

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \sum_{f,t,i} d_0 \left(V_i(f,t) \| \sum_j \lambda_{ij}\,(f)\, P_j(f,t) \right) \qquad (8)$$

where d_0 is the Itakura-Saito divergence [8]. More details can be found in our previous work.

Finally, the entries λ_{ij} of the interference matrix and the PSDs $P_j(f,t)$ for each j can be obtain using the classical Non-negative Matrix Factorization (NMF) methodology. It can be shown that the corresponding update rule for the two parameters are:

$$P_j\,(f,t) \leftarrow P_j\,(f,t) \cdot \frac{\sum_{i=1}^{I} P_i\,(f,t)^{-2}\, V_i\,(f,t)\, \lambda_{ij}\,(f)}{\sum_{i=1}^{I} P_i\,(f,t)^{-1}\, \lambda_{ij}\,(f)} \qquad (9)$$

$$\lambda_{ij}(f) \leftarrow \lambda_{ij}(f) \cdot \frac{\sum_{t=1}^{T} P_i(f,t)^{-2} V_i(f,t) P_j(f,t)}{\sum_{t=1}^{T} P_i(f,t)^{-1} P_j(f,t)} \qquad (10)$$

We can see that the parameters are refined iteratively so as to best match the observations. However this optimization problem (8) is non-convex, so that both optimization methods we proposed are sensitive to initialization. As already discussed in [5], a rational and good initialization for both of the parameters is to exploit the information of the *close-microphone*. This approach lead the the `mimMIRA` algorithm presented in our previous work.

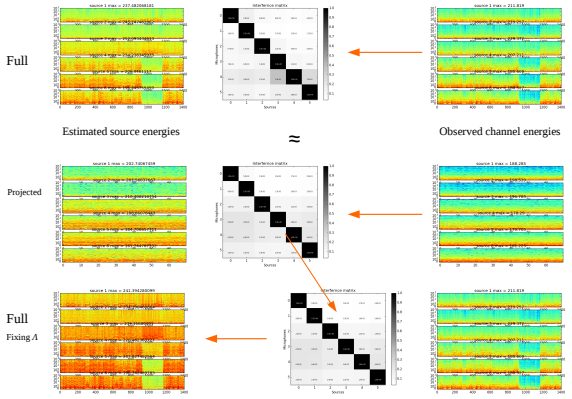

Fig. 2. Block diagram of the proposed approach: instead of estimating both $\Lambda(f)$ and the $\{P_j(f,t)\}_j$, $\Lambda(f)$ is estimated in a projected smaller subspace and kept fixed for estimating $\{P_j(f,t)\}_j$ from the original mix

3. RANDOM PROJECTION

The approached we proposed are able to process small-scale data. However end-user of this application, such as sound engineer, would be able to provide enough enhancement in a reasonable time. It was thus necessary to extend our work to large-scale recordings. More precisely, it can be seen that time and space complexity of the algorithm that use (9) and (10) is $\mathcal{O}(F \times T \times I \times J)$. Typical size of the inputs for 3 minutes long multi-track live recordings are: $F = 4096, T = 1000, I = 30, J = 25$. This typical values in an NFM-based algorithm yield an senseless application, both due to time and space occupancy. In these sense, a speed up is more required than and desirable.

The bottleneck is found in the updating rule 10. In fact, learning $\Lambda(f)$ require a summation over t: this force the algorithm to have access to the whole data every time, often is bigger than the central memory. If the $\Lambda(f)$ is known a priori, the algorithm can be significantly speeded up.

The idea is the compress useful information from the original mix and learn an good approximation of the $\Lambda(f)$, dimensionally smaller. Thus use this approximation to estimate the $\{P_j(f,t)\}_j$. This idea is illustrated in figure 2

To perform this compression, we use the random projection (RP) approach. For every mixture i-th, we define

$$M_i(f,r) \triangleq \sum_{t=1}^{T} X_i(f,t)\; Q_i(r,t) \qquad (11)$$

as the projection of $X_i(f,t)$ onto a lower r-dimensional subspace, where $Q_i(r,t) \sim \mathcal{N}(0,1)$ are the entries of a matrix with dimension $R \times T$ such that $R \ll T$. As a consequence of Gaussian assumption of $X_i(f,t)$, we can derive the distri-

bution of the entries M_i using typical statistical formula [1]:

$$
\begin{aligned}
M_i(f,r) &\sim \mathcal{N}\left(0, \sum_t Q_i(r,t)^2 \sum_j \lambda_{ij}(f) P_j(f,t)\right) \\
&\sim \mathcal{N}\left(0, \sum_j \lambda_{ij}(f) \underbrace{\sum_t P_j(f,t) Q_i(r,t)^2}_{\triangleq M_j(f,t)}\right) \\
&\sim \mathcal{N}\left(0, \sum_j \lambda_{ij}(f)\, M_j(f,t)\right)
\end{aligned}
\tag{12}
$$

Yet we can find again a the same formalization of (5). Moreover, similarly to (2) we can define the spectrogram of $M_i(f,t)$ as $|M_i(f,t)|^2$.

Hence, instead of processing the input data for their entire length, we can simply replace respectively $P_j(f,t)$ and V_i in (9) and in (10) by $M_j(f,t)$ and $|M_i(f,t)|^2$. Note that (12) states that the entries $\lambda_{ij}(f)$ are the same of (4), that is the interference matrix is not affected by the random projection. This is allowed by the Gaussian assumption, which may not hold in general.

In this way the algorithm can run faster on smaller input data and it returns an estimation of the interference matrix. Now it can be use as a prior knowledge in the original algorithm, yielding to an huge simplification: the real data can be now used and only the PSDs have to be estimated with (9). The implementation of this algorithm is named `fastMIRAND`.

4. EXPERIMENTAL EVALUATION

In order to compare the effectiveness of the proposed method, we compare the performance of our algorithm with respect to to its original version as in []. The two algorithm were applied on a whole pop rock multitrack live recording session of Huey Lewis and the News *Power of Love* at the Montreux Jazz Festival 2000 (length: 510"). This recording features 40 microphones recording 30 voices. It has a sample-rate of 48 kHZ and a depth of 16 bits/sample. The overall size of this multitrack recording is almost `1.2 GB` and it was provided by the Montreux Jazz Digital Project and EPFL.

Instead of conducting a perceptual evaluation, we decided to compare the interference matrix and the PSDs estimated by `fastMIRAND` with respect to the ones estimated by `mimMIRA` as function of the random subspace dimension R.

We first tested the effect of the reduced dimensionality using different values of R. We choose to follow an exponential law, that is $R = 2^k$ with $k = 0, 1, \dots 13]$. This allowed as understand the behavior for the extreme values of this parameter. At each R, all the parameters are computed anew. Figure 3 show the the *reconstruction error*, i.e. the cost function defined in (8), as function of the number of iteration. In this figure it is clearly seen that random projection yields a similar results: dimensionality reduction by random projection make the algorithm to converge at the same point of `mimMIRA` already after a few iteration. This occurs even with very low values of R: the normalized reconstruction error between the two approach for $R \geq 8$ is almost lower then 20%.

Unfortunately, a similar reconstruction error is just a sufficient and not necessary condition to assert the equality of the algorithms. In fact it is does not take into account the structure of the $P_j(f,t)$ nor $\Lambda(f)$, but only their product as in (8). To investigate that, we highlight the distance between the interference matrices: the target is the output of the `mimMIRA` versus the estimated from the `fastMIRAND` varying the dimension R and recording length T, i.e. the number of frames. A variation over T was chosen in order to understand if different sizes of input data can affect the choice of the R. Ad-hoc value for R and T have been chosen in order to highlight the performance of the `fastMIRAND`. Figure 4 show the results as phase transition map, typical visualization method for compressed sensing approach. It is easy to notice that a good approximation for $\Lambda(f)$ is obtained even for small values of the dimension R whatever the number of frames T, e.g. $R = 512$. Lower values lead to worts performances as expected, because to much information is missing once it is projected in a very small subspace.

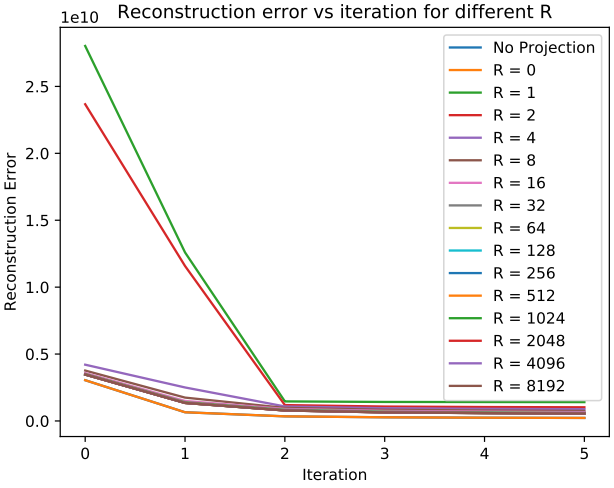

Fig. 3. Reconstruction error as function of the number of iteration of the algorithm for different size of R

The greatest point of interest is the computational complexity of the methods. Figure 5 shows the time in seconds measured with a 16-core desktop computer with 16GB of

[1] Given two random variable $\mathcal{X}$ and $\mathcal{Y}$ and a scalar α, if $\mathbb{E}(\mathcal{X}) = \mathbb{E}(\mathcal{Y}) = 0$, then $var(\mathcal{X}\mathcal{Y}) = var(\mathcal{X})var(\mathcal{Y})$, and $var(\alpha\mathcal{X}) = \alpha^2 var(\mathcal{X})$

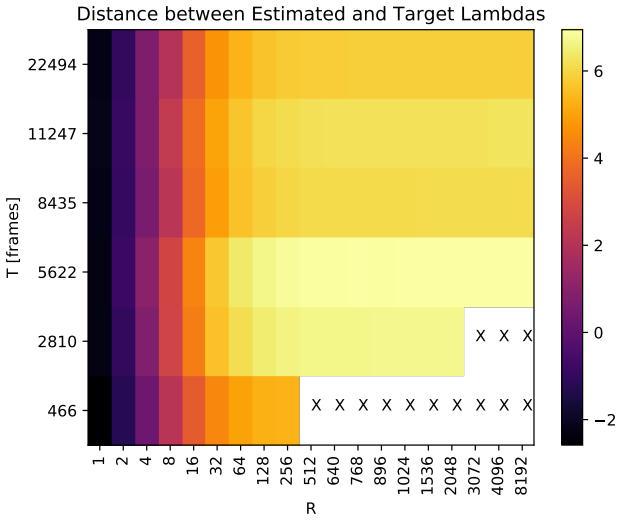

Fig. 4. Phase transition map. Distance between the matrix estimated in the projected subspace and the one estimated from the original mix is plotted as function of the subspace dimension R and the length of the input data T in frames

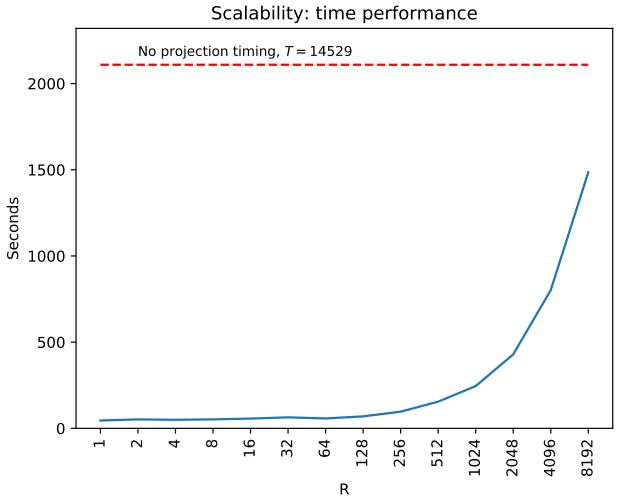

Fig. 5. Time consumption for the original (in dottet lines) and the proposed approach (in blue) as function of subspace dimension R. The simulation is run on a typical sound-engineer desktop computer

RAM, which is an ordinary setup for professional sound engineer. It is clear that while the `mimMIRA` takes more than 30 minutes, the proposed approach can yield a good approximation in only few minutes.

Listening test conducted by ourself reveals quite worst quality respect to the original methods, in particular in term of both isolation and artifacts. However for some application, such as orchestral (where signal are mostly harmonics) or drums (where signal are mostly percussive) recordings, performances are comparable to the previous proposed method.

5. CONCLUSION

In this paper, we have proposed a simple, yet effective way to reduce the computational load of an algorithm for interference reduction in live multitrack recordings. It is based on the random projection of the input data in a smaller dimensional subspace. This allow the algorithm to estimate one parameter, the interference matrix, which require to much computational load in the original input data space. Once the parameter is estimated, the original algorithm can be run frame-wise on the original data. Our evaluation indicates that in addition to being very simple to implement, this approach behave well in estimating the require parameter.

6. REFERENCES

[1] Christian Uhle and Josh Reiss, "Determined source separation for microphone recordings using IIR filters," in *Proceedings of the Audio Engineering Society Convention(AES)*, November 2010.

[2] Elias K. Kokkinis and John Mourjopoulos, "Unmixing acoustic sources in real reverberant environments for close-microphone applications," *Journal of the Audio Engineering Society*, vol. 58, no. 11, pp. 907–922, 2010.

[3] Alice Clifford and Joshua Reiss, "Microphone interference reduction in live sound," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, September 2011.

[4] Elias K. Kokkinis, Joshua D. Reiss, and John Mourjopoulos, "A Wiener filter approach to microphone leakage reduction in close-microphone applications," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 3, pp. 767–779, 2012.

[5] Thomas Prätzlich, Rachel M Bittner, Antoine Liutkus, and Meinard Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 584–588.

[6] Laurent Benaroya, Frédéric Bimbot, and Rémi Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[7] Antoine Liutkus, Roland Badeau, and Gäel Richard, "Gaussian processes for underdetermined source sepa-

ration," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155 –3167, July 2011.

[8] Cédric Févotte and Jérôme Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[9] Julio J Carabias-Orti, Maximo Cobos, Pedro Vera-Candeas, and Francisco J Rodriguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 184, 2013.

[10] Thomas Prätzlich, Meinard Müller, Benjamin W Bohl, Joachim Veit, and Musikwissenschaftliches Seminar, "Freisch utz digital: Demos of audio-related contributions," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain*, 2015.

[11] Diego Di Carlo, Ken Déguernel, and Antoine Liutkus, "Gaussian framework for interference reduction in live recordings," in *AES International Conference on Semantic Audio*, Erlangen, Germany, June 2017.

[12] Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, and Frédéric Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1775–1787, 2015.

[13] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[14] Sanjoy Dasgupta, "Experiments with random projection," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 143–151.

[15] Ngoc Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," *Latent Variable Analysis and Signal Separation*, pp. 73–80, 2010.

Chapter 3

Conclusion and Future Work

3.1 Conclusions

In this thesis, the problem of interference reduction in multitrack audio applications was addressed. In the literature It was formulated for the first time inside the signal processing framework as a blind source separation problem and as noise suppression, both in time and time-frequency domain. Extending a previous work, the Gaussian Process framework for source separation was formally applied: two consistent novel methods for interference reduction were derived, fixing some ad-hod heuristics of previous methods. It is applied to the case of multiple sources and microphones based only on the close-microphone assumption. As in its original formation, the latent PSDs of the sound sources and interference matrix are estimated, yet in a optimal way. A perceptual evaluation on real-world live recordings show that the proposed methods behaves well when compared with the state-of-the-art.

The main drawback of using of this approach is that it is typically slow, as much as general source separation approach. Since the execution time is an important issue, especially for end-user application, an effective way to reduce the computational load of proposed approach is introduced. This is based on projecting the original data onto a smaller dimensional subspace using random matrices. This provide a good approximation of the interference matrix, that can be achieve incredibly faster then the original method. Once this parameter is learn, the PSDs of the sources can be estimated with a frame-wise processing on the real data. Our experimental evaluation indicates that this approach yields to similar performance to the original algorithm.

3.2 Exodos

Some positive results have been obtained and an application able to yield good isolation have been developed. However, there is still room for improvements and thus much work could still be done to obtain even better results. Thus improvements can be done in several part:

- A more *rigorous perceptual* evaluation should be conducted on the presented method, as well on the state-of-the-art ones. Unfortunately crow-sourcing evaluation in this

subject introduce to much variability, thus a team of expert could provide more accurate results. Because of real live recording, BSS-eval framework can not be use to assert the quality of the methods, perceptual evaluation is hence necessary.

- As for source separation, the NMF approach is found quite limited and *neural network* are now studied as a new way to estimate model's parameters [26]. Moreover, an *informed* approach can be considered: using more prior knowledge about the original recordings, such as type of instruments and environment description, could lead to better results at the cost of more complicated model.

- *Random projection* approach open new possibilities that should be investigated. New parameters are now added to the model, such as the distribution of the random matrix. Hence many other projection approach can be used. This field of research is know as *compressed sensing* and *channel coding*, which are hot research topic right now.

Thanks to the presented results now interesting application can be developed. The possibilities for sound engineers are only bounded by fantasy and creativity. Moreover other fields of research can be interested in the discussed topic.
One of the most straightforward applications are in the biomedical, telecommunication and electronics field. In fact The theoretical approach behind them all relies in signal processing as well for sound and music computing. As explained in the introduction to this work, IR can be used as pre-processing tool for isolating signal of interest, no mater of the information they carry. In a similar way, random projection can be use as fast and good approximation for computational resource-demanding implementation of algorithms which are now limited by the size of the raw data.

Bibliography

[1] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.

[2] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[3] J. W. Tukey and N. Wiener, "The extrapolation, interpolation and smoothing of stationary time series with engineering applications.", *Journal of the American Statistical Association*, vol. 47, no. 258, p. 319, Jun. 1952, ISSN: 01621459. DOI: 10.2307/2280758. [Online]. Available: http://www.jstor.org/stable/2280758?origin=crossref.

[4] T. Prätzlich, R. M. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings", in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 584–588.

[5] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation", *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[6] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source", in *Acoustics speech and signal processing (icassp), 2010 ieee international conference on*, IEEE, 2010, pp. 425–428.

[7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech", *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[8] H. Levitt, "Noise reduction in hearing aids: A review", *Journal of rehabilitation research and development*, vol. 38, no. 1, p. 111, 2001.

[9] P. Stavroulakis, *Interference analysis and reduction for wireless systems*. Artech House, 2003.

[10] P. Comon and C. Jutten, *Handbook of blind source separation: Independent component analysis and applications*. Academic press, 2010.

[11] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions”, *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[12] R. A. Harshman, “Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis”, 1970.

[13] T. Hastie and R. Tibshirani, *Generalized additive models*. Wiley Online Library, 1990.

[14] F. Pérez-Cruz, S. Van Vaerenbergh, J. J. Murillo-Fuentes, M. Lázaro-Gredilla, and I. Santamaria, “Gaussian processes for nonlinear signal processing: An overview of recent advances”, *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 40–50, 2013.

[15] P. A. Alvarado and D. Stowell, “Gaussian processes for music audio modelling and content analysis”, in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, IEEE, 2016, pp. 1–6.

[16] C. Uhle and J. Reiss, “Determined source separation for microphone recordings using iir filters”, in *In 129th Convention of the Audio Engineering Society*, Citeseer, 2010.

[17] A. Clifford, J. D. Reiss, *et al.*, “Microphone interference reduction in live sound”, in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[18] E. K. Kokkinis and J. Mourjopoulos, “Unmixing acoustic sources in real reverberant environments for close-microphone applications”, *Journal of the Audio Engineering Society*, vol. 58, no. 11, pp. 907–922, 2010.

[19] E. K. Kokkinis, J. D. Reiss, and J. Mourjopoulos, “A wiener filter approach to microphone leakage reduction in close-microphone applications”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 767–779, 2012.

[20] L. Benaroya, F. Bimbot, and R. Gribonval, “Audio source separation with a single sensor”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[21] E. Kokkinis, A. Tsilfidis, T. Kostis, and K. Karamitas, “A new dsp tool for drum leakage suppression”, in *Audio Engineering Society Convention 135*, Audio Engineering Society, 2013.

[22] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[23] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodriguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings", *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 184, 2013.

[24] T. Prätzlich, R. M. Bittner, A. Liutkus, J. P. Bello, and M. Müller, "Interference reduction for multitrack music recordings", 2017.

[25] A. liutkus, "Processus gaussiens pour la séparation de sources et le codage informé", PhD thesis, 2012, p. 261.

[26] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, Jun. 16, 2016, ISSN: 2329-9290. DOI: 10.1109/TASLP.2016.2580946. [Online]. Available: http://ieeexplore.ieee.org/document/7492604%20https://hal.inria.fr/hal-01163369, published.