Principled methods for mixtures processing

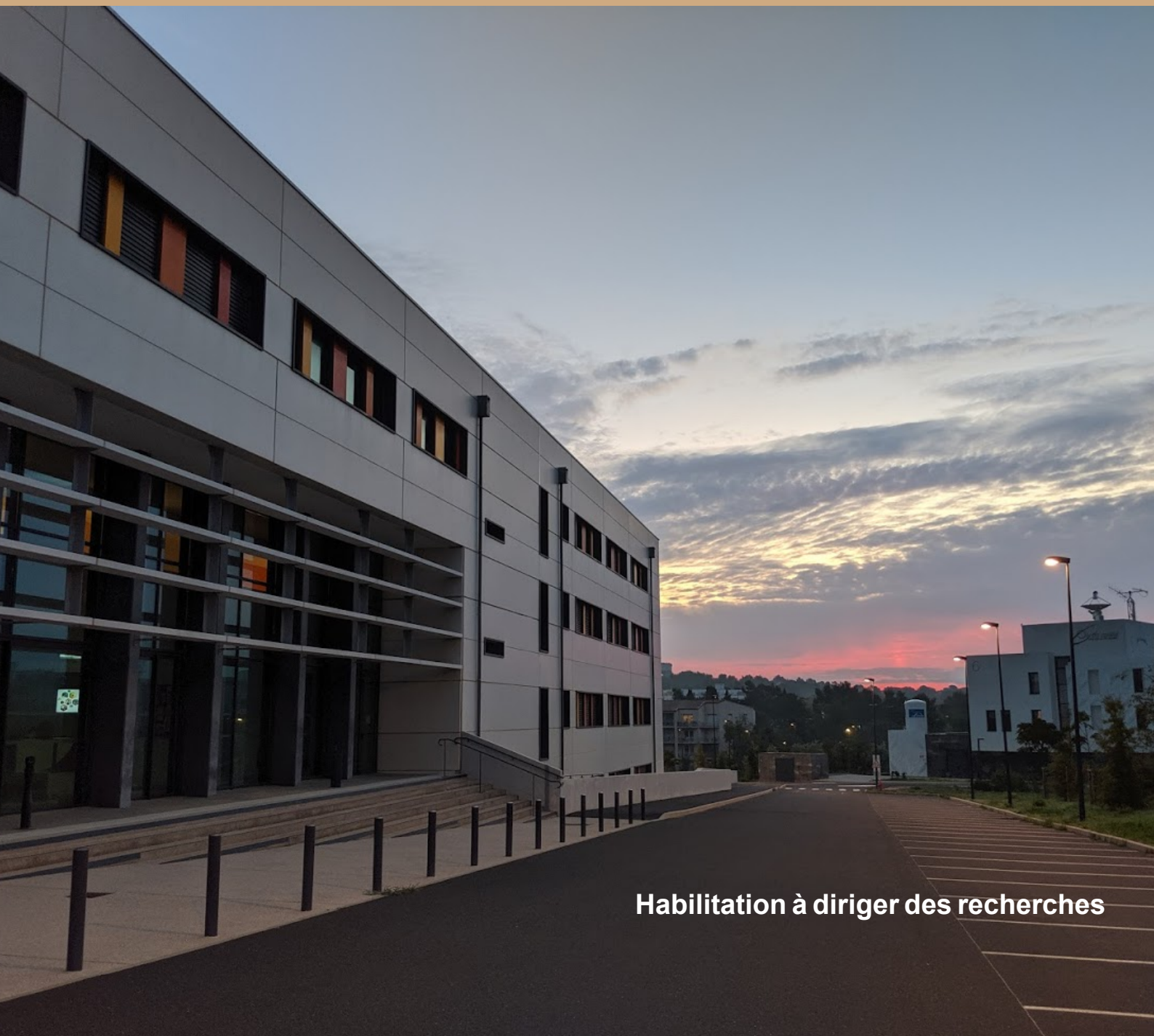
Antoine Liutkus

Université de Montpellier
LIRMM. Laboratoire d'informatique, de robotique et de microélectronique de Montpellier

Antoine Liutkus

Principled methods for mixtures processing

an overview of my past and future research



Habilitation à diriger des recherches

Antoine Liutkus

Principled methods for mixtures processing

an overview of my past and future research

Habilitation à diriger des recherches, February, 2022

LIRMM. Laboratoire d'informatique, de robotique et de microélectronique de Montpellier,
Université de Montpellier,
161 Rue Ada, 34095 Montpellier , France
www.lirmm.fr

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

Acknowledgements

First, I would like to thank Christian Jutten, Rémi Gribonval, Tuomas Virtanen, Cédric Févotte and Alexey Ozerov for accepting to be part of my jury. You all had a deep impact on my research curriculum and it is a real pleasure and honor to have you playing this role for me.

Second, I absolutely must acknowledge here the hundred researchers I collaborated with in the past 15 years. Although I cannot mention all of you individually, please know that doing research together was the real driving force that motivated my efforts throughout the years. I love spending time with you all, and the best moments of my career happened when great ideas popped in while discussing science with you, notably late at night in remote places. I always come back from conferences with a broken voice and a deep hangover, but also with a year's worth of ideas.

Third, my special thoughts go to Arie Nugraha and Mathieu Fontaine, for being such extraordinarily easy and talented Ph.D. students. Having you as close collaborators for these three years we spent together was a very rewarding experience and it is always a great pleasure and pride for me to read your new papers. I can't wait for you to have Ph.D. students of your own, so that I can finally be an academic grandpa.

Fourth, I would like to thank my family: brother, sisters, nephews, nieces and especially my parents Antoine and Martine. In this special occasion, I particularly must acknowledge the unfailing supervision from my mother. Her passion for science and pedagogy was inspiring all the way.

Finally, my deepest love goes to my wife Cheryl and to our three marvelous children Elias, Maeva and Daria. Thank you so much for making me feel so happy and fulfilled. My theory of *permanent golden age* has never been more accurate than since I met you. Ech hun iech vun Häerze gär.

Abstract

This document is my thesis for getting the *habilitation à diriger des recherches*, which is the french diploma that is required to fully supervise Ph.D. students. It summarizes the research I did in the last 15 years and also provides the short-term research directions and applications I want to investigate.

Regarding my past research, I first describe the work I did on probabilistic audio modeling, including the separation of Gaussian and α-stable stochastic processes. Then, I mention my work on deep learning applied to audio, which rapidly turned into a large effort for community service. Finally, I present my contributions in machine learning, with some works on hardware compressed sensing and probabilistic generative models.

My research programme involves a theoretical part that revolves around probabilistic machine learning, and an applied part that concerns the processing of time series arising in both audio and life sciences.

Contents

1 Past research: probabilistic methods for mixtures processing

Foreword. In this chapter, I will overview the research I did in the past 15 years. Since it corresponds to a lot of material (15 journal papers, 60 conferences and 4 book chapters), I decided to only focus on its guiding principles and general consistency, rather than to linger on its many technical details, which would require a massive document. Likewise, I decided not to cite articles I didn't write to keep this document light. Of course, this shouldn't give the impression that I am unaware of the tiny role I played in the big picture: the several papers in appendix comprise many hundreds of references, serving as appropriate pointers for the interested reader.

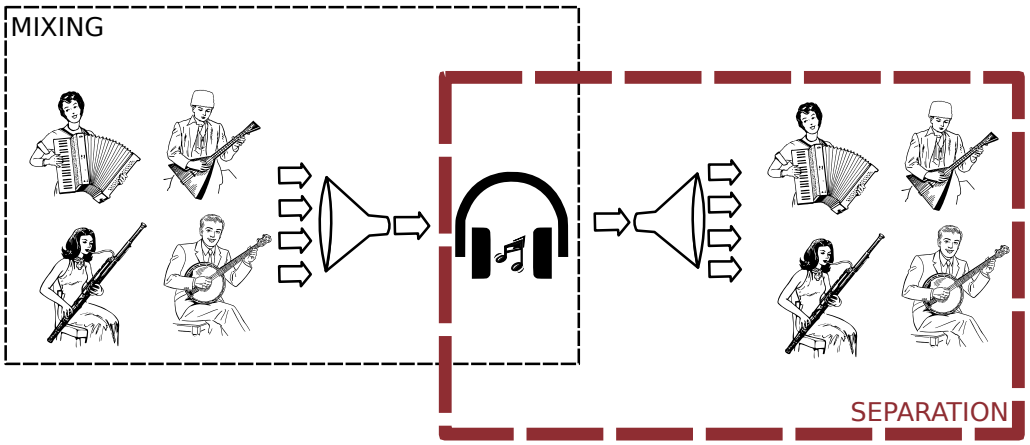


Figure 1.1: Music source separation aims at recovering the isolated instruments.

Source separation. The main application I was occupied with for all these years is music source separation. I met it early on in 2007 as a research engineer at Audionamix[1] and became passionate about it all the way until recently. In a nutshell, it is depicted in Figure 1.1 and consists in *demixing music*, i.e. recovering the isolated instrumental *sources* (bass, drums, guitar, vocals, etc.) from the mixture. As such, its goal is to invert the *mixing* procedure, which is either governed by the acoustic principles of sound propagation, or by non-physical and non-linear audio-engineering practice. As a very challenging inverse problem, it has been attracting an intense research effort for decades. I wrote several overview papers on the topic [J.Can19, B.Par18, J.Raf18], and refer the interested reader to page 104 for a detailed overview.

Outline. Additionally to my personal interest in music production, the main reason why I particularly enjoyed working on source separation is because it is a great playground for many different aspects of **theoretical research in probabilistic signal processing**. In section 1.1, I summarize the different models I proposed.
Equipped with probabilistic models for waveforms, the core challenge becomes to estimate their parameters, often boiling down to the "spectrograms" for the sources, which must be inferred from the mixture. On this matter, I considered both **kernel methods and deep neural networks**, that were still not widely used for source separation at that time. Managing a successful interaction between those powerful parametric models and music filtering occupied a significant part of my research activity from that point, that I summarize in section 1.2.
Although most of my research activity was applied to audio and source separation, my core contributions concern the design of new **probabilistic data manipulation techniques**. This lead me to have some contributions in other domains, notably matrix factorization, compressed sensing and lately generative models. I summarize these aspects in section 1.3.

[1] www.audionamix.com, formerly: MIST Technologies.

1.1 Probabilistic models for source separation

Throughout this document, bold uppercase denotes matrices, bold lowercase denotes vectors and normal typesetting denotes scalars. Italic typesetting denotes higher dimensional tensors. For indexing, I find it most readable to use named axis, so that the letter used indicates the dimension over which slicing is done. For instance, if $\mathcal{P}$ is an $M \times N \times R$ tensor, $\mathbf{P}_m$ is a $N \times R$ matrix, and $\mathbf{p}_{mr}$ is a N-dimensional vector. Finally, functions are written in italics like $\mathit{Y}_j, \mathit{k}_j$.

1.1.1 Background

Notations. The waveform from a music track is called a *mixture* in source separation, and I write it with a tilde: $\widetilde{\mathbf{X}}$. It is a real matrix of dimension $L \times I$, where L is the number of samples and I the number of channels. For stereo signals, we have $I = 2$. It is the sum of J source signals, corresponding to the individual instruments:

$$\widetilde{\mathbf{X}} = \sum_j \widetilde{\mathbf{Y}}_j. \tag{1.1}$$

A common preprocessing step is to take the Short-Term Fourier Transform (STFT) on both sides, yielding $F \times T \times I$ complex-valued tensors $\mathcal{X}$ and $\mathcal{Y}_j$, where F is the number of non-redundant frequency bins and T the number of time frames. Since the STFT is a linear operation, we get:

$$\mathcal{X} = \sum_j \mathcal{Y}_j. \tag{1.2}$$

A source separation method then produces sources *estimates* $\widehat{\mathcal{Y}}_j$ from the sole observation of the mixture $\mathcal{X}$, with the objective of having them as close as possible to the original signals $\mathcal{Y}_j$. Signals in the time domain are then obtained through inverse STFT.[2]

The Wiener filter. When I started to work on the topic, most research focused on the monophonic case $I = 1$, and it was already understood that choosing all $y_{jft} \in \mathbb{C}$ as independent and distributed with respect to (wrt.) an isotropic complex Gaussian distribution leads to a convenient treatment. We write this as:

$$y_{jft} \sim \mathcal{N}_c\left(0, v_{jft}\right), \tag{1.3}$$

where the real and nonnegative $F \times T$ matrix $\mathbf{V}_j$ is called the power spectral density (PSD) of source j. Its entries v_{jft} can be understood as the energy of source j at time t and frequency f. Basic Bayesian machinery would then show that the minimum mean squared error (MMSE) for $\mathbf{Y}_j$ in this case is:

$$\widehat{\mathbf{Y}}_j = \frac{\mathbf{V}_j}{\sum_{j'} \mathbf{V}_{j'}} \bullet \mathbf{X}, \tag{1.4}$$

where $\mathbf{A}/\mathbf{B}$ and $\mathbf{A} \bullet \mathbf{B}$ denote element-wise division and multiplication, respectively. This filter was derived by N. Wiener more than 60 years ago.[3] A nice feature of the method is that source estimates add up to the original mixture ($\sum \widehat{\mathbf{Y}}_j = \mathbf{X}$), which turns out to be an important property for professional users.

Challenges. This concise introduction is enough to describe what I perceived were the main challenges for source separation back in 2007 when I met the topic.

- **On the theoretical side**, the Gaussian assumption (1.3) is convenient, but I didn't really understand it back then. Of course, the Wiener filter (1.4) made sense to me as a dispatching of the mixture content according to each source's estimated energy. However, I didn't like picking a model in the Time-Frequency (TF) domain while what I wanted were waveforms. For this reason, I became interested in probabilistic models for waveforms that would underlie the basic assumption (1.3) and possibly extend it.

[2] Note that recent research in *end-to-end* separation replaces the STFT filterbank with learned transforms, but the core story remains the same.

[3] The original derivation did not rely on a model like (1.3), but rather on linearity assumption of the estimator combined with weak stationarity.

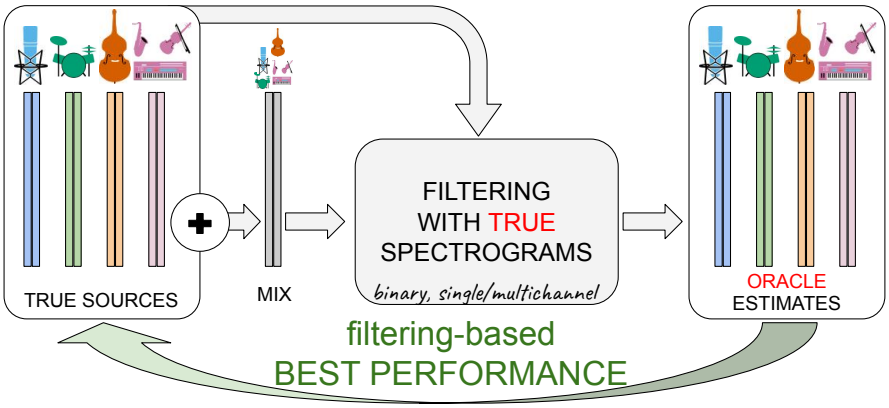


Figure 1.2: Oracle evaluation means using a separation procedure with the parameters computed from the true sources. The classical example is using the Wiener filter (1.4) with the spectrograms of the true sources. It allows to know what is the best quality a filtering method can lead to.

This lead me to propose Gaussian processes (GP) as an appropriate model for source separation, and later on α-harmonizable models as well as hybrid models. They are reviewed in Sections 1.1.2-1.1.7 below.

- **In terms of performance**, the Wiener filter provides very good separation quality whenever the (nonnegative) sources PSDs $\mathbf{V}_j$ are well estimated. Pushing this to the *oracle* limit illustrated in Figure 1.2, the difference with the true sources is actually almost inaudible, validating the approach as sufficient for all practical purposes.
  The natural research question was hence how to estimate these sources PSDs from the sole observation of the mixture. For this purpose, state of the art at that time was Nonnegative Matrix Factorization (NMF). It decomposes the mix spectrogram $|\mathbf{X}|^2$ into additive low-rank terms. Its baseline version in the single channel case ($I = 1$) reads:

$$|\mathbf{X}|^2 \approx \sum_k \mathbf{w}_k \mathbf{h}_k^\top, \tag{1.5}$$

  where each elementary $\mathbf{w}_k \mathbf{h}_k^\top$ supposedly stands for the modulated spectrum of a single sound element, to be assigned to some source. After estimation with appropriate methods, these estimated power spectra may be used for separation with Wiener filtering.
  I used this NMF model for several years, notably for audio coding. Then, I focused on alternatives that include non-parametric kernel models for PSDs and deep learning. I describe this line of research in the next Section 1.2.

1.1.2 Gaussian processes for source separation

While I was reading the excellent book *Gaussian processes for machine learning* by Rassmussen *et al.*, I came across the plots that I reproduce here in Figure 1.3. Without mentioning this explicitly, the authors are actually doing source separation on their data to decompose it as the sum of ”long trends” and ”seasonal” components! I was enthusiastic about this finding because it looked like the most general Gaussian framework to handle source separation and it was the starting point of several years of research, that I summarize now. This material was notably published in [J.Liu11, LBR11] that is included here on page 117, and stands as the foundation of a large part of the work I did afterwards.

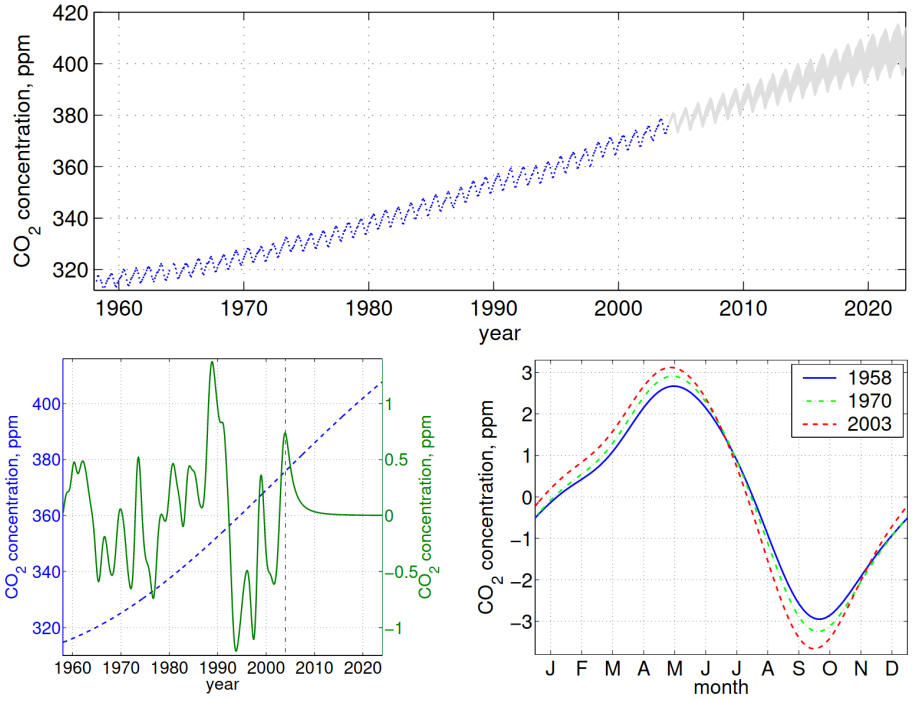


Figure 1.3: Excerpt from the book *Gaussian processes for machine learning* by Rassmussen et al. (p. 119-120). Carbon dioxyde concentration over the years (up) is decomposed as a sum of several independent contributions (below).

Gaussian processes Let $\mathbb{L}$ be an arbitrary set that we call the *index set* and whose entries stand for *coordinates* (think $\mathbb{L}$ as "locations"). Our aim is to model mixtures of random functions from $\mathbb{L}$ to $\mathbb{R}$, allowing separation of any kind of signals, defined with possibly high dimensional coordinates like Euclidean spaces ($\mathbb{L} = \mathbb{R}^D$), thus going much beyond the classical case of regularly sampled time series ($\mathbb{L} = \mathbb{N}$). Actually, literature on geostatistics in general and on *kriging* in particular was also a great inspiration to this research.

The mathematical machinery underlying the definition and manipulation of stochastic processes can be quite involved and I need some simplified notations to keep the math readable. Let $\mathbf{l} \in \mathbb{L}^L \triangleq [l_1, \ldots, l_L]$ denote a vector of L coordinates taken from $\mathbb{L}$, and let $Y : \mathbb{L} \to \mathbb{R}$ be a function defined on $\mathbb{L}$. We write $Y(\mathbf{l})$ for the L-dimensional vector of its values taken on $\mathbf{l}$. We also write $\mathbf{y}$ for this same object, when the locations $\mathbf{l}$ considered are clear from the context:

$$\mathbf{y} \triangleq Y(\mathbf{l}) \triangleq [Y(l_1), \ldots Y(l_L)]\,, \tag{1.6}$$

Likewise, if $k : \mathbb{L} \times \mathbb{L} \to \mathbb{R}$ is a real-valued function defined on $\mathbb{L} \times \mathbb{L}$, which is often called a *kernel*, we write $\mathbf{K} \triangleq k(\mathbf{l}, \mathbf{l})$ for the $L \times L$ matrix corresponding to its restriction to $\mathbf{l} \times \mathbf{l}$:

$$\mathbf{K} \triangleq k(\mathbf{l}, \mathbf{l}) \equiv [k(l_m, l_n)]_{m,n}\,. \tag{1.7}$$

Equipped with these few notations, we can now define a (centered) Gaussian process (GP), say Y. It is a collection $\{Y(l) \in \mathbb{R}\}_{l \in \mathbb{L}}$ of random variables (r.v.), such that its values taken on any finite and given set of locations is a Gaussian (centered) random vector. Mathematically, we write this as:

$$Y \sim \mathcal{GP}(k) \Leftrightarrow \forall L \in \mathbb{N}, \forall \boldsymbol{l} \in \mathbb{L}^L, \mathbf{y} \triangleq Y(\mathbf{l}) \sim \mathcal{N}\left(0, \mathbf{K} \triangleq k(\mathbf{l}, \mathbf{l})\right)\,. \tag{1.8}$$

This rather technical definition simply means that a GP can be handled like a Gaussian random vector for all practical purposes, i.e. whenever we are interested in its values at some finite set of coordinates, which is always the case in practice. The required covariance matrices are then "filled

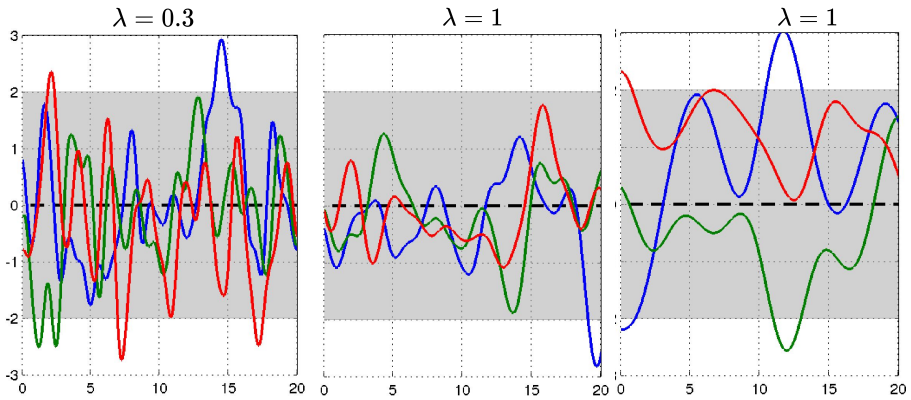


Figure 1.4: Some realizations of a centered Gaussian process with the squared exponential covariance function (1.9), for various values of λ. From [B.Liu12].

up" by applying the *covariance function* $k : \mathbb{L} \times \mathbb{L} \to \mathbb{R}$ on the desired points. This covariance function thus plays the central role in the definiton of a GP. It is defined as: $\forall (l, l'), k(l, l') = \mathbb{E}[Y(l)Y(l')]$. I understand it as a *similarity measure*: $k(l, l')$ is high whenever $Y(l)$ is expected to be similar to $Y(l')$. A typical example is the squared exponential covariance function:

$$k(l, l') = \exp(-\|l - l'\|^2 / \lambda^2), \tag{1.9}$$

that is appropriate for modeling *smooth* functions since two locations l and l' that are close will yield highly correlated values, and independent values otherwise. The *lengthscale* parameter λ controls the degree of smoothness.

Several basic facts must be known regarding GP. First, they can be easily sampled. This means that one can easily *randomly draw functions* that comply with the prescribed covariance structure. Examples can be found in Figure 1.4 for the SE covariance function (1.9). This can be understood as rolling a dice, except that we don't get a single number between 1 and 6 each time, but rather the values of a whole new function on the points of interest. Second, they allow a host of applications in both regression and classification. Their application to source separation is described at length in my Ph.D. manuscript [B.Liu12], as well as in several papers [J.Liu11, LBR11].

GP for source separation. Let us now consider J independent GP $Y_j \sim \mathcal{GP}(k_j)$. A first fact is that their sum, that we call the *mixture*, $X = \sum Y_j$ is also a GP:

$$Y_j \sim \mathcal{GP}(0, k_j) \text{ all independent } \Rightarrow X \triangleq \sum_j Y_j \sim \mathcal{GP}\left(0, \sum_j k_j\right) \tag{1.10}$$

Then, let us consider a particular set $\mathbf{l} \in \mathbb{L}^L$ of locations on which the mixture $\mathbf{x} \triangleq X(\mathbf{l})$ is observed. The *source separation* problem consists in inferring the value of the different sources $\mathbf{y}_j$ at these points.[4] If we know the true covariance functions k_j, it is readily shown that the posterior distribution of $\mathbf{y}_j \mid \mathbf{x}$ is:

$$\mathbf{y}_j \mid \mathbf{x} \sim \mathcal{N}\left(\mathbf{K}_j \mathbf{K}_x^{-1} \mathbf{x}, \mathbf{K}_j - \mathbf{K}_j \mathbf{K}_x^{-1} \mathbf{K}_j\right) \text{ with } \mathbf{K}_x = \sum_{j=1}^{J} \mathbf{K}_j. \tag{1.11}$$

Retrospectively, these considerations flow somewhat naturally from the actual definitions and don't seem to bring much into the picture compared to classical GP regression. However, GP were not

[4] The most general treatment allows for inference of the value of the sources at other locations than those on which the mix is observed, as well as the joint separation of all sources, but I decided not to provide all the details here that can be found elsewhere, notably in [J.Liu11, LBR11, LOBR12].

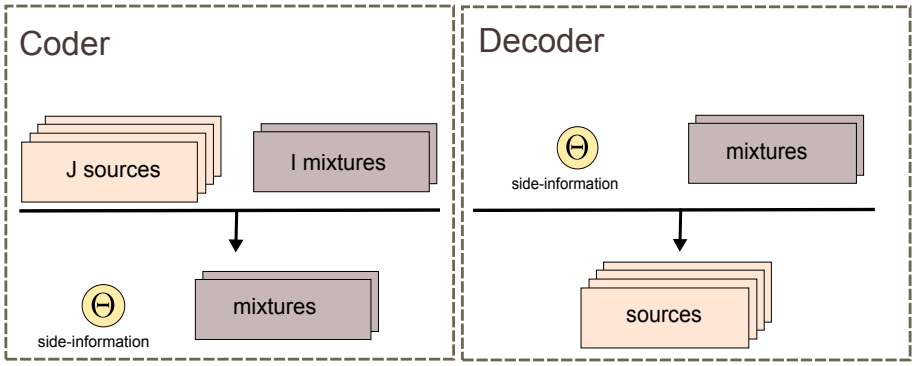


Figure 1.5: Informed source separation, a particular case of audio coding.

widely known back then in the source separation community, so that the connection was timely and I'm happy that it had a significant impact. It notably provided a way to justify separation methods with models picked *in the time domain* and to extend them to arbitrary index sets. I used this property several times afterwards myself [LBR11, LDA+12, PPL17, SLB+16].

The stationary case . A special case occurs whenever the sources are *stationary*, which means their covariance function $k_j(l, l')$ is a function of the *difference* $l - l'$ between its operands. When the coordinates $\mathbf{l} \in \mathbb{L}^L$ are located on a regular grid, all covariance matrices $\mathbf{K}_j$ are Toeplitz, and can approximately be diagonalized in the $L \times L$ Fourier basis. **This straightforwardly leads to the Wiener filter** (1.3) as the posterior mean for the sources. The Wiener-Khintchin theorem states that the covariance functions k_j and the PSDs are Fourier pairs. I spent some time proving these results again, notably when the index set is a general Euclidean space $\mathbb{R}^D$.

Finally, although this theory holds for stationary signals, it had to be adapted for **signals that are only *locally* stationary**, like typical audio signals. As a take-home message, I finally understood model (1.3) as resulting from two simplifying assumptions: i/ all the frames of the audio signals are independent. This is notoriously false, mostly due to the common overlap between them, but it stands as a convenient assumption all the same. ii/ each frame comprises a stationary Gaussian process and is sufficiently long to reasonably approximate its covariance matrix as circulant. It can be noted that some authors, notably K. Yoshii, took this understanding as a starting point for developing new models that would not make these simplifying assumptions.

1.1.3 Informed source separation

My Ph.D topic was *Informed Source Separation* (ISS). Its workflow is depicted in Figure 1.5. The J original separated sources are available at a first *encoding* stage, where a side information is computed. During *decoding*, it is used in conjunction with the mixture to reconstruct the sources. The applications of that setup arise in *active listening* applications, where the music producer would like to offer some features to the final audience, like muting and equalizing the different instruments, or adapting playback according to the available loudspeakers [MBB+12].

Parametric single-channel ISS. Assuming the true sources are monophonic, my first contribution on this topic was to compress the true source spectrograms jointly with nonnegative tensor factorization (NTF), as:

$$|y_{jft}|^2 \approx \sum_k w_{fk} h_{tk} q_{jk}, \tag{1.12}$$

using a principled method from the state of the art (Majoration-Maximization) due to C. Févotte. An example of such a model in action can be seen in Figure 1.6. Then the (small) parameters **W**, **H**, **Q** are sent to the decoder and used along with the mixture for Wiener filtering as in (1.4). This idea actually proved extremely effective in terms of bitrate, requiring only a few additional kilobits per seconds (kbps.) to recover the sources almost perfectly from the mixture [LBR10]. Later on, I also proposed to simply encode the log-spectrograms of the sources in the JPEG format, which actually proved very effective, both in terms of computational complexity and quality [J.Liu12].

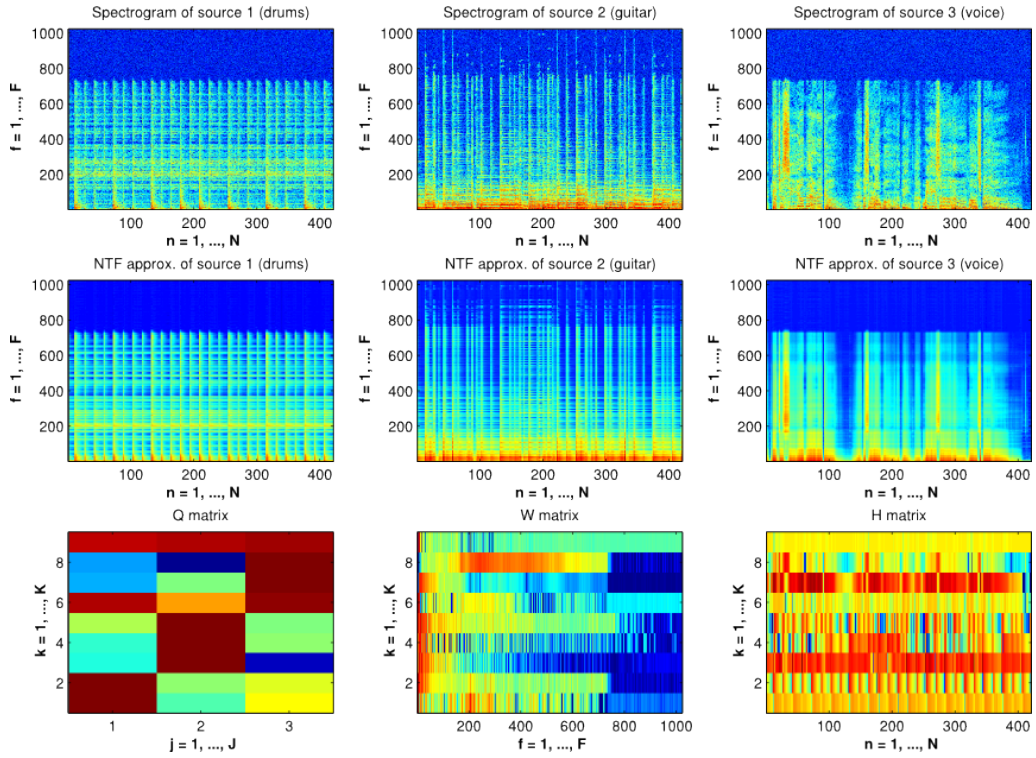


Figure 1.6: Nonnegative Tensor Factorization (NTF) for joint compression of the spectrograms of the sources as in (1.12). From [J.Oze13].

Multichannel ISS. Later on, I generalized these approaches to allow for multichannel signals. For a given TF entry (t, f), the mixing model (1.2) becomes $\mathbf{x}_{ft} = \sum_j \mathbf{y}_{jft}$, that are all I-dimensional complex vectors. For stereo signals, we have $I = 2$. The Local Gaussian Model (LGM), proposed in 2010 by N. Duong *et al.* assumes:

$$\mathbf{y}_j \sim \mathcal{N}_c\left(0, v_{jft}\mathbf{R}_{jf}\right), \tag{1.13}$$

where $v_{jft} \geq 0$ is still called a PSD, while $\mathbf{R}_{jf} \succeq 0$ is called the *spatial covariance matrix* (SCM). It encodes the correlations between the channels. When it is rank-1, we obtain the common *convolutive model* $\mathbf{x}_{ft} = \sum_j \mathbf{a}_{jf} s_{jft}$.

Just like in my previous work on single channel signals, I worked to understand what were the assumptions on time-domain signals that lead to the LGM (1.13). It turns out it can be understood as modeling a sound source as *non punctual*, meaning it is composed of many parts vibrating independently with the same PSDs [B.Liu12, SLP+12].
In any case, I generalized my previous work on single channel ISS to also allow for stereo signals, and I benchmarked all the methods proposed by the community [LGS+12].

Further ISS developments. Continuing this work on ISS lead to several publications done in a very stimulating collaborative environment. Along with my co-authors, we significantly improved quality of the source estimates for real mixtures. Our advances involved compressing the sources spectrograms with very fast randomized methods [RCL17], quantization-aware parameter estimation [RLB17] to optimize the bitrate required to transmit the NTF parameters in (1.12), separation directly on the MP3 bitstream [ZGL13]. A patent was granted for those techniques [P.Gir10]. In a few years, we were able to come up with methods to recover good estimates for the original source signals from real mixtures, **with less than 2kbps of additional bitrate** [LBR13].

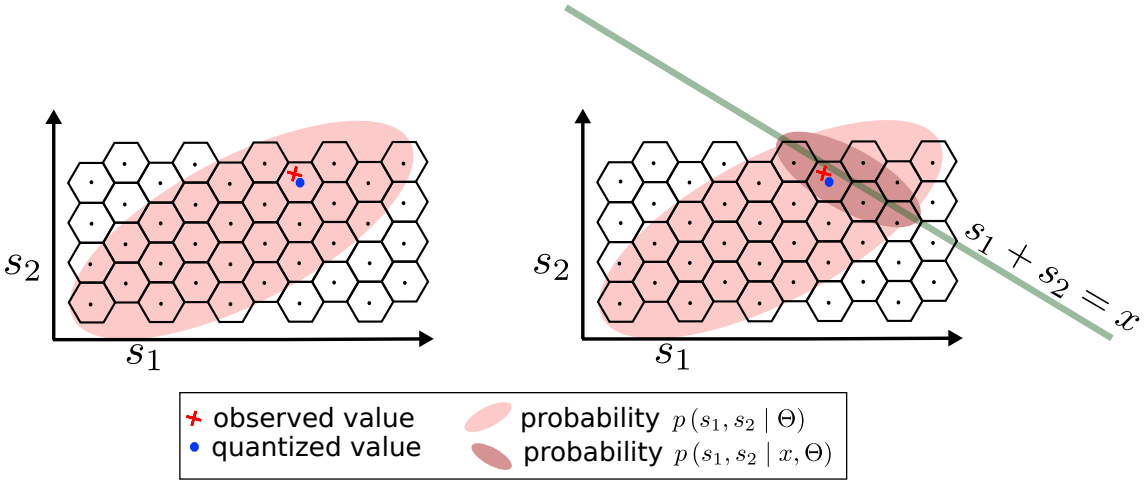


Figure 1.7: Coding-based Informed Source Separation, when source separation meets source coding. Left: standard source coding. Right: posterior source coding that exploits the mixture.

ISS as parametric audio coding. In all the methods referenced above, it is impossible to reach perfect reconstruction, because performance of the Wiener filter is bounded. A key fact about the Gaussian formulations (1.11) and LGM (1.13) are indeed to provide a full *posterior (Gaussian) distribution* for the sources given the mixture. This means they not only provide a way to reconstruct the original signals in the MMSE sense, they also *provide an estimate of its error*, which will never be 0, unless the mix contains at most one source.

In this respect, it was pointed out to me by A. Ozerov that the Wiener filter could be considered as a very particular case of *parametric coding*, which is a large family of compression methods that consist in the following steps:

- We choose a family of signals, say $F(\Theta)$, parameterized by some scalars Θ.
- To compress a signal $\mathbf{y}_j$, we compute

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\| \mathbf{y}_j - F(\Theta) \right\|^2, \tag{1.14}$$

  which is transmitted to the decoder. In real applications, the loss function may include some perceptual modeling.
- At the decoder, we reconstruct the signals through $F(\widehat{\Theta})$, which is a good approximation to $\mathbf{y}_j$ by construction (1.14).

Taking $\Theta = \{\mathbf{V}_1, \dots \mathbf{V}_J\}$ and $F_j(\Theta) = \mathbf{V}_j / \left(\sum_{j'} \mathbf{V}_{j'}\right) \bullet \mathbf{X}$, we indeed notice that Wiener filters *can* be seen as parametric coding, with the original twist that both the encoder and the decoder share the mixture $\mathbf{X}$ as a common *side information*.
With this realization, we discovered that ISS had already been investigated for several years under the name "spatial audio object coding" (**SAOC**), by the audio coding community. State of the art methods there involved sophisticated arithmetic coding techniques, but very basic PSD models for the sources, which didn't exploit the long-term redundancies inherent to music signals.

Posterior source coding. With this realization came the next step that, I believe, was one of the most elegant works I contributed to and that was coordinated by A. Ozerov. Its core idea is to leverage *source coding*, a branch of information theory, to allow for arbitrary quality of the reconstructed sources and thus to go beyond the limited quality of ISS. We do it by exploiting the posterior distribution (1.11) for the sources.

In Figure 1.7, I picture the core idea. Let's assume we want to transmit two scalar values, say $s_1 \in \mathbb{R}$ and $s_2 \in \mathbb{R}$, that may for instance be the Modified Discrete Cosine Transform (MDCT)

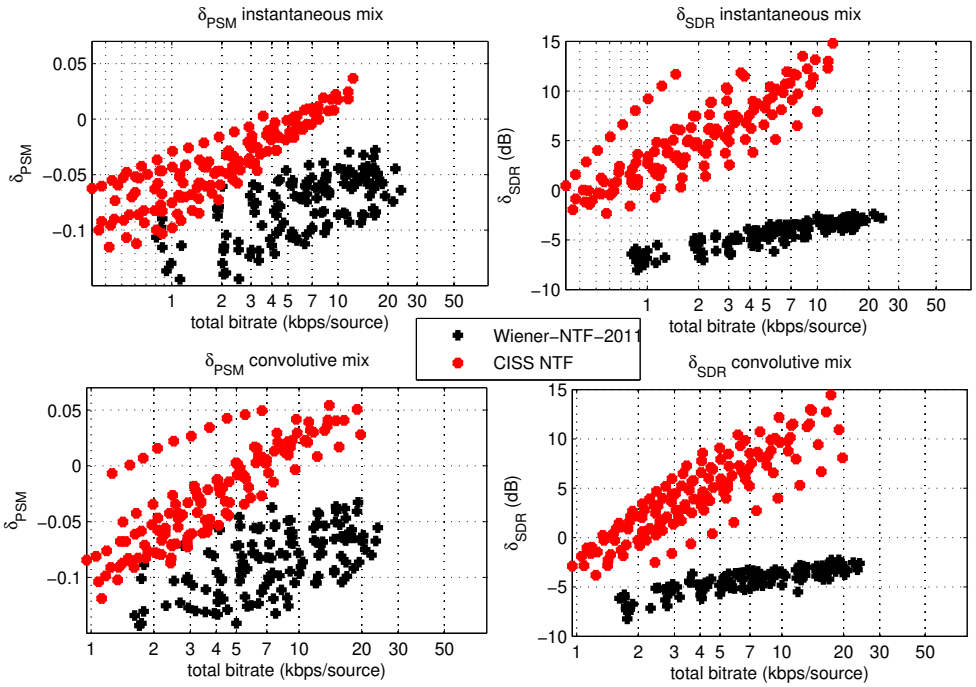


Figure 1.8: Performance of CISS. Up: simple mixtures, Down: convolutive (more realistic) mixtures. Left: perceptual scores. Right: signal to distortion ratio. 0 in the y-axis always indicates the performance of ideal Wiener filter (1.4).

values of the sources at some given TF point. After picking a desired distortion, we can *quantize* the sources, so that all the points within a given *cell* are encoded as its centroid. Cells can be arbitrarily small depending on the desired distortion. Then, if a joint distribution $\mathbb{P}(s_1, s_2)$ is available as in Figure 1.7 (left), source coding provides an optimal way to encode the cell position, in terms of bitrate minimization. Basically, cells that are likely will require less bits than the other ones. In this context, **posterior source coding** as proposed in [OLBR11, J.Oze13, LOBR12] and illustrated in Figure 1.7 (right) replaces model $\mathbb{P}(s_1, s_2)$ by its *a posteriori* version $\mathbb{P}(s_1, s_2 \mid x)$ derived as in (1.11), with $x = s_1 + s_2$. As can be seen visually, taking the mixture into account strongly restricts the number of likely cells, for the same price to pay in terms of models parameters. This idea was coined in as Coding-based ISS (CISS) and leads to both a very strong reduction in bitrate and arbitrarily good performance, as exemplified in Figure 1.8. We detail it at length in [J.Oze13, B.Liu12]. A version that includes perceptual models was presented in [KOLG14].

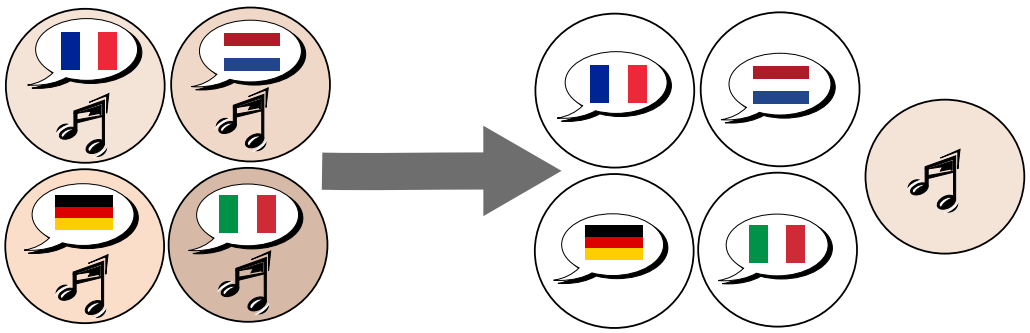


Figure 1.9: Separation of international versions from the same movie [LL10]. Exploiting the redundancy, we can isolate the individual dialogues, even if they are never observed alone.

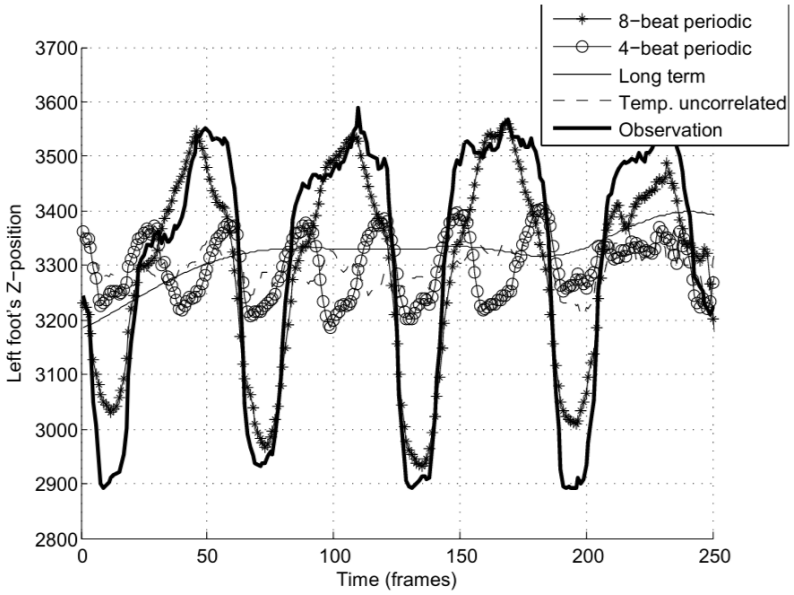


Figure 1.10: The vertical position of the left foot from one dancer decomposed as a sum of several latent components. From [LDA+12].

1.1.4 Gaussian follow-ups

After the initial realization that GP are perfectly appropriate to achieve source separation [J.Liu11, LBR11], I had occasional collaborations on their use for various types of signals, that I briefly review now.

- **Biomedical signals.** C. Damon, a fellow researcher from Telecom ParisTech was working on biosignals analysis (EEG and MEG). We applied Gaussian source separation based on NTF model to separate useful signals from artifacts in [DLGE13a, DLGE13b].
- **Interference reduction.** Along with my student D. Di Carlo, we applied the Gaussian separation framework for computationally effective reduction of interferences in live recordings from the Montreux Jazz Festival [DCDL17]. This echoes the early investigations I did with P. Leveau on the related problem of dialogue-music separation from several international versions of the same movie [LL10], which is illustrated in Figure 1.9.
- **Alternative representations.** I had collaborations with several researchers on the topic of using other representations than the STFT for Gaussian-based music separation. This includes constant-Q transform [FLBR12], a new *Common-Fate Transform* we proposed that is inspired from Gestalt theory [SLB+16] and some related multi-resolution and modulation-based representations [PPL17].
- **Dance movements.** My fellow researcher A. Dremeau at Telecom ParisTech handled data corresponding to the spatial positions of several joints from the body of dancers during a performance, yielding interesting space-time data. We applied GP separation to them to decompose movements into explanatory latent components [LDA+12]. One example is depicted in Figure 1.10.

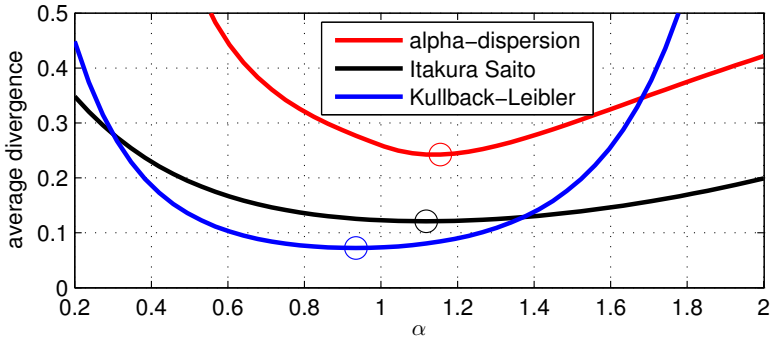


Figure 1.11: Divergence between the α-spectrogram of the mix and the sum of the α-spectrograms of the sources, according to several metrics [LB15]. Additivity of power-spectrograms ($\alpha = 2$) is clearly not happening.

1.1.5 Alpha-stable processes

As related above, I spent most of my Ph.D. working on the Gaussian framework for source separation, notably for its many interesting features in terms of (informed) audio coding or for the separation of signals defined on non-usual index sets.
However, it turns out **there are cases where the Gaussian model is not appropriate**. In particular, several widespread and effective practices in audio processing that I review below are not straightforwardly compatible with Gaussian assumptions [B.Liu12].

Having identified missing parts of the puzzle, I spent a significant amount of my time looking for theoretical grounds on which these practices could stand. I found out that a satisfying answer lied in **α-stable distributions and processes**. Their introduction to source separation stands out as the second core contribution from my past work, that I present in this section. I must acknowledge the help from many collaborators on the way, among which M. Fontaine and R. Badeau played a very prominent role.

Established tricks not compatible with Gaussian assumptions. When the signals are taken as locally stationary Gaussian processes, the key objects that are manipulated are second order statistics: the power spectrograms of the sources are assumed to add up to yield that of the mixture, and they are used to construct Wiener filters, through the simple ratio procedure (1.4). Multichannel models as (1.13) involve slightly more sophisticated machinery but *in fine* work in the same way: *covariance matrices* are used instead of scalar variances.

However, the fact is that a large part of the audio separation research was not fitting in this canvas, hence departing from the Gaussian ”story”.

- **Decomposing magnitude spectrograms**. From a general perspective, audio modeling could be understood as hand-picking a specific parametric model like the NTF (1.12) for the PSDs of the sources, and train them under the constraint that their sum would fit the spectrogram of the mixture. It was a known fact from the literature that the Gaussian assumption would provide an appropriate loss function for this strategy, namely the Itakura-Saito divergence for power spectrograms, see e.g. [B.Liu12] and references therein.
  However, a large part of the community was considering **magnitude spectrograms** instead, so that the model would not be compatible with Gaussian assumptions. Regardless of the lack of theoretical foundations, it is indeed the case that *fractional* α-spectrograms (magnitude of STFT raised to the power $\alpha \in (0, 2]$) are better candidates to comply with the additivity assumption for additive sources, as depicted in Figure 1.11. This experimentally validates the idea of using them instead of power spectrograms.
  In this context, some researchers like P. Smaragdis or T. Virtanen *did* propose a probabilistic model called **Probabilistic Latent Component Analysis (PLCA)** that would justify this approach theoretically under Poisson distributions, but I must say was not fully satisfied with it. Indeed, even if it *does* justify additivity, nothing in PLCA may be invoked to understand why

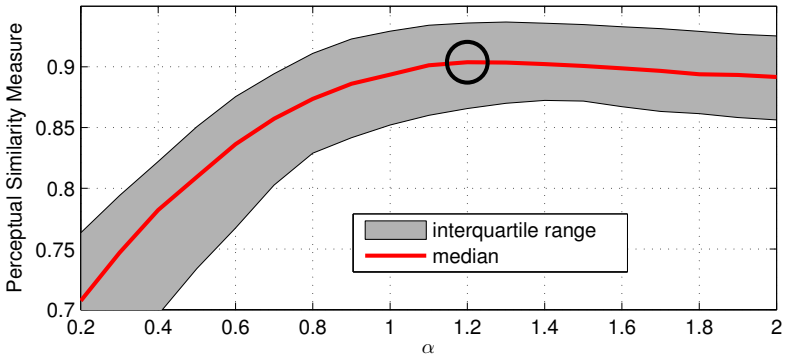


Figure 1.12: Perceptual quality of oracle separation of a mixture through a *ratio mask* involving the α-fractional spectrograms of the sources. $\alpha = 1$ is magnitude, $\alpha = 2$ is power. The best overall quality is obtained here somewhere around $\alpha = 1.2$ [LB15].

the magnitude rather than any other fractional α-spectrogram should be used: it doesn't come with a corresponding probabilistic model for *waveforms*, contrarily to the Gaussian case. This leaves the practitioner with the uncomfortable need to just **ignore all the preprocessing when choosing the probabilistic model**.
For this reason, even if leads to good performance in practice, I felt there was room for a deeper understanding regarding the use of fractional spectrograms.

- **Filtering with fractional spectrograms**. In the Gaussian case, as soon as the sources spectrograms are estimated, they can readily be used to construct a Wiener filter (1.4) or its multichannel variant, to recover the sources signals. Once more, although this procedure is only justified theoretically for power spectrograms, it is common practice to just use it also with any fractional α-spectrogram, notably magnitude $\alpha = 1$. Being understood that this is not a Wiener filter anymore, the method is often referred to in a loose way as a **ratio mask** and is widely acknowledged as working well. I illustrate this in Figure 1.12, taken from [LB15], where we see the average perceptual quality of the estimates as a function of α.

- **Parameterized Wiener filters**. As mentioned above, ratio masks simply consist in using fractional α-spectrogram in (1.4), instead of only the power $\alpha = 2$, which is the only one supported by the Gaussian theory. Another filtering trick that has been widely used for decades in speech enhancement is called the *parameterized Wiener filter* (see [FLGB17]). It is mostly used in speech enhancement, hence with two sources only: speech **S** and noise, with respective PSD estimates **V** and **N**. It replaces (1.4) by:

$$\widehat{\mathbf{S}} = \frac{\mathbf{V}}{\mathbf{V} + k\,\mathbf{N}} \bullet \mathbf{X}, \tag{1.15}$$

where the parameter $k > 0$ mitigates the importance of the estimated noise in the resulting filter. This "trick" is widely used with $k \in (0, 1]$ to attenuate the perceptual *overprocessing* often felt when using the standard Wiener filter.
Although this parameterized Wiener filter only departs slightly from the standard one (1.4), we were not aware of a solid understanding for it.

α-stable distributions and processes. While I was investigating the alternatives to Gaussian processes, I spent some time reading the incredibly thorough book *Stable Non-Gaussian Random Processes*, by Samorodnitsky and Taqqu. Although it was pretty challenging for me to get into this topic due to its mathematical prerequisites, it proved very rewarding. When I landed on their Theorem 4.1.2, reproduced in Figure 1.13, I suddenly realized that this α-stable framework could very well exactly be what I was looking for. This realization was the starting point of a whole episode of my research curriculum, leading to many collaborations (notably the Ph.D. investigations of M. Fontaine), 3 journal papers and 13 conference papers.

Theorem 4.1.2 *Let X_1 and X_2 be two jointly $S\alpha S$ random variables with $1 < \alpha \leq 2$. Then*

$$E(X_2|X_1) = \frac{[X_2, X_1]_\alpha}{\|X_1\|_\alpha^\alpha} X_1 \quad a.s. \tag{4.1.4}$$

Figure 1.13: Excerpt from *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, by Samorodnitsky and Taqqu (p 175). This page unlocked a host of new research for me, when I felt it could provide a theoretical grounding for ratio-masks.

For the text to be self consistent, I need to shortly introduce α-stable processes here. Since these distributions were not so common in music signal processing, we had to introduce them many times and I decided to take inspiration here from our short paper [LB15] that is reproduced on page 122, with a slightly more rigorous treatment. Longer developments may be found in [J.Fon20].

Let **v** be a random vector of length L. We say it is strictly stable if for any positive numbers A and B, there is a positive number C such that

$$A\,\mathbf{v}^{(1)} + B\,\mathbf{v}^{(2)} \stackrel{d}{=} C\,\mathbf{v}, \tag{1.16}$$

where $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are independent copies of **v** and $\stackrel{d}{=}$ denotes equality in distribution. In a source separation context, the *stability* property (1.16) is fundamental. It basically means that provided the sources are modeled as stable, so will be their mixture. It can be shown that for any random vector **v** satisfying (1.16), there is one constant $\alpha \in (0\,2]$ called the *characteristic exponent* such that C in (1.16) is given by:

$$C = (A^\alpha + B^\alpha)^{1/\alpha}.$$

We then say that **v** is α-stable. If **v** and $-\mathbf{v}$ furthermore have the same distribution, **v** is called symmetric α-stable, abbreviated as SαS. An important result is that the simple property (1.16) of an α-stable random vector permits to derive its characteristic function (ch.f.). As special cases, $\alpha = 2$ and $\alpha = 1$ respectively coincide with the Gaussian and Cauchy distributions.

α-stable distributions have an important number of desirable properties. One of the most famous is their ability to model data with very large deviations, making them a practical model for impulsive data in the field of robust signal processing. In practice, the closest α is to 0, the heavier are the tails of an α-stable distribution.

Just like in the Gaussian case, we say that a collection $\left\{\tilde{Y}(l)\right\}_{l\in\mathbb{L}}$ of r.v. is an α-stable random *process* if the vector $\tilde{\mathbf{Y}}(\mathbf{l}) \triangleq \left[\tilde{Y}(l_1), \ldots, \tilde{Y}(l_L)\right]$ is α-stable for any choice and any number of sample locations $l_1, \ldots, l_L$.

Isotropic complex SαS random variable Because it will be useful in the sequel, I mention here that a complex r.v. $y = z_1 + iz_2$ is called SαS if the random vector $[z_1\, z_2]$ is SαS. A particular case of interest in our context is when a complex SαS r.v. $y \in \mathbb{C}$ is isotropic, or circular, abbreviated $S\alpha S_c$, meaning that:

$$\forall\theta \in [0\,2\pi[\,, \exp(i\theta)\, y \stackrel{d}{=} y.$$

It can be shown that in the Gaussian case $\alpha = 2$, this is equivalent to real and imaginary parts z_1 and z_2 being independent and identically distributed (i.i.d.) Gaussian r.v., whereas for the case $\alpha < 2$, isotropy leads to the particular ch.f.:

$$y = z_1 + iz_2 \sim S\alpha S_c \Leftrightarrow \mathbb{E}\left[\exp\left(i\left(\theta_1 z_1 + \theta_2 z_2\right)\right)\right] = \exp\left(-\sigma^\alpha |\boldsymbol{\theta}|^\alpha\right), \tag{1.17}$$

where $|\boldsymbol{\theta}|$ is the Euclidean norm of the vector $[\theta_1\, \theta_2]$, and $\sigma > 0$ is a scale parameter. The real and imaginary parts of an isotropic complex SαS r.v. are *not* independent in general. As can be

seen, the isotropic complex SαS distribution is only parameterized by the scale parameter σ. For convenience, we write it $S\alpha S_c\left(\sigma^\alpha\right)$ and trivially have:

$$y_1 \sim S\alpha S_c\left(\sigma_1^\alpha\right) \text{ and } y_2 \sim S\alpha S_c\left(\sigma_2^\alpha\right), \text{ with } y_1 \perp y_2 \Rightarrow y_1 + y_2 \sim S\alpha S_c\left(\sigma_1^\alpha + \sigma_2^\alpha\right), \tag{1.18}$$

where $\perp$ denotes independence.

Stationary harmonizable α-stable processes An harmonizable process $\tilde{Y}(l)$ is defined as the inverse Fourier transform of a complex random measure Y with independent increments:[5]

$$\tilde{Y}(l) = \int_{-\infty}^{\infty} \exp\left(i\omega l\right) Y(d\omega), \tag{1.19}$$

where the r.v. $Y(d\omega) \in \mathbb{C}$ may be understood as the (complex) spectrum of $\tilde{Y}$, taken at frequency ω. Stating that Y has independent increments means that:

$$\forall \mathcal{A}, \mathcal{B} \subset \mathbb{R}, \mathcal{A} \cap \mathcal{B} = \emptyset \Rightarrow Y(\mathcal{A}) \perp Y(\mathcal{B}). \tag{1.20}$$

Let us now split the interval $[0, 1]$ into (many) F non-overlapping frequency *bins* $\{\Omega_1, \ldots, \Omega_F\}$ with centroids ω_f. Provided that: i/ F is large enough, ii/ signals have a limited bandwidth and iii/ are regularly sampled, we get:[6]

$$\tilde{Y}(l) \approx \sum_{f=1}^{F} \exp\left(i\omega_f l\right) Y(\Omega_f), \tag{1.21}$$

where all $Y(\Omega_f) \in \mathbb{C}$ are independent complex r.v. If the number of samples is sufficiently large, the discrete Fourier transform (DFT) of $\widetilde{\mathbf{y}} \triangleq \widetilde{Y}(\mathbf{l})$, that we write: $\mathbf{y} = [y_1, \ldots, y_F]$ approximates $Y(\Omega_f)$. In short, provided the frames are sufficiently long, the DFT coefficients of an harmonizable process can be considered independent. This is a very desirable property because it means that once the DFT has been computed, all its coefficients may be considered independently, instead of requiring a joint distribution of all the time-domain samples.

Now, it is a classical result that when Y is an isotropic complex Gaussian random measure with independent increments, $\tilde{Y}$ in (1.19) is furthermore stationary and we land back on the Gaussian story mentioned above. However, it is not the only way of guaranteeing that an harmonizable process $\tilde{Y}$ is stationary. In particular, a very important result in our context is that taking Y as an isotropic complex $\mathcal{S}\alpha\mathcal{S}_c$ random measure is equivalent to having $\tilde{Y}$ being both a stationary and an $\mathcal{S}\alpha\mathcal{S}$ random process, which is the natural extension of the Gaussian case to $\alpha < 2$. We write it $Y(d\omega) \sim S\alpha S_c\left(\sigma^\alpha\left(d\omega\right)\right)$, where $\sigma^\alpha\left(d\omega\right) \geq 0$ is the (nonnegative) *control* measure of the process. We call it its fractional power spectral density (α-PSD). We have:

$$y_f \triangleq Y(\Omega_f) \sim S\alpha S_c\left(\sigma_f^2 \triangleq \sigma^\alpha(\Omega_f)\right) \tag{1.22}$$

The main interest of the α-harmonizable model is to account for signals that both include large deviations and are stationary. It is thus interesting for audio signals, because they are stationary on short time-frames and often feature large dynamic ranges.
Following the same route as for Gaussian processes, we can split the signals into frames that we assume independent and α-harmonizable. This yields our final locally α-harmonizable model, with time varying spectral measures Y_t, boiling down to the following assumption:

$$y_{ft} \sim S\alpha S_c\left(\sigma_{ft}^\alpha\right). \tag{1.23}$$

As a last remarkable fact, just like the PSD is estimated empirically as the power spectrogram $|\, y_{ft} \,|^2$, the α-PSD is estimated by the α-spectrogram $|\, y_{ft} \,|^\alpha$, up to a constant that only depends on α.[7] As far as I know, our paper [LB15] was the first to introduce such processes for audio in a principled way.

[5] For conciseness, I detail the case $\mathbb{L} = \mathbb{R}$ here. The same goes for $\mathbb{L} = \mathbb{R}^D$.

[6] There are some trivial technicalities for real signals due to the fact that their spectrum is Hermitian, but I skip those here at the cost of rigour. They amount to $Y(\Omega_f) \approx Y(-\Omega_f)^\star$.

[7] These aspects are better discussed in [LOMG15].

Fractional spectrograms and the α-Wiener filter. From the derivations above, we immediately see that if the sources $\tilde{Y}_j$ are modeled as independent α-harmonizable, their mixture is also α-harmonizable and we have:

$$x_{ft} = \sum_j y_{jft} \sim S\alpha S_c \left(\sum_j \sigma^\alpha_{jft} \right). \tag{1.24}$$

This is the point where we can invoke the result depicted in Figure 1.13 to write that:

$$\mathbb{E}\left[y_{jft} \mid x_{ft} \right] = \frac{\sigma^\alpha_{jft}}{\sum_{j'} \sigma^\alpha_{j'ft}} x_{ft}, \tag{1.25}$$

which exactly generalizes (1.4) to α-PSD. The proof for this result is actually quite demanding and is given in [LB15]. Getting to it was a nice achievement in my view, because it meant I now had an understanding for the soft-masking procedure that manipulates α-spectrograms.

Likewise, the α-harmonizable model comes with a probabilistic interpretation for the additive α-spectrograms depicted in Figure 1.11. As a particular case, our results make it clear that a model assuming additive magnitude spectrograms should rather focus on the isotropic complex Cauchy distribution, rather than on Poisson assumptions. We exploited this model several times [LFB15, FNB+19].

Single-channel applications Once we had a nice framework for justificating the use of α-PSD for filtering, part of my work along with colleagues concerned the estimation of these α-PSD from the α-spectrograms of the mixture. Since NMF-based models like (1.5) were state of the art at this time, our developments focused on combining them with the α-stable framework.[8]
As a first natural step, we leveraged this new probabilistic framework to derive **new cost-functions for NMF** through maximum likelihood estimation. It turns out that only the Cauchy ($\alpha = 1$) and Levy ($\alpha = 0.5$) non-Gaussian cases are tractable in closed form in this way, but both have interesting features. Cauchy NMF [LFB15] proved interesting for denoising and appears as some kind of nonnegative "robust PCA". Levy NMF [MBL17b, MBL17a] is a principled framework for separating nonnegative random variables, which was an original problem.
Since I was not fully convinced that methods based on maximum likelihood were mandatory, we also departed from the fully probabilistic framework to also consider alternatives like **fractional lower-order moment-matching** methods. In that case, the probabilistic model provides an analytic formula regarding the moments of the data (fractional, logarithmic, etc), and we use an optimization method with an unrelated cost-function to estimate those. Although arguably less elegant, the method had interesting results [LOMG15, FLGB17].

1.1.6 Multichannel α-stable signals

The single channel α-harmonizable model was now well understood and already opened some interesting tracks of research. However, it was desirable for practical applications to also handle multichannel signals, which *in fine* means extending the scalar α-stable model to multivariate observations. Apart from the Gaussian case that was state of the art for some years already, notably through (1.13), the only non-Gaussian model that is amenable to classical likelihood methods is multivariate Cauchy, that we discussed in in [FNB+19].
Except from those, α-stable random vectors were not so commonly considered in the audio literature to the best of my knowledge and demanded new ideas and methods that occupied a significant part of the research effort I did with collaborators.

[8] I also considered training neural networks with cost-functions derived from this framework [J.Nug16], but I present this line of work later.

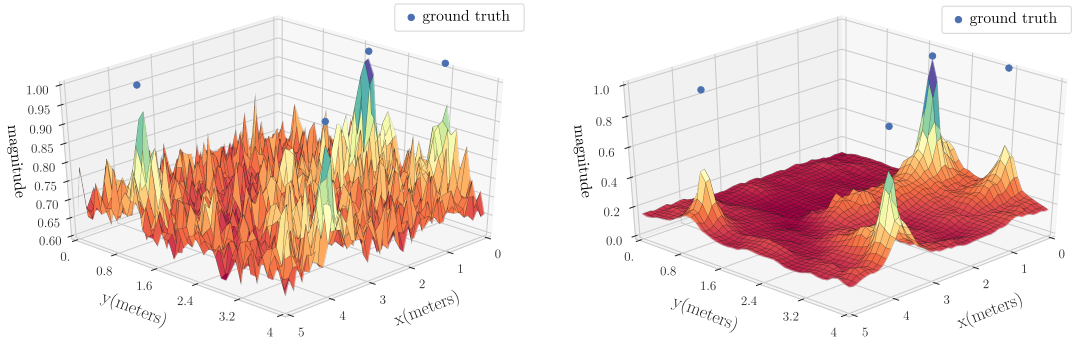


Figure 1.14: (left) Imaging with the Steering Response Power classical method. (right) Method proposed in [FVLB17a] based on a α-stable model. Blue dots stand for the true positions of the sources. Figure from M. Fontaine's Ph.D. thesis.

Punctual multichannel extensions A first natural step for me was to assume some *narrow-band convolutive model*, as was usual in the literature. If $I \in \mathbb{N}$ is the number of channels (or microphones) and $\mathbf{y}_{ft} \in \mathbb{C}^I$ stand for the concatenation of the STFT of the signals waveforms, at TF bin (f, t). The model reads:

$$\mathbf{y}_{ft} = \mathbf{a}_f s_{ft}, \tag{1.26}$$

where the multichannel signal $\mathbf{y}_{ft}$ is now called the *spatial image* of some single channel *source* $s_{ft} \in \mathbb{C}$ and $\mathbf{a}_f \in \mathbb{C}^I$ is a *mixing filter* that is powerful enough to account for gains and phases differences between channels. Assuming all J sources have their own mixing filters, we get to:

$$\mathbf{x}_{ft} \approx \sum_{j=1}^{J} \mathbf{a}_{jf} s_j(f,t) = \mathbf{A}_f \, \mathbf{s}_{ft}, \tag{1.27}$$

which is a very common model in audio. The original twist we considered is to pick an α-harmonizable model for the sources $\tilde{S}_j$ as above and didn't consider the actual mixture, but rather *projections* of it $\langle \mathbf{u}, \mathbf{x}_{ft} \rangle$, where $\mathbf{u} \in \mathbb{C}^I$. Indeed, the distribution of such projections is given by:

$$\forall \mathbf{u} \in \mathbb{C}^I, \langle \mathbf{u}, \mathbf{x}_{ft} \rangle \sim S\alpha S_c \left(\sum_j | \langle \mathbf{u}, \mathbf{a}_{jf} \rangle |^\alpha \, \sigma^\alpha_{jft} \right), \tag{1.28}$$

where σ^α_{jft} is the α-PSD for source j. From this point, assuming the projections to be independent and learning both the mixing filters $\mathbf{a}_{jf}$ and the α-PSD simultaneously could be an option, but it turns out rather challenging.
Instead, a successful idea was to pick an *acoustic model*, that acts as a given family of mixing filters $\{\mathbf{a}_f(p)\}_{p \in \mathbb{P}}$ indexed by a position $p \in \mathbb{P}$, where $\mathbb{P}$ is a finite set of P positions. It can correspond to a grid within a room $\mathbb{P} \subset \mathbb{R}^3$ or a set of panning and phase: $\mathbb{P} = [0, 1] \times [0, 2\pi]$. Anyways, we now assume we know the filters and that there is a source $\tilde{S}_p$ at each position. The model becomes:

$$\forall \mathbf{u} \in \mathbb{C}^I, \langle \mathbf{u}, \mathbf{x}_{ft} \rangle \sim S\alpha S_c \left(\sum_p | \langle \mathbf{u}, \mathbf{a}_f(p) \rangle |^\alpha \, \sigma^\alpha_{pft} \right), \tag{1.29}$$

where the only unknowns are now the α-PSD σ^α_{pft} of the sources, that can actually be understood as a kind of *heatmap* of the acoustic strength of the signal originating from position $p \in \mathbb{P}$. Concerning the actual choice of the projections vectors $\mathbf{u}$, we usually took a set of P projections $\mathbf{u}_p = \mathbf{a}_f(p)$, but random projections also fit in the canvas.

This approach opened the way to many different investigations. First, M. Fontaine along with collaborators successfully exploited this framework for acoustic imaging [FVLB17a, FVLB17b], with a method illustrated in Figure 1.14. Second, we also found it useful for the separation of stereo music signals, under the name PROJET: projection-based estimation technique [J.Fit16, FLB16, FRL17]. Much more details and developments can be found in these papers.

General diffuse case. From the point of view of the mixture, each source signal in the previous punctual model originates from a fixed and unique *direction* $\mathbf{a}_f(p)$, so that the directions that actually feature "energy" correspond to a set of measure 0 within the set of all possible positions. It turns out that this consideration is in line with the way multivariate $S\alpha S_c$ distributions are defined in the general case. Let $\mathbb{S}^I$ be the I-dimensional sphere in $\mathbb{C}^I$. We call its elements $\boldsymbol{\theta} \in \mathbb{S}^I$ *directions*. An $S\alpha S_c$ random *vector* is defined through its ch.f., which reads:

$$\mathbf{y} \sim S\alpha S_c\left(\Gamma_{\mathbf{y}}\right) \Leftrightarrow \forall \mathbf{u} \in \mathbb{C}^I, \phi_{\mathbf{y}}(\mathbf{u}) \triangleq \mathbb{E}\left[\exp\left(i\Re\left\langle \mathbf{u}, \mathbf{y}\right\rangle\right)\right] = \exp\left(-\int_{\boldsymbol{\theta}\in\mathbb{S}^I} |\left\langle \mathbf{u}, \boldsymbol{\theta}\right\rangle|^{\alpha} \Gamma_{\mathbf{y}}\left(d\boldsymbol{\theta}\right)\right), \tag{1.30}$$

where $\Re z$ is the real part of $z \in \mathbb{C}$ and the main object of interest is $\Gamma_{\mathbf{y}}$. It is a nonnegative measure on $\mathbb{S}^I$ that is commonly called the *spectral measure* of the r.v. $\mathbf{y}$, but that we decided to rename its *spatial measure* in a signal processing context, to avoid any confusion with what is usually called "spectral" in this community. If $\Theta \subset \mathbb{S}^I$ is a subset from the sphere, I understand $\Gamma_{\mathbf{y}}(\Theta) \geq 0$ as giving the average scale of the contributions originating from directions within Θ. It is trivial to see that:

$$\mathbf{y}_1 \sim S\alpha S_c\left(\Gamma_1\right) \perp \mathbf{y}_2 \sim S\alpha S_c\left(\Gamma_2\right) \Rightarrow \mathbf{y}_1 + \mathbf{y}_2 \sim S\alpha S_c\left(\Gamma_1 + \Gamma_2\right), \tag{1.31}$$

generalizing the additivity property for α-stable vectors. This is illustrated in Figure 1.15.

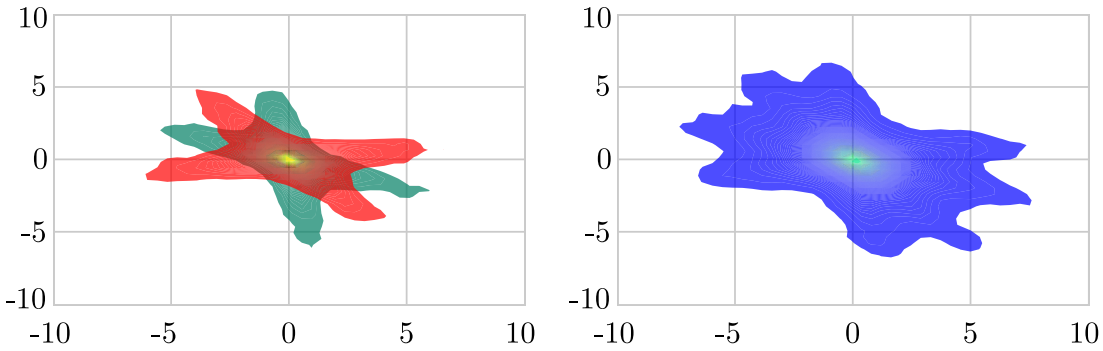


Figure 1.15: (left) The spatial measures of two bivariate α-stable random vectors. (right) The spatial measure of their sum, which is the sum of the spatial measures. From [J.Fon20]

It turns out that the **punctual particular case** (1.27) combined with α-harmonizable sources leads to the spatial measure $\Gamma_{\mathbf{x}}$ being a sum of Dirac. Without loss of generality, let's assume that all mixing filters $\mathbf{a}_{jf}$ are unit length.[9] I omit the f, t indices for readability:

$$\mathbf{x} = \mathbf{A}\,\boldsymbol{s} \text{ with } s_j \sim S\alpha S_c\left(\sigma_j^{\alpha}\right) \Rightarrow \Gamma_{\mathbf{x}} = \frac{1}{2}\sum_{j=1}^{J}\left[\delta\left(\mathbf{a}_j\right) + \delta\left(-\mathbf{a}_j\right)\right]\sigma_j^{\alpha}. \tag{1.32}$$

This result validates the interpretation of the spatial measure as a very appropriate object for analyzing multivariate mixtures.
In the general case, we came out with a **representation theorem** [J.Fon20], that decomposes a multivariate α-stable vector into an infinite sum of independent contributions originating from all directions, each one with a scale parameter given by the spectral measure:

$$\mathbf{y} \sim S\alpha S_c\left(\Gamma_{\mathbf{y}}\right) \Leftrightarrow \mathbf{y} \overset{d}{=} \int_{\boldsymbol{\theta}\in\mathbb{S}^I} \boldsymbol{\theta}\, Y\left(d\boldsymbol{\theta}\right), \text{ with } \forall \Theta \subset \mathbb{S}^I, Y\left(\Theta\right) \sim S\alpha S_c\left(\Gamma_{\mathbf{y}}\left(\Theta\right)\right). \tag{1.33}$$

This result is proved in [J.Fon20]. The random measure Y on the sphere, with independent increments, can be understood as the spatial analogous to the spectrum we have in (1.19). This decomposition and model (1.30) in general opens many interesting research questions in probabilistic signal processing. Some of them were already investigated by M. Fontaine during his Ph.D. [J.Fon20], where several multichannel filtering methods based on α-stable models were proposed, that generalize Wiener filtering and strongly outperform it whenever $\alpha < 2$.

[9] Any $\|\mathbf{a}_{jf}\| \neq 1$ could equivalently be assigned to σ_{jft}

Conditionally Gaussian α-stable vectors. Finally, we also considered Elliptically-contoured α-stable random vectors, which are a particular case of (1.30) for which the ch.f. can be written as:

$$\mathbf{y} \sim \mathcal{ES}_c\left(\mathbf{\Sigma_y}, \alpha\right) \Leftrightarrow \forall \mathbf{u} \in \mathbb{C}^I, \phi_{\mathbf{y}}(\mathbf{u}) = \exp\left(-\left|\frac{\mathbf{u}^\star \mathbf{\Sigma_y} \mathbf{u}}{2}\right|^{\alpha/2}\right), \tag{1.34}$$

where $\cdot^\star$ denotes complex conjugation and $\mathbf{\Sigma_y}$ is a $I \times I$ positive semi-definite *scatter matrix*. It coincides with the covariance matrix in the Gaussian $\alpha = 2$ case. Unfortunately, the sum of $\mathcal{ES}_c$ r.v. is not $\mathcal{ES}_c$ itself except in the Gaussian ($\alpha = 2$, any I) and the scalar (any $\alpha \in (0, 2]$, $I = 1$) cases, so that we cannot easily just model sources and mixtures with such a model.
However, elliptically contoured α-stable vectors have a very remarkable property: they are conditionally Gaussian. In short, they behave like a **Gaussian random vector whose covariance is randomly perturbed** by a nonnegative noise with heavy tails. This is written as:

$$\mathbf{y} \sim \mathcal{ES}_c\left(\mathbf{\Sigma_y}, \alpha\right) \Leftrightarrow \begin{cases} \phi \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\left(2\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}\right), \\ \mathbf{y} \mid \phi \sim \mathcal{N}_c\left(0, \phi\mathbf{\Sigma_y}\right), \end{cases} \tag{1.35}$$

where $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$ is the positive $\alpha/2$ stable distribution, for nonnegative variables. It is a particular case of α-stable scalar distributions which is totally skewed to the right and that includes Levy as a special case [MBL17b]. $\phi \geq 0$ in (1.35) is called the *impulse* variable in our papers [LSL$^+$17, FSL$^+$18, J.Sim15, SEL$^+$18]. As can be seen, its distribution only depends on α and it should really be understood as a random perturbation over a Gaussian model. Along with collaborators, I exploited the conditionally-Gaussian property of α-stable r.v. many times.
A first idea in [LSL$^+$17] was to take the ***joint* distribution of sources and mixtures as** $\mathcal{ES}_c$. The rationale for this choice was to introduce some **robustness**: In a Bayesian optimization context, randomly initialized parameters are typically very bad in the first phases of training, so that the probabilistic model should allow for large deviations around its current mode, which is precisely what the impulse variable ends up doing: it gets large whenever the model is poorly accounting for the observations. This idea proved extremely effictive in a CISS scenario, where the impulse variables are estimated again at the decoder and lead to a much closer fit to the data, causing a drastic reduction in bitrate.
Another idea considered in [SEL$^+$18, J.Sim15] is to exploit conditional Gaussianity to derive a **Markov-Chain Monte Carlo (MCMC) algorithm for NMF** in a α-stable context.
Finally, we also used conditional Gaussianity to do inference in hybrid models, as I discuss later.

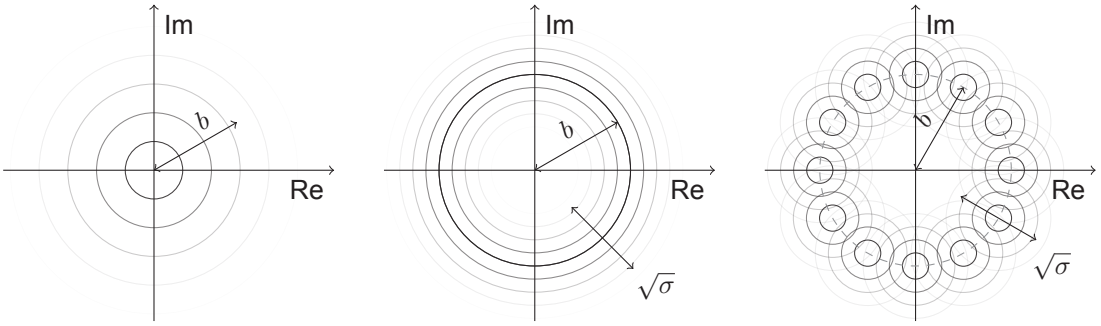
1.1.7 Hybrid models

As can be seen, my efforts to propose probabilistic models for time series and audio was mostly occupied with α-stable models, including the important Gaussian case.
However, the models I presented so far have the particularity of being both *unimodal* and *the same for all sources*. These two particularities eventually appeared as limitations, that I addressed with colleagues, mostly through **mixture models**, as I present now.

Separation with (approximately) known magnitude: the BEADS model In the informed source separation (ISS) scenario presented above, the decoder is provided with *very good* approximations for the sources magnitudes in the STFT domain while the phases are unknown.[10] Likewise, the most recent advances in source separation based on deep learning in the frequency domain lead to methods capable of also providing very good estimates for those magnitudes blindly.

In both cases, this calls for better probabilistic models than those presented above, along with corresponding separation methods. The reason for this is summarized in Figure 1.16. Both Gaussian and α-stable models actually put the highest probability mass on 0 and are thus incompatible with a the prior on *magnitude* that would be expected in such cases. On the contrary, a "donut-shaped" distribution seems more appropriate and even allows for some prescribed error on the magnitude. Its problem is: it is hard to use for separation.

[10] Due to their wildly uncorrelated random nature, the phase coefficients for the sources are typically very costly to transmit, contrarily to the largely redundant magnitudes.



(a) The Local Gaussian Model is tractable, but inconsistent with a prior on magnitude.

(b) A *donut*-shaped distribution is not tractable, but complies with the prior.

(c) The BEADS model combines advantages of both (12 components)

Figure 1.16: The BEADS model, from [LRD18].

In [LRD18], we proposed a Gaussian mixture model (GMM) called BEADS (Bayesian expansion to approximate the donut shape...) that is both tractable and complies with a prior on magnitudes. Basically, each source is modeled as a mixture of Gaussian components located along a circle of the complex plane with the prescribed magnitude. The mixture is readily shown to also be a GMM and we can compute the posterior distribution in a tractable way, as illustrated in Figure 1.17.
It turns out the BEADS model didn't have much impact for now, but I find it nice and wanted to include it here.

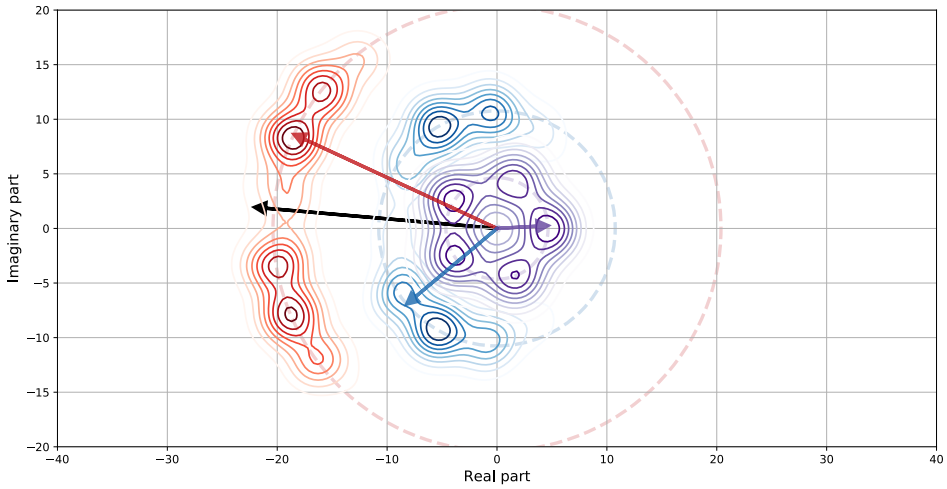


Figure 1.17: Separation with the BEADS model. The true sources are drawn with colored arrows and the mix is in black. Dotted lines are their true magnitudes, serving as a prior for BEADS separation. The contour plots for the marginal posterior distribution of each source can be seen to coincide with the magnitude prior. The joint posterior distribution of the sources is multimodal, with modes that approximate the positions of the original sources very well, up to symmetry ambiguities.

Multi alpha denoising The work presented above on α-harmonizable processes made a step towards more flexibility in the modeling, allowing to choose a characteristic exponent α for the sources. The smaller α is, the heavier are the tails of the model. However, experience showed it suffers from a weakness: **all sources have to share the same characteristic exponent** α. This is counter-intuitive, because we expect different sources to have different degrees of impulsiveness and it seems desirable to have a model that is flexible enough to allow for different α_j. I did some collaborations on this topic [FLGB17, SEL$^+$18, FSL$^+$18], that I summarize now.

With the motivation of picking a different characteristic exponent α_j per source, a typical model we would like to choose in the STFT domain is:

$$\mathbf{y}_{jft} \sim \mathcal{ES}_c(\mathbf{\Sigma}_{jft}, \alpha_j), \tag{1.36}$$

where the symmetric elliptical α-stable distribution $\mathcal{ES}_c$ has been presented earlier. The difficulty introduced by this model is that the distribution for the mixture $\mathbf{x}_{ft}$ is not readily tractable.

Still, the **conditional Gaussianity** of α-stable variables that I mentioned above in (1.35) can be invoked to the rescue. Introducing the latent impulse variables ϕ_{jft}, we have:

$$\forall j, \mathbf{y}_{jft} \sim \mathcal{ES}_c\left(\mathbf{\Sigma}_{jft}, \alpha_j\right) \Leftrightarrow \begin{cases} \phi_{jft} \sim \mathcal{P}\frac{\alpha_j}{2}\mathcal{S}\left(2\left(\cos\frac{\pi\alpha_j}{4}\right)^{2/\alpha_j}\right), \\ \mathbf{y}_j \mid \phi_{jft} \sim \mathcal{N}_c\left(\mathbf{0}, \phi_{jft}\mathbf{\Sigma}_{jft}\right), \end{cases} \tag{1.37}$$

so that, conditionally on the impulse variables ϕ_{jft}, the mixture $\mathbf{x}_{ft}$ has distribution:

$$\mathbf{x}_{ft} \mid \{\phi_{jft}\}_j \sim \mathcal{N}_c\left(\mathbf{0}, \sum_j \phi_{jft}\mathbf{\Sigma}_{jft}\right). \tag{1.38}$$

While we only considered the scalar case in [FLGB17, SEL$^+$18] and introduced the multichannel case in [FSL$^+$18]. As can be seen from (1.38), provided we successfully estimate the impulse variables, the mixture distribution has a classical Gaussian distribution and the MMSE estimate for, say, source $\mathbf{y}_j$ is given by the classical (multichannel) Wiener filter:

$$\mathbb{E}\left[\mathbf{y}_{jft} \mid \mathbf{x}_{ft}, \boldsymbol{\phi}_{ft}\right] = \phi_{jft}\mathbf{\Sigma}_{jft}\left(\sum_{j'} \phi_{j'ft}\mathbf{\Sigma}_{j'ft}\right)^{-1}\mathbf{x}_{ft}. \tag{1.39}$$

Of course, the main difficulty of the approach now lies in the fact we are left to estimate not only the model parameters $\mathbf{\Sigma}_{jft}$ but also the numerous impulse variables ϕ_{jft}. Along with several collaborators and students, we considered two main routes to deal with this issue.

The first method proposed in [FLGB17] consists in **simply replacing the impulse variables by their most probable value**, after noticing that their prior distribution (1.37) does not depend on any parameter but α_j. We used the median $\hat{\phi}_j = \mathbb{M}\left[\phi_{jft}\right]$ as an estimate, which is computed beforehand over a population drawn wrt. the prior distribution in (1.37). As crude as it may seem, we showed that the resulting Wiener filter actually behaves very similarly to the optimal one, and the method appears as a nice theoretical understanding for the parameterized Wiener filter (1.15). Indeed, we then have:

$$\hat{\mathbf{y}}_{jft} = \mathbf{\Sigma}_{jft}\left(\sum_{j'} \frac{\hat{\phi}_{j'}}{\hat{\phi}_j}\mathbf{\Sigma}_{j'ft}\right)^{-1}\mathbf{x}_{ft}, \tag{1.40}$$

so that the "mysterious" $k > 0$ coefficient in (1.15) could be interpreted as a ratio over average impulse variables, and hence be estimated automatically. The core advantage of this approach is computational efficiency: using multiple α_j doesn't come at any additional computational burden.

The second method we developed in [SEL$^+$18, FSL$^+$18] is to marginalize over the impulse variables $\boldsymbol{\phi}_{ft}$ whenever needed during training and inference through **MCMC strategies**, notably exploiting the Metropolis-Hastings algorithm. Acceptance probabilities involve Gaussian likelihoods and didn't prove so computationally demanding.

Mixtures of α-stable distributions A particularly interesting aspect of α-stable distributions to model audio signals that I didn't mention still is they offer sufficient variability to actually become interesting *nonparametric* options. More precisely, they are sufficiently permissive in terms of dynamics to to be appropriate models for the *marginal distribution* of the sources STFT $\mathbf{Y}_j$:

$$y_{jft} \sim \mathcal{ES}_c\left(\sigma_j, \alpha\right), \tag{1.41}$$

where the scale parameter crucially does not depend on the TF bin f, t anymore. This approach proved effective enough for source localization, where each location in space came with a source modeled with such a unique scale parameter [FVLB17a, FVLB17b].

In [KDL18], we went beyond this model (1.41) to pick a frequency-dependant characteristic exponent α_{jf}:

$$s_{jft} \sim \mathcal{ES}_c\left(\sigma_j, \alpha_{jf}\right), \tag{1.42}$$

and combined it with a punctual mixing model (1.27). Due to the sparse nature of the α-stable distributions, we furthermore assumed that only one source $z(f,t) \in \{1, \ldots, J\}$ has a significant contribution at any TF bin f, t, all other taken together having a small energy and hence being appropriately modeled as a Gaussian additive term $\mathbf{e}_j(f,t)$. This lead us to a *mixture model*:

$$\mathbf{x}_{ft} = \sum_{j=1}^{J} \mathbb{1}(z(f,t) = j) \left[\mathbf{a}_j(f) s_j(f,t) + \mathbf{e}_j(f,t)\right], \tag{1.43}$$

where $\mathbb{1}$ is the indicator function. This model being chosen, we exploited recently proposed methods based on **sketching** to estimate its parameters and obtained interesting results. In short, the method can be understood as estimating parameters based on generalized moments matching.

I note here that such moment matching methods are particularly suited for α-stable models, which lack a closed-form expression for their likelihood but not for their characteristic function. On this matter, I highlight here that our research on localization [FVLB17b, FVLB17a] also exploited such methods. Since I personally feel that sketching methods are very timely because they scale to very large datasets, it looks likely to me that α-stable distributions may gain some momentum from this perspective.

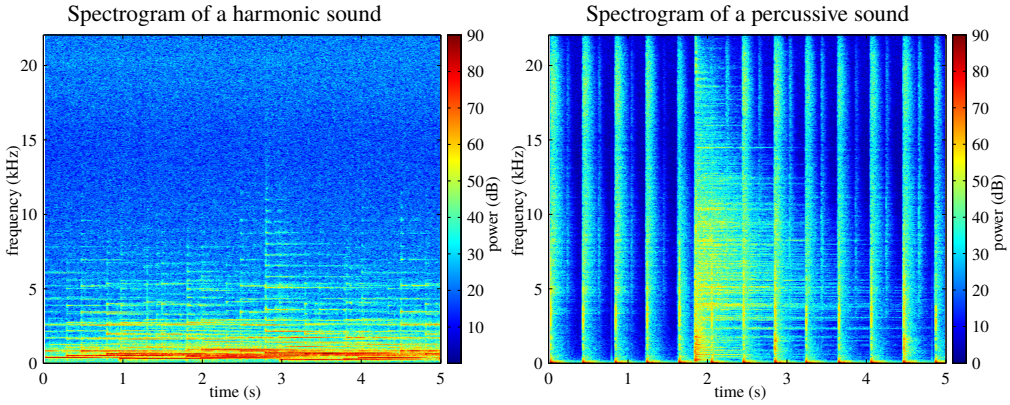


Figure 1.18: Harmonic sounds have spectrograms that are typically *locally constant* along time, while percussive sounds have spectrograms that are locally constant along the frequencies. This motivated local parametric models.

1.2 Spectrogram models and community service

The line of research I presented above in Section 1.1 was concerned with the numerous probabilistic models I proposed for waveforms that could be useful in a source separation context, because they lead to new filtering procedures, new cost-functions for training their parameters or because they allow for signals defined on non-standard index sets.

However, this somewhat theoretical research on probabilistic models had to be balanced with a more applied research on **spectrogram models**, to be combined with all these filtering procedures. Without good estimates for the spectrograms of the sources, all these sophisticated signal models are indeed useless in practice. In terms of applications, the core difficulty lies in choosing a model that complies with any available *prior information* [LDDR13] and allows to estimate the sources α-PSD from the mixture α-spectrogram. I spent a significant part of my research and supervision effort on these aspects.

At first, I mostly used NMF methods from the state of the art, including the many variants that were specifically designed for music signals and that we reviewed in [J.Raf18]. Although I did some theoretical contributions on the matter [LFB15, J.Sim15], I can say I was mostly a user of such methods.
Then, I proposed an orignal kernel-based approach called Kernel Additive Modeling (KAM), that had some impact due to its ability to flexibly model spectrograms in a nonparametric way.
Finally, I embraced deep neural networks for modeling spectrograms due to their impressive superiority in accomplishing this task as compared to what I was doing before.

It turns out I also was the general chair for the international Signal Separation Evaluation Campaign (SiSEC) at that time, which I decided to strongly modernize so that it would be compliant with the recently proposed deep learning methodologies. This lead me to be strongly involved in community service, a work that actually proved quite rewarding.

1.2.1 Nonparametric spectrogram models

Median-based separation Once again, the starting point of my work on kernel local modeling was a very effective technique whose theoretical grounding seemed weak to me. In 2010, D. Fitzgerald proposed to separate harmonic sounds from percussive ones by simply applying running median filters on the mixture spectrogram along the time and frequency axis. The result was then used for Wiener filtering and was impressively effective. This idea is called harmonic-percussive source separation (HPSS) and is illustrated in Figure 1.18, where typical spectrograms for harmonic and percussive sounds are shown.

The following year, Z. Rafii proposed another method for the separation of singing voice from an accompaniment background that is assumed repetitive. The method was called Repeating

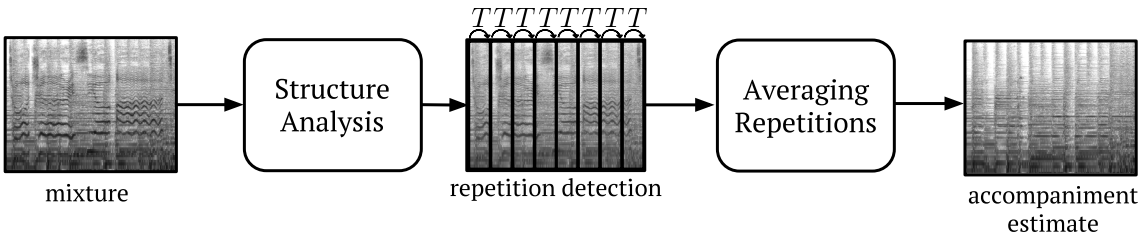


Figure 1.19: Separating singing voice from a repeating accompaniment. From [J.Raf18].

Pattern Extraction Technique (REPET) and is illustrated in Figure 1.19. In a first stage, the period of the repeating pattern is estimated, and then its spectrogram is taken as a median over all its repetitions. The approach was easily implemented and lead to remarkable performance boost as compared to NMF-based methods on the same signals. On the same heuristics grounds and with these new collaborators, I stepped in that line of research at that point and extended REPET to handle *locally repetitive* accompaniment in [LRB+12]. It notably allowed to process full-length tracks and obtain unprecedented separation quality.

As yet another median-filtering based separation method, Z. Rafii also proposed the so-called REPET-sim approach to estimate the accompaniment contribution for each time frame as a median over *frame-specific neighbours*, defined as the other frames within the same mixtures that have a similar spectrum. They are identified through some hand-picked similarity metric [B.Raf14].

In any case, even if these methods did offer very exciting performance, I was frustrated by their *ad-hoc* flavour: they didn't come with a principled framework that would allow their understanding or their extension to more challenging scenarios. It was obvious to me that all the median-filtering based techniques reviewed above did share some common features and were all particular cases of some more general framework that was still to be written down. This motivated me to investigate in many different directions. The closest established framework I could find and that raised my attention was **(nonparametric) local regression**, dating back to the 1980's and exemplified by the famous LOcally Estimated Scatterplot Smoothing (LOESS) method, that I shortly present now.

Local regression. Let $\mathcal{D} = \{l_n \in \mathbb{L}, z_n \in \mathbb{R}\}_{n=1,\dots,N}$ be our observed data. It consists in N signal values $z_n \in \mathbb{R}$ observed at locations $l_n \in \mathbb{L}$. In early applications, we typically have $\mathbb{L} = \mathbb{R}$. The objective of local regression is to infer the value of the signal at other locations. There are many ways to consider this problem, including GP as presented in Section 1.1.2, but local regression takes a different route. First, it picks some parametric model for functions:

$$F(\cdot \mid \theta) : l \in \mathbb{L} \mapsto F(l \mid \theta) \in \mathbb{R}, \qquad (1.44)$$

that can be as simple as, say, a linear model $F(\mathbf{l} \mid \{\mathbf{a}, b\}) = \mathbf{a}^\top \mathbf{l} + b$. As is classical, the value at a new location $l \in \mathbb{L}$ would then be approximated as $F(l \mid \theta)$ with parameters that need to be estimated.

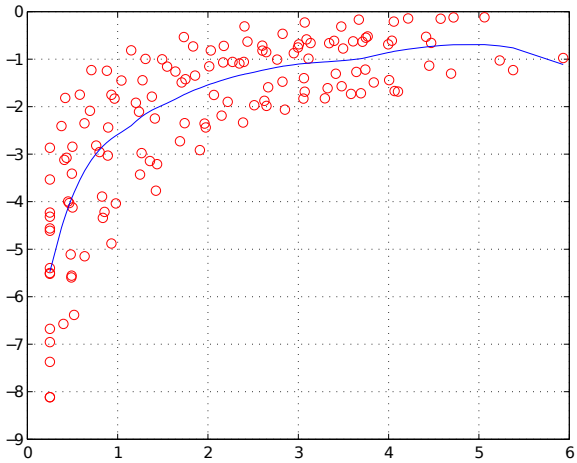


Figure 1.20: A typical example of LOESS smoothing of a scatterplot with a local linear model.

Then, the main idea of local regression is to avoid training a unique set of global parameters, that are assumed to hold for all locations. On the contrary, **each location $l \in \mathbb{L}$ comes with its own parameters** θ_l, so that the signal is estimated as $F(l, \theta_l)$. The rationale for this choice is that even if the data is quite sophisticated globally, it can be approximated by a simple model locally. I simply see this as some generalization for Taylor series expansion, where complex functions are locally approximated by polynomials.

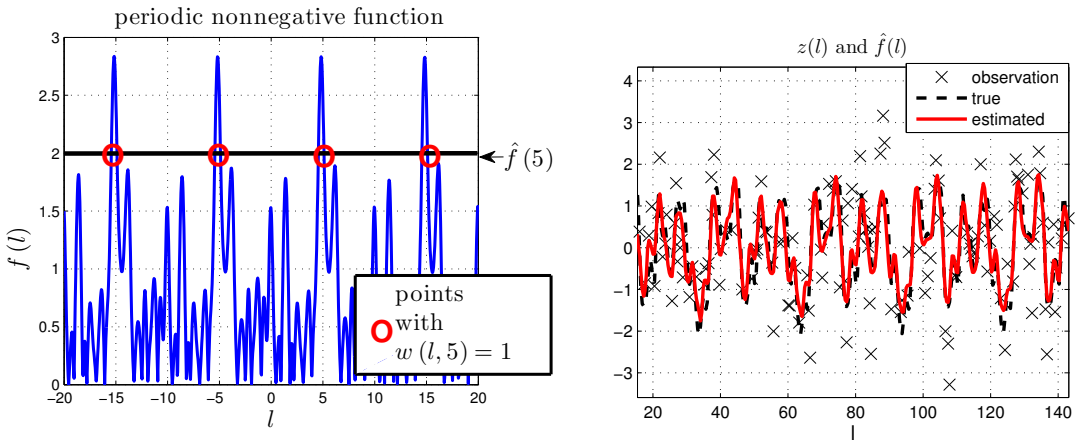


Figure 1.21: (left) A periodic proximity kernel allows to model arbitrary periodic signals as constant functions with periodic proximity. (right) An example of a locally periodic signal that is denoised with such a pseudo-periodic kernel. From [J.Liu14b]

With this idea of local modeling in mind, the main question becomes **the estimation of the local parameters** θ_l given the data $\mathcal{D}$. This is typically done by a weighted scheme, with data-points that are "close" to l having a stronger importance than points far from it:

$$\theta_l \leftarrow \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w(l_n, l) \left\| F(l_n \mid \theta) - z_n \right\|^2, \tag{1.45}$$

where the squared error can be replaced by any cost function that is deemed more appropriate and where $w(l_n, l) \geq 0$ is a *proximity kernel* that is high whenever z_n should play an important role when estimating θ_l, typically when l_n is close to l. A typical example is given in Figure 1.20.

Kernel additive modeling. Inspired by the local regression paradigm, I proposed the KAM framework, that puts all the median-based approach reviewed above under the same umbrella. It is detailed in [J.Liu14b], which is reproduced on page 136 and that I quickly summarize here.

The first core idea of KAM is to detach the proximity kernel $w(l_n, l)$ in (1.45) from the actual $\mathbb{L}$-distance $\|l - l_n\|$, departing from what is classical in the local regression literature. The idea is exemplified in Figure 1.21 (left), where we see that a strongly varying but periodic signal can be considered constant, provided the notion of proximity itself becomes periodic. As we show in [J.Liu14b], this twist allows to denoise signals simply even under very adverse noise conditions. This is made possible by additionally using cost-functions like absolute error $\mid F(l_n \mid \theta_l) - l_n \mid$ in (1.45), which are minimized with median filters.

The second idea I used was to generalized the **backfitting algorithm** that I found in the old *Generalized additive models* book by Hastie and Tibshirani. It allows to decompose the observation as the sum of J components defined nonparametrically through their proximity kernels w_j. I coined the resulting algorithm "Kernel backfitting" and it was good enough to generalize to many cases of interest that we described.

Equipped with the KAM framework, it was easy to **formalize and generalize all the median-based separation methods**, corresponding to different choices for the kernels as depicted in Figure 1.22. The kernel for REPET-sim could additionally be understood as the correlation between the mixture at frame t and t': $w(t', t) = \langle \boldsymbol{x}_t, \boldsymbol{x}_{t'} \rangle / \sqrt{\|\boldsymbol{x}_t\| \|\boldsymbol{x}_{t'}\|}$. The initial studies [J.Liu14b, LRP+14] came with a model with the four sources depicted in Figure 1.22 and I believe we can say it settled a new state of the art for two years in the domain.

KAM follow-ups. Although introducing the KAM framework was an achievement of its own in my opinion, it also came with a host of new research directions that I was happy to investigate with collaborators and students in the following years, notably thanks to the "Jeune Chercheur" grant I

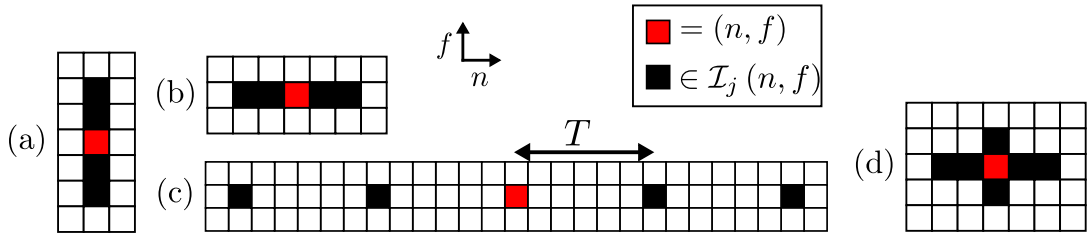


Figure 1.22: Examples of kernel to model audio spectrograms. (a) Percussive sound. (b) Harmonic sounds. (c) Repetitive sounds. (d) Locally continuous. From [J.Liu14b]

obtained for this purpose by the French national agency for research. We used KAM for a refined version of the HPSS method, allowing for iterative and multichannel separation [FLR+14]. We also considered exploiting user input to help regarding the design of the proximity kernels in [RLP15] and we found out that KAM could also be useful for interference reduction, which consists in reducing the *leakage* of one source to all microphones in live music recordings [PBLM15, DCLD18]. We also applied it for dialogue-background separation in [KLC15].
Finally, we addressed the scalability issues of KAM in [LFR15], which stem out of the fact that the whole source spectrograms must be kept into memory during separation, owing to the nonparametric nature of the model. This causes memory practical issues for large files. The contribution on this matter was to compress the parameters through randomized singular value decompositions (SVD), to significantly reduce the memory footprint of the method at virtually no cost in terms of performance.

1.2.2 Audio modeling with neural networks

As a researcher in probabilistic signal processing with an expertise in audio source separation, I could describe my daily activity as proposing new *models* for audio signals that would make sense for various physical or psychoacoustic reasons, and then to propose new methods to estimate the parameters of these models in light of actual mixtures, leading to original separation techniques.

However, **deep learning** provoked a profound paradigm shift in my research field in just a few years. Researchers have been proposing source separation methods exploiting deep neural networks (DNN) as early as 2014[11] and although the resulting methods were initially quite far from beating KAM in terms of robustness, scalability and sheer performance, I rapidly felt that they were about to become the *de facto* standard in the domain. On this matter, I really must acknowledge here the particularly accurate foresight of my colleague E. Vincent, that proposed me back in 2014 to co-supervise the Ph.D. of A. Nugraha on the topic of deep learning for source separation, which really triggered a new phase of my research career.

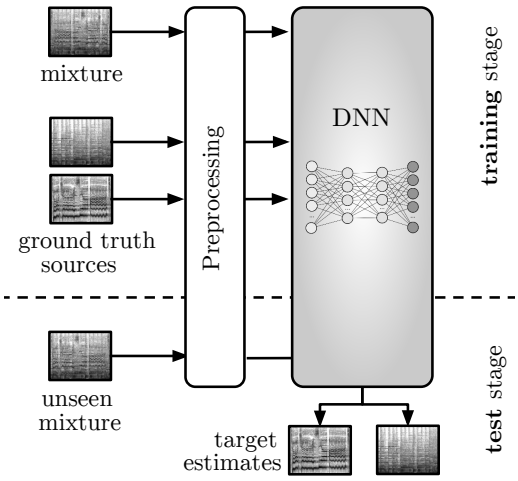


Figure 1.23: Supervised training of a DNN-based separation method. From [J.Raf18]

Supervised training for source separation. Sticking to the established probabilistic frameworks mentioned before, the objective of DNN was long taken as inputing the (magnitude) spectrogram from the mixture and outputing an estimate for the (magnitude) spectrogram of the sources.[12] Once those estimates are given by the DNN, we may combine them directly with the mixture phase or use them to construct a Wiener filter (1.4).

[11] As far as I know, *Singing-voice separation from monaural recordings using deep recurrent neural networks* by P.-S. Huang *et al.* is the first DNN-based method proposed for vocals extraction in music.

[12] A rigorous treatment would say we produce an estimate for the (magnitude) *spectral density* of the source, but the word ”spectrogram” is widely used (incorrectly) in this context instead.

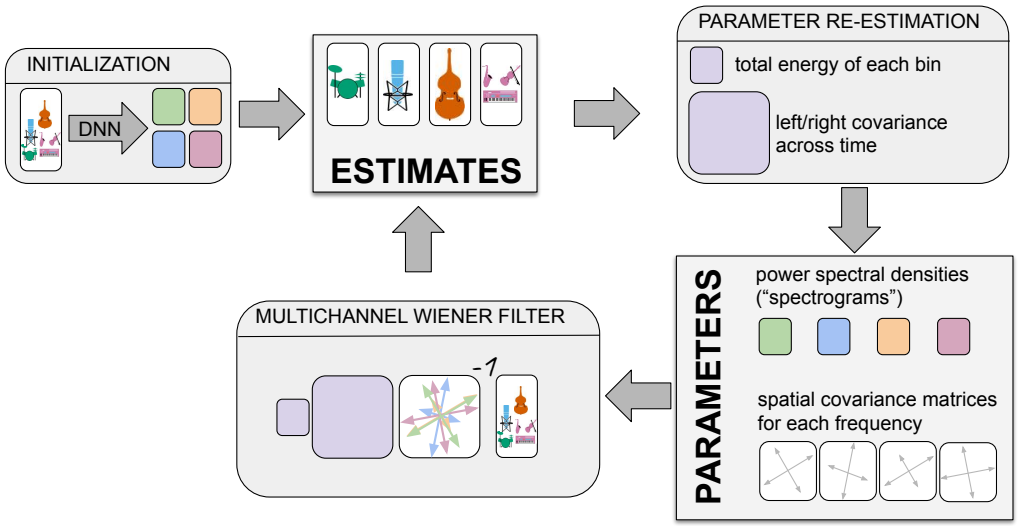


Figure 1.24: The audio separation system in [J.Sto19] is the combines a first phase of DNN-based separation followed by an EM algorithm for refining the parameters, notably for better stereo consistency.

From a general perspective, the general workflow for training a supervised DNN method is given in Figure 1.23. In a first **training stage**, we leverage available training data (mixtures + true separated sources) to learn a mapping between the mix and target spectrogram. Skipping the details, this mapping is taken as a deep neural net, hence as a parametric nonlinear function comprising millions of parameters. Once its training is finished, we may use the model during the second stage called **testing** or *inference*. A new unknown mixture is provided to the model that runs with the parameters identified during training kept fixed, and its outputs serve as an estimate for the targets. Although training can take from a few days up to a few weeks, inference is actually very fast and can typically be done in real time.

As can be seen, the main feature of this approach is to be *data-driven*: it leverages training data to devise effective separation methods whereas previous work mostly relied on (appropriate but limited) *intuitions*. From the perspective of 2015, this new methodology raises several interesting questions. First, **what kind of network** is appropriate for modeling audio spectrograms? Second, how could we **combine deep learning with signal processing** to incorporate DNN into a digital signal processing (DSP) pipeline? Third, how should we train the models?
In the following few years, I spent a significant amount of my time along with my students and collaborators trying to answer those questions.

Deep neural network models. As of the current day, there are new deep architectures that are proposed for source separation every week and it is largely out of the scope of this document to present them. However, when A. Nugraha started his Ph.D. under my co-supervision, state of the art was quite limited and basically consisted in a recurrent neural network inputting each frame of the mix spectrogram at a time. Just like classical Markov models, this suffered from a **limited ability to exploit long term dependencies**. For this reason, we introduced a feedforward multilayer perceptron (MLP), that **simply inputs whole chunks of the mixture spectrogram at a time** to predict the central frame for the target. It comprised three layers, yielding millions of parameters already.

Combining DNN with DSP. Until very recently, most DNN architectures that were proposed for source separation handled (magnitude) spectrograms as input/output, leaving the rest as pre-post processing. This fits nicely in the classical scheme I presented above for Gaussian-based separation, and the first system we proposed in [J.Nug16] is an iterative method were each iteration comprises a first step of *DNN-based spectrogram denoising* (partly replacing the Maximization-step) followed by a regular multichannel filtering procedure, as an Expectation step. It required the

training of one network per source and per iteration. This combination of DNN and DSP appears to me as a nice achievement. It replaced KAM as the best performing method for music separation for 2 years and [J.Nug16] stands as my most impactful paper today with nearly 250 citations. It is reproduced on page 149.

Later on, we proposed a more effective weighted variant in [NLV16, B.Nug18] for re-estimating the spatial covariance matrix, and my latest work on the matter [J.Sto19] considers the slightly simplified architecture depicted on Figure 1.24, where DNN are only used once to produce the initial estimates (by just combining their output with the mixture phase), before these are fed in a classical EM algorithm.

Does architecture matter? Even if the model we used in [J.Nug16] is particularly simple, it proved remarkably effective and it took a long while for competing systems to really outperform it by a large margin.[13] This fact actually had me believe for some time that the particular architecture used for DNN-based modeling did not have any real importance nor was it a promising track for research, at least for the time being. We even observed in [SLI18] that most DNN-based systems at that time did not show significantly different performance. Of course, I was proved wrong later on when breakthroughs finally popped in, notably with end-to-end methods. In retrospect, my belief is that architecture comes third, after doing the DSP right[14] and training the model correctly.

Training the model. It was widely believed early-on that the actual cost-function used for training a DNN would be important. I guess the reason I clung to this idea initially was more aesthetic than practical: I believed it was the way the different probabilistic models I was working on could get into the picture. In [J.Nug16], a significant part of the experiments concerned trying out many different cost functions, and I must say I soon realized that this kind of research was not going to be so exciting. These days, everyone is just training their models through MMSE with amazing performance.
On the way, I realized that other aspects were way more important, including other network architectures, regularization schemes, learning-rate scheduling, batching, speed and parallelization, etc.

DNN follow-ups. In the following years, I had some collaborations regarding new deep-learning models for signal separation. This includes combining a deep auto-encoder model for the speech signal with a "garbage" α-stable model for noise. This idea was first proposed in [LSL$^+$19] in the single channel case, and then in [FNB$^+$19] in a multichannel Cauchy context, where an additional NMF model came in for the noise for more flexibility. I think these studies nicely exemplify what probabilistic signal processing has finally become after several years of cohabitation with deep learning. In essence, the dust has settled down and deep-neural nets play the role of nonlinear parametric models, only with unprecedented modeling capabilities, coming at the cost of long training times.

1.2.3 Community service

As described above, the first phase of my (our) research on deep learning for music separation involved proposing and training new separation systems that would perform well, like I was doing before but exploiting this new training-based methodology. Still, I rapidly felt that this way of proceeding would not allow me to grasp the big picture as much as I wanted to, notably because experience showed that **it takes a lot of work to have one single model working sufficiently well for publication**.
For this reason, I decided to take another route and to delve into **community service** so that I could be playing a role in the big picture while lacking the large team it would require to make many advances a year on the topic. The fact is that deep learning for music processing was a

[13] Two years later, our good old MLP was still better than newly proposed methods leveraging fancy autoregressive *wavenet* variants...

[14] A nice paper I reviewed was obtaining bad performance just because the overlap-add procedure was not doing perfect reconstruction.

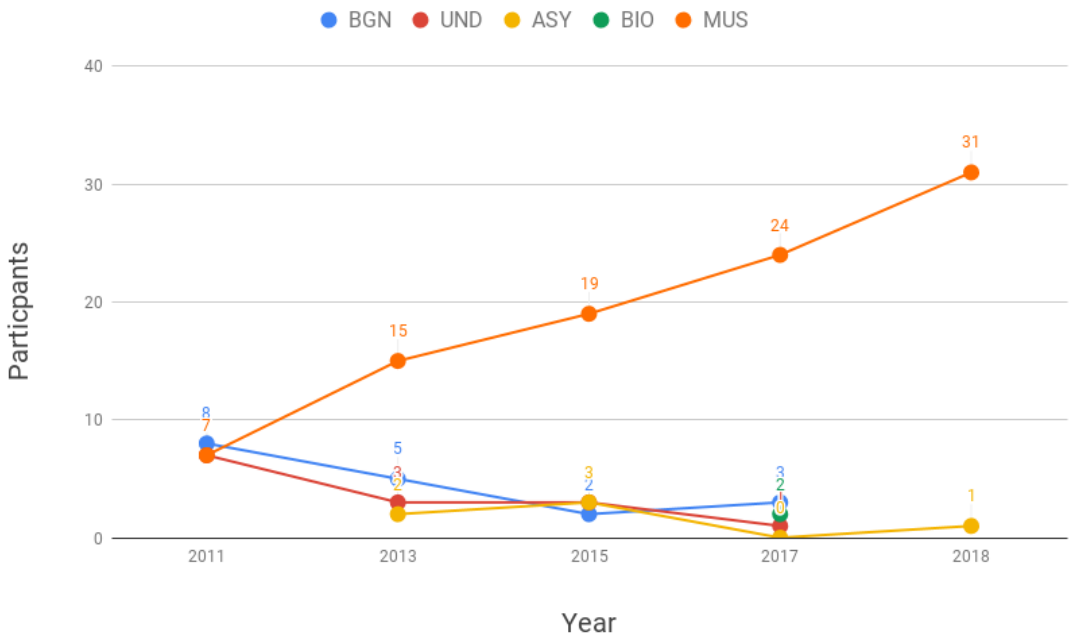


Figure 1.25: Number of participants of SiSEC over the years for each task. I was involved in MUS starting from 2015.

very recent topic and attracted a lot of interest and momentum. However, it lacked a community and all that comes with it: data, software, benchmark, etc.

It turns out that, back in 2014, N. Ono offered me the opportunity to become the general chair of the international **Signal Separation Evaluation Campaign (SiSEC)**, which is organized every 18 months since 2007 as a satellite event for the LVA-ICA conference.[15] I gladly accepted the challenge, because I felt it would be the perfect canvas for my plans.

Challenges and roadmap. Although I could not participate in the first two occurrences of SiSEC because I was working in the industry and that possibly not proposing the "best system" was deemed too risky by the management, I had closely been following this international evaluation campaign since its beginning in 2007, when it was initiated by E. Vincent. I could get involved as soon as I joined academia in 2010.
The objective of this evaluation campaign is to compare existing methods for source separation, based on voluntary submission of separation results by researchers. It was targeted at any source separation applications, including speech, music, but also biological signals. However, as can be seen of Figure 1.25, only the **music demixing task** (MUS) was really getting attention, at least in regard to the modest size of the music separation community.

Due to the taking over of deep learning, the community was at a turning point. Since it was calibrated for evaluating music demixing systems in the "model-based" era, SiSEC was not at all appropriate for deep learning research and was on the verge to simply collapse, for several reasons that became the different directions for a large part of my research effort in the recent years.
This work clearly does not consist in so many theoretical contributions, but rather in setting up an appropriate canvas for others to do research properly. Still, I am happy with the impact it has and I would say that such a sustained *community service* is definitely rewarding in the long run. I really must acknowledge here that I learned a lot from my colleague F.-R. Stöter on all these matters, who I feel has a very clear vision of what modern reproducible research should be.

[15] I just spent some time finding out that the we could write: SiSEC is a sesquiennial event...

Lack of data. Before the advent of deep learning, data was not necessary for the design and training of most source separation methods, but only for their evaluation. Up to 2013, the dataset considered by SiSEC for this purpose comprised excerpts of up to 10 tracks, which was already considered a big step forward compared to the 2mn30 total duration of the dataset it used before. I identified two problems with this very limited amount of data.
First, even before deep learning was there, I was feeling that we were **strongly overfitting** over the few tracks from SiSEC. One remarkable aspect of KAM in my opinion was indeed its *robustness* as compared to the state of the art. I could not show this aspect with only the few test tracks from SiSEC, that had actually been chosen because they were easily separated. For this reason, I spent some time preparing the `ccmixter` dataset [S.Liu14], which consists in 50 full length tracks comprising separated vocals and accompaniment and that I introduced in [J.Liu14b, LFR15]. As opposed to the only other "large" dataset that was available at the time (MIR1K), its main advantages were to feature full-length and realistic tracks.
Second, as soon as learning-based methods became dominant, I was thinking that if nothing was done to equip the community with a decent public dataset for training and evaluation, there would be two unfortunate consequences: SiSEC would die in the very short term and advances on the matter would be driven by private datasets and non-reproducible research.
Experience showed that relying on professional sound engineers for preparing a music dataset was not scalable because it was way too costly in regard of our limited academia budgets. However, it turns out that S. Mimilakis and myself are amateur music producers, so that we decided it was worth it to spend the time it would take to prepare it ourselves. Z. Rafii had already downloaded the separated stems for 100 tracks from the extraordinary *Mixing secrets dataset* resource,[16] that served as a starting point. For each song, it comprised from 5 to 60 separated stems. The desiderata in preparing our new dataset were the following:

- **Realistic mixing.** The songs should not be mixed with any source separation assumptions in mind, or involve simplified and nonrealistic methods. For preparing them, we only considered artistic considerations and used professional software only, including highly nonlinear tools.
- **Usable.** The core objective of preparing the dataset was to have *ground-truth* data for training and evaluating music separation systems. Even if we used as much nonlinear processing and effects like distortion, reverberation and compression as needed to prepare the (stereo) sources, we avoided *mastering*: the sources had to sum up to the mix. This introduced some very particular and difficult constraints to still maintain mixtures that would be realistic.
- **Systematic.** I wanted to create a dataset that would be usable at scale in practice, and that would go beyond just vocals/accompaniment separation as was classical at that time. For this reason, we decided to split the stems arbitrarily into 4 groups: vocals, bass, drums and "other". Although it is arguable in many ways and is regularly (rightfully) criticized, this simplified scenario has the benefit of promoting research on the separation of music into many other sources than just vocals.

After approximately 2 months of intense work later, we came out with the DSD100 (demixing secrets dataset), that we could use for SiSEC 2016 [LSR+17]. After augmenting it with 50 additional tracks that notably came from Medley DB,[17] it became the MUSDB18 dataset [S.Raf17]. I can say the effort was definitely worth it, since this dataset stands today as the *de facto* standard for benchmarking research in music filtering. It has been downloaded approximately 10 000 times.

Lack of software. Along with deep learning came another important change, which is software. Just like most other researchers in the topic I switched from MATLAB to Python. It turns out I did the change early on in 2010, and I already had time to program some commonly used routines in this new language: STFT and inverse, multichannel Wiener filtering, randomized SVD, NMF/NTF, etc.
I was fully sharing the view of my colleague F.-R. Stöter that there was room to provide the community with a new set of public tools for music separation, which would serve a double objective: i/ make it easy for a machine learning researcher to do music processing. ii/ make it easy for a researcher in music processing to do machine learning.

[16] https://www.cambridge-mt.com/ms/mtk.
[17] https://medleydb.weebly.com/

The http://sigsep.github.io website and related github community serve as the landing point for most of the open-source software we released for music separation. It gathers several kinds of resources.

- **Data and postprocessing tools.** I believe that one key to the success of MUSDB18 is the fact that it comes with a mature Python package to manipulate it easily: `musdb` [S.Sto19a], which is readily installed through `pip`, the Python package manager. It allows a very pythonic manipulation of the dataset. We also released `norbert` [S.Liu19], a Python implementation for multichannel Wiener filters, for the interested researchers to finally have a tested implementation for this subtle procedure that is so detrimental to performance when done wrong.
- **Evaluation tools.** The quality of source separation is measured through objective metrics that can basically be understood as some kind of signal to error ratios computed by comparing the true sources with the estimates. Among them, BSSeval metrics stand out as the *de facto* standard in the community, although it is widely acknowledged that there are better alternatives [SRG+16]. The official implementation of these metrics used to be in MATLAB and I took the responsability for their Python implementation, as the general chair of SiSEC at the time. This resulted in `museval` [S.Sto19b], which stands as BSSeval v4. It features some particular tricks to significantly speed up evaluations, as presented in [SLI18].

open-unmix. Experience recently showed to me that the line between research and engineering is sometimes quite blurry, especially when it comes to deep learning. The software presented above probably rather stands on the ”engineering” side. Still, we also worked hard on a development project that really took more than a year, and that I would definitely call research: along with colleagues from Sony research, we implemented an open-source (MIT-licensed) baseline implementation for music separation, named `open-unmix`.
With model-based method, there were established implementations that could be used freely by the whole community. At its beginnings, the research of DNN-based music separation was really not reproducible: neither the implementations nor the weights of the networks were available to anyone, and the training data was also private. This made it virtually impossible to assess any improvement. As mentioned earlier, it was indeed quite likely that up to several dB of performance would be lost or gained through some pre/post processing.
For all these reasons, we aimed at `open-unmix` being a fully open-source separation model along with its processing pipeline. The desiderata were the following: i/ It should be trained on MUSDB18 only. ii/ It should not feature anything too exotic but only standard baseline models (we picked a simple Bidirectional Long Short Term Memory network). iii/ It should reach state of the art performance, so that anyone proposing a new method should basically beat it to prove a contribution. iv/ Its weights should be freely available to allow anyone to just test the model off-the-shelf.
Putting all these desiderata together actually proved a very demanding effort but we were eventually proud to release `open-unmix` [J.Sto19]. It is available as a `pip` package, and has already met an honorable audience.

SiSEC workflow. Before I became the general chair of SiSEC, evaluation results were sent over the internet through mail to the organizers (usually under non-consistent filenames), who painfully evaluated each one of them. Although it was enough for its purpose at the time, this workflow clearly had to be improved and take benefit from automatized methodologies.

For SiSEC 2015 and 2016, I already implemented scripts that would allow the participants themselves to run all their evaluations, so that we would only be left to analyze the results [ORK+15, LSR+17].
For SiSEC 2018, we completely changed the way to proceed and participants had now to open a pull-request on the evaluation git repository to do a submission [SLI18]. We reached an unprecedented number of participants that year and organizing this campaign was a very stimulating collaboration with researchers worldwide. The scores obtained by all those participants are now hardcoded into `museval`, so that anyone working on the topic can automatically compare his work to state of the art without having to rerun any of those experiments.

In Figure 1.26, I display the performance of the best system in SiSEC along the years, during the period where I chaired it. As can be seen, music separation witnessed an incredible jump of

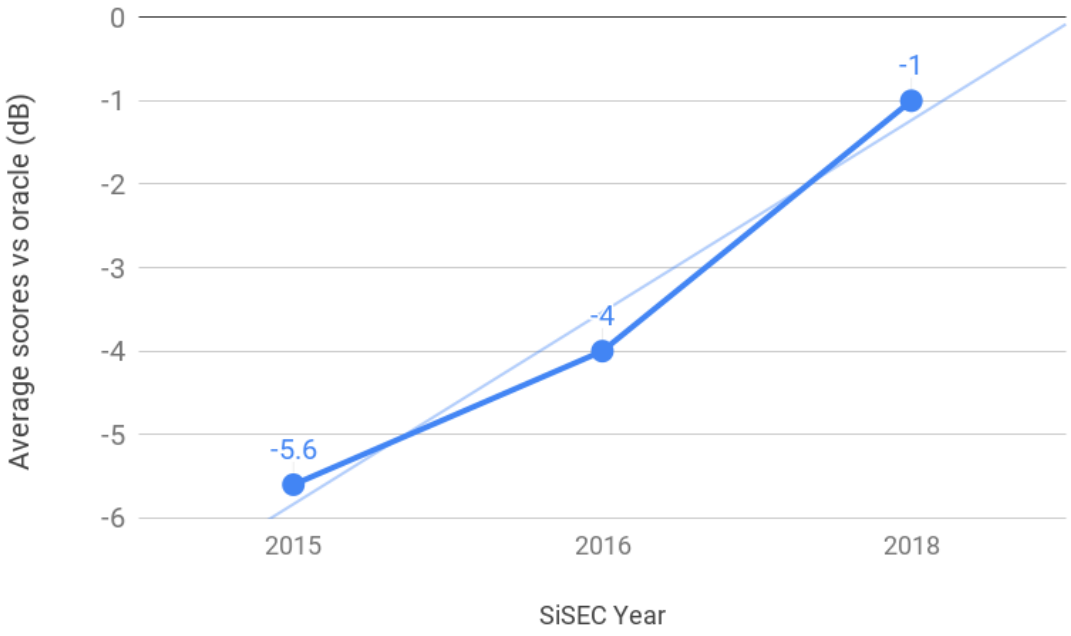


Figure 1.26: Performance of the best system evaluated in SiSEC along the years. Higher is better, and the reference 0 dB score is set as the performance of Wiener filtering with the true sources spectrograms.

5 dB performance in the meantime. We actually showed that the best systems we had in 2018 were not behaving significantly differently than oracle Wiener filtering, which was considered a glass-ceiling.

After all these years being involved in helping music separation become a machine learning task, I felt my time as a general chair for SiSEC was over, and I handed the flag to Y. Mitsufuji from Sony Corporation in 2019, that will probably continue expanding its impact even further.

1.2.4 Outreach

The last part of the work I want to mention about community service is my effort in widening the audience of source separation.

Scientific outreach. Along with students and with other colleagues, we created a lot of teaching material that is freely available on http://sigsep.github.io and that was the basis for two conference tutorials (ISMIR'18 and EUSIPCO'19), as well as one invited overview paper [J.Raf18] and a book chapter [J.Can19]. I was also involved in a journal paper in the french scientific outreach journal *interstices* [J.Liu16].

Transfer to the industry. Although I cannot describe this due to non-disclosure agreements, I can mention that I am the co-author of a software called UMX-PRO that has been transferred to a north-american company for several hundred thousand euros. This software is a complete solution for audio source separation. All other details regarding this software transfer are confidential.

1.3 Signal processing for machine learning

The overwhelming majority of the research I presented above concerns music signal processing and separation. This stems out of the fact that I am a passionate about music production and always felt that music demixing would be a key technology to revise the way we think about music creation and usage [J.Car19], and would unlock a new phase for intellectual property in arts.[18]

This said, my personal interest as a researcher was never really limited to that particular application and my scientific motivation always rather was in looking for better theoretical foundations for the practices I met and didn't understand. On the way, I got interested in many different disciplines, and I was always glad to start a collaboration on topics other than audio whenever I had the opportunity.

In this section, I quickly review my activity in those domains where audio was not the core focus. These works all share several features. First, they all comprise rather strong experimental *and* theoretical parts. Second, as all of my other studies, they were done in collaboration: depending on the context, I was either the theoretical fellow, or the one rather doing the experiments. Third, they involve probability theory applied to signal processing or machine learning.

1.3.1 Robust matrix factorization methods

As I mentioned above in Section 1.1, I got involved in the use of NMF for source separation, notably for applications in audio-coding. However, once the α-stable models had shown some interesting properties, it became natural to some colleagues and students to study how they could be used for deriving **new matrix factorization algorithms**.

In this line of research, I already mentioned the Cauchy NMF study that I coordinated myself [LFB15]. However, two other cases are worth mentioning here. The first one is the Levy NMF [MBL17a, MBL17b], proposed by P. Magron, which appears as a nice way to do source separation of nonnegative random variables with perfect reconstruction performance. Indeed, just like Gaussian or Cauchy approaches enforce nonnegative energetic models over (complex or real) observations, the Levy approach also estimates an energy-based model, but it differs in the fact that the observations themselves are also nonnegative. This notably lead to yet another generalization of the Wiener filter applicable to nonnegative observations.

As a very general contribution in [SEL$^+$18, J.Sim15], U. Simsekli proposed to exploit the conditional Gaussianity (1.35) to estimate parameters for α-stable NMF models with any characteristic exponent α. This is done through an MCMC-based algorithm. Interestingly, the approach is powerful enough to also estimate the actual exponent $\alpha \in (0\,2]$ to use, providing one principled answer to this natural question.

When inspecting the different sources of citations obtained by these rather theoretical papers, I was glad to see the methods we proposed attracted some interest outside of the audio processing community, notably in data-analytics contexts. Even if I've not been using factorization methods too much in audio recently, I still think they are very attractive in many respects, notably thanks to their ability to extract redundant patterns from possibly very noisy data and to be applicable in cases where training data is very scarce, which is the case in many applications.

1.3.2 Sparsity: hardware compressed sensing

After I defended my Ph.D in audio coding and source separation, I wanted to investigate applications that would be very different from audio, to get a grasp of the use of signal processing in other domains. I was very lucky that L. Daudet offered me the opportunity to join him at Institut Langevin to work on **hardware compressed sensing**, which was a totally new and disruptive idea at the time. I feel privileged that I could lead the data-analytics aspects of what was to become the very exciting adventure of optical computers.[19] In this section, I briefly summarize my work on this aspect, that is described in full details in [J.Liu14a], that I reproduced on page 162.

[18] As an amateur musician, I used to be pretty involved in the open-license movement, promoting *copyleft* licenses like `Creative Commons` from 2005 onwards.

[19] L. Daudet is now CTO at lighton: `http://lighton.io`.

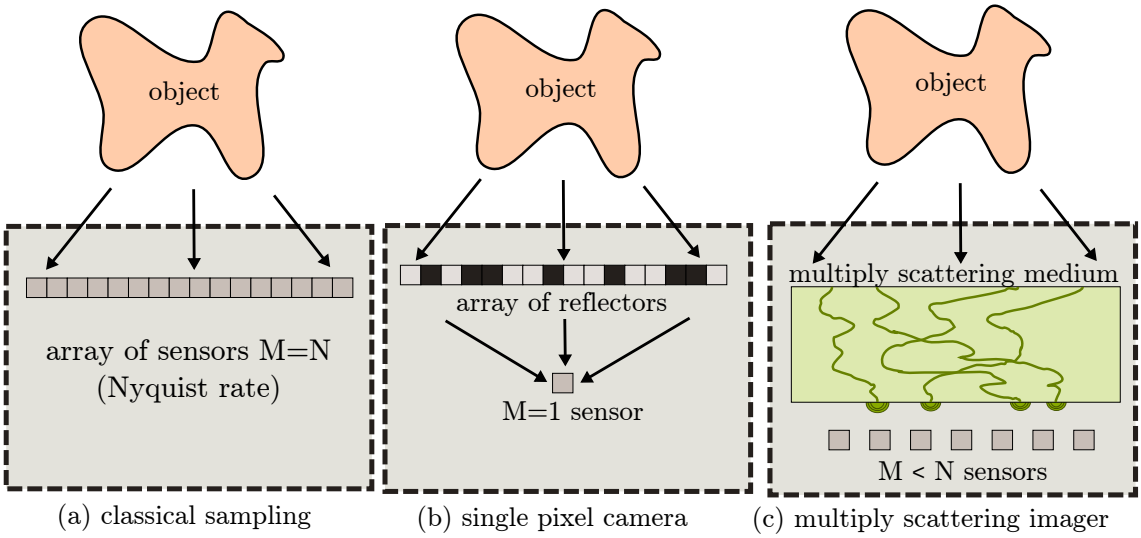


Figure 1.27: General idea for [J.Liu14a]. Classical sampling takes regular local snapshots. The single pixel camera takes random projections with a human-crafted array of reflectors, one at a time. The proposed method just lets light originating from the object go through a scattering material, that acts as an analog random mixer.

The natural randomness of optics of turbid media In my view, the most remarkable aspect of the year I spent at Institut Langevin was to work daily with very talented physicists whose research topics were extremely different from mine. This created a very stimulating environment where probabilistic signal processing was as a very exotic solution to problems raised by very non-classical optics.

In any case, my understanding of the physics remained limited. For all practical purpose, it was sufficient for me to understand the few basic principles that are illustrated on Figure 1.27. The objective was to create an **imaging device** (a camera) that is able to capture an input signal with only a very small number of sensors. For this purpose, the groundbreaking idea was to depart from traditional Shannon sampling theory where many samples are regularly taken along the signal to leverage *compressed sensing* instead, where each sample would contain some global information from all parts of the object. Although this idea had already been proposed previously, it was only through the carefully designed pseudo-random mixtures of the *single pixel camera*. Here, the mixing was achieved at the speed of light by simply having the signal go through a thin layer of **multiply scattering material**. As far as I understand, this could be any kind of physical *non-transparent* material like a thin layer of white paint, except that it should ideally remain still at a microscopic level for some time and not be *opaque*, i.e. it should let the signal go through! More details and references are given in [J.Liu14a].

From a pragmatic data-analytics perspective, the whole experimental pipeline could be summarized in the following way. Let $\mathbf{x} \in \mathbb{C}^N$ be the signal to capture. It corresponds to some complex valued optical wave-front. When this light wave goes through a scattering material, it undergoes a very large number of diffractions due to the chaotic nature of the material. If we mesure the wave front at some output, say y_m, physics guarantee we can model it as $y_m = \sum_{n=1}^{N} h_{mn} x_n$, so that gathering M such measures together as a vector $\mathbf{y}$, we get:

$$\mathbf{y} = \mathbf{H}\mathbf{x}, \tag{1.46}$$

where $\mathbf{H} \equiv [h_{mn} \in \mathbb{C}]_{mn}$ is called the *transmission matrix* (TM) and characterizes the scattering material, just like focal length characterizes a perfect lens. A crucial fact about the setup is that the entries for the TM for a strongly scattering material can very adequately be modeled as i.i.d Gaussian.

Experimentally, it turns out that the acquisition process is way more complicated than (1.46), because it's rather $|\ \mathbf{y}\ |^2$ that is observed in practice. However, obtaining linear observations was made possible by clever optical considerations.

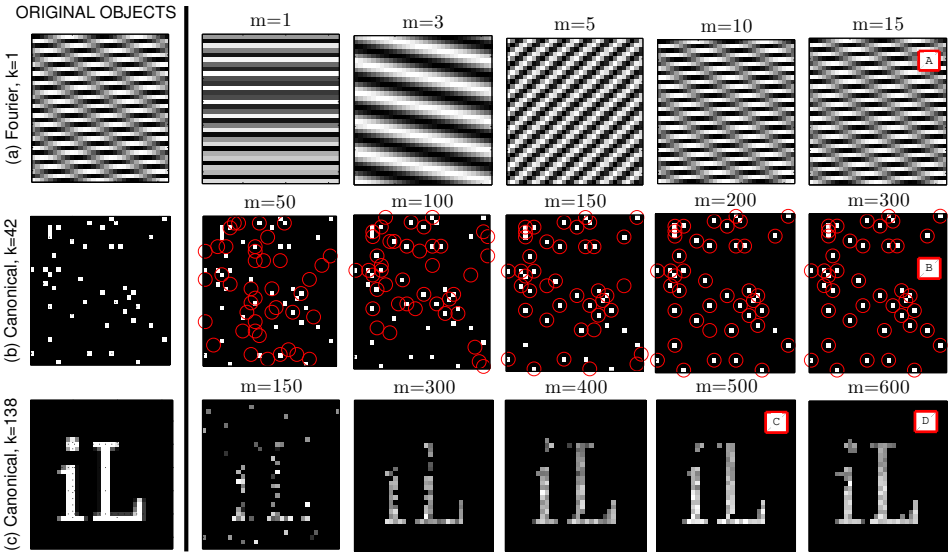


Figure 1.28: Examples of reconstruction of signals ($N = 1024$) that went through a real multiply-scattering material. k is the sparsity of the signals (number of nonzero coefficients required to describe them) while m is the number of sensors used for acquisition. Labels A, B, C and D correspond to the position of these examples within the plot in Figure 1.29.

Proof of concept for physical compressed sensing Due to his strong involvement in the community of *sparsity-based signal processing*, it occurred to L. Daudet that the acquisition process (1.46) actually behaves like **a hardware implementation for compressed sensing**. To summarize this whole branch of information theory in one sentence, I can say it guarantees perfect reconstruction of a wide family of signals, provided they are acquired through a sensing matrix like in (1.46) that is sufficiently random. For one year, I worked hard to show that scattering materials could be considered as good candidates to perform such sensing, at the speed of light.

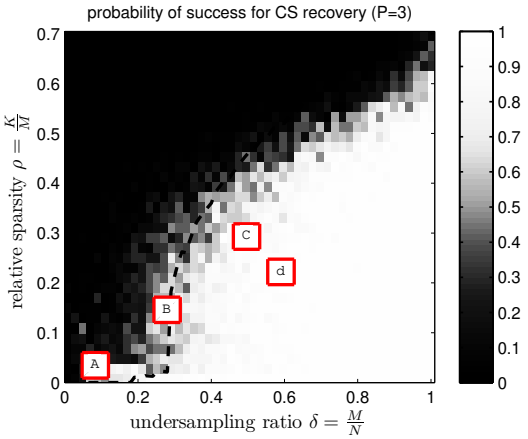


Figure 1.29: Experimental phase transition between failure and success of signal reconstruction, as a function of the sampling ratio and relative sparsity.

The results for this experimental study were numerous. First, we were happy and impressed to observe that the behaviour for our hardware sensing system was actually following some phase transition between perfect and totally failed reconstruction as a function of the number of sensors, as predicted by the theory. This is illustrated in Figure 1.29 and some examples of reconstructions are given on Figure 1.28. This study was the first to show that compressed sensing could be implemented at the speed of light, and I must say I am positively surprised by the impact it had.

Second, obtaining the good results reported above involved several crucial aspects that gave rise to separate studies with various collaborators and students. Actually estimating the Transmission Matrix was a challenge of its own and led me to have several collaborations [J.Dre15, J.Liu15]. Then, it turned out that exploiting *multiple measurements* was a key aspect, even if each one involved only a very limited number of sensors [LMGD14].

After I left Institut Langevin to work with Inria, the adventure continued and optical computing is now a reality that can be leveraged to train large-scale deep learning models.

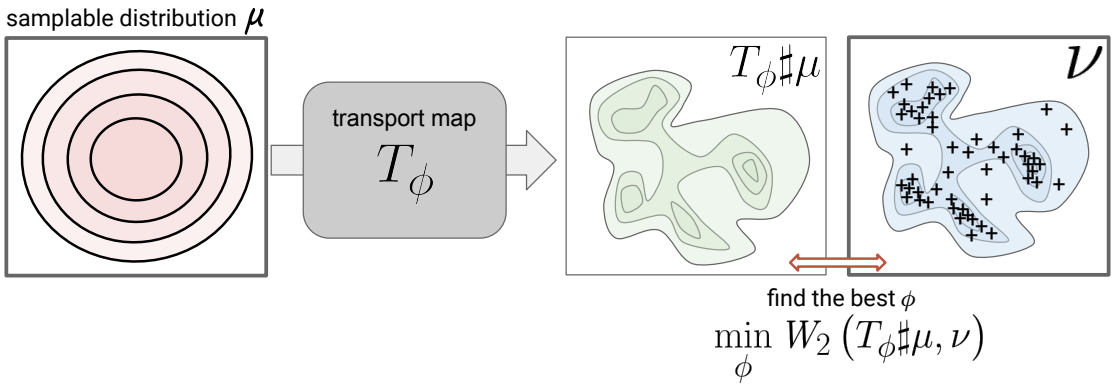


Figure 1.30: Optimal transport as an appropriate framework for generative modeling. The usual pipeline consists in applying some parametric transport map T_Θ (a neural network) on samples from a *source distribution* μ so that the resulting *pushforward* measure $T_\Theta \sharp \mu$ is as close as possible to the *target* ν in some sense.

1.3.3 Probabilistic methods for generative models

As I will detail better in the next chapter focusing on my research programme, I recently got interested in generative models, because I am feeling they are the key to go beyond the *filtering paradigm* I have been considering before. Although I already was involved in collaborations about deep generative models [FNB+19, LSL+19], the topic was largely new to me when it comes to actually *reconstruct* missing parts of signals realistically.

Extensively reading about DNN-based generative modeling and trying out my several options, it seemed clear that the much acclaimed *Generative Adversarial Networks* (GAN) by I. Goodfellow *et al.* looked like an inviting way to go. Still, I felt I really needed to better understand their theory before having a real clue of what to do with them. After some reading in many directions, the *Wasserstein GAN* from M. Arjovsky *et al.* caught my attention back in 2017, because it was the first time I felt there would be some solid theoretical framework that I would like to delve into and that would be an appropriate foundation for such generative models, and it was **optimal transport**. In the following page, I will introduce this fascinating research topic, to motivate the research done in [LŞM+19], which is reproduced on page 172.

Optimal transport and generative modeling. From our perspective, generative modeling may be formalized as in Figure 1.30. Let $\Omega \subset \mathbb{R}^D$ be our sample space, i.e. the space on which our signals are defined. First, we draw input samples $\mathbf{x}_i$ from an easy *source distribution* $\mathbf{x}_i \sim \mu$ like a Gaussian. Then, we have them go through some processing, say a parameterized function T_Θ that is called a *generator*, to yield output samples $T_\Theta(\mathbf{x}_i)$ whose distribution is called the *pushforward distribution* and written $T_\Theta \sharp \mu$. T_Θ is also called the *transport map* from μ to $T_\Theta \sharp \mu$. The objective of training is to enforce that $T_\Theta \sharp \mu$ is as close as possible to some prescribed *target distribution* ν, which is generally not known in closed-form but rather through many examples $\mathbf{y}_i \sim \nu$.

At this point, the only missing part of the story is to choose an appropriate **cost function** to compare $T_\Theta \sharp \mu$ and ν. There are many candidates for this purpose like Kullback Leibler or Total Variation, but one point of the Wasserstein-GAN paper is to show they suffer from severe "gradient saturation": whatever the current set of parameters Θ, the value for the distance between $T_\Theta \sharp \mu$ and ν is most probably identically maximal, unless we are already very close to a solution, making training very difficult.

Instead, they show that we may use the *Wasserstein distance* for comparing probability measures, which is better behaved and was introduced by L. Kantorovich. Informally, it tells how hard it is to *transport samples* from μ to ν. If it is small, then μ and ν are very similar.
To quantify this, the classical route is to first introduce the concept of a *transport plan* γ, illustrated in Figure 1.31. It is a measure on $\Omega \times \Omega$ and we think of $d\gamma(\mathbf{x}, \mathbf{y})$ as the amount of mass transported from $\mathbf{x}$ to $\mathbf{y}$. It introduces the idea that we may transport samples stochastically, and is a relaxation

of the deterministic transport *map T* seen above that we can write as $(Id \times T)\sharp\mu$. By construction, a transport plan should satisfy: $\forall A \subset \Omega$, $\gamma(A \times \Omega) = \mu(A)$ and $\gamma(\Omega \times A) = \nu(A)$, i.e. the marginals of γ coincide with μ and ν: before we start transporting, we have μ, and after we're finished, we have ν.

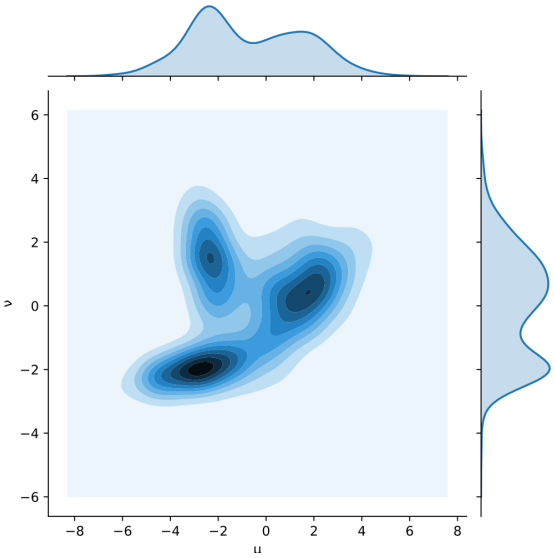


Figure 1.31: Two one-dimensional distributions μ and ν, plotted on the x and y axes, and one possible joint distribution that defines a transport plan between them. The joint distribution/transport plan is not unique. From Wikipedia, By Lambdabadger (CC BY-SA 4.0.).

Let now $\mathcal{C}(\mu, \nu)$ be the set of all such transport plans. For $p > 0$, the p-Wasserstein distance between μ and ν, written $\mathcal{W}_p(\mu, \nu)$ is defined as follows:

$$\mathcal{W}_p(\mu,\nu)^p \triangleq \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int_{\Omega\times\Omega} \|\mathbf{x}-\mathbf{y}\|^p d\gamma(\mathbf{x},\mathbf{y}), \tag{1.47}$$

I understand equation (1.47) in the following way. Each time we transport a sample $\mathbf{x} \sim \mu$ into a point $\mathbf{y}$, we penalize the move by cost $\|\mathbf{x}-\mathbf{y}\|^p$, because we ideally would not like to carry samples on long distances. The total loss induced by some plan $\gamma \in \mathcal{C}(\mu,\nu)$ thus becomes $\int_{\Omega\times\Omega} \|\mathbf{x}-\mathbf{y}\|^p d\gamma(\mathbf{x},\mathbf{y})$ and we define the actual distance between these distributions as the loss of the smarter plan. It can be shown to exist and to define a metric space over probability measures.

Equipped with the definition of the Wasserstein distance, generative learning may now be formulated as:

$$\Theta^\star \leftarrow \underset{\Theta}{\operatorname{argmin}}\, \mathcal{W}_p\left(T_\Theta\sharp\mu, \nu\right). \tag{1.48}$$

Whenever this loss function (1.48) is minimized, T_Θ is a transport map from μ to ν. Very fortunately, Y. Brenier proved that such a map exists and is unique in a very general setting (see [LŞM+19] and references therein). However, this existence theorem is not constructive, and we are left with the difficult problem of estimating it, which is highly non-trivial in the general case.

GAN theory: primal and dual optimal transport The *tour de force* due to Arjovsky is to make a connection between (1.48) and GAN training. For convenience, let $\mu(\Theta) \triangleq T_\Theta\sharp\mu$. It can be shown that the 1-Wasserstein metric (1.47) accepts a dual formulation, that reads:

$$\mathcal{W}_1\left(\mu(\Theta), \nu\right) = \sup_{\|f\|=1} \mathbb{E}_\nu\left[f\right] - \mathbb{E}_{\mu(\Theta)}\left[f\right], \tag{1.49}$$

where the supremum is taken over all the 1-Lipschitz functions $f : \Omega \to \mathbb{R}$. It turns out that this supremum is indeed attained by a function I write ψ, which is called the *Kantorovich potential* between μ and ν. Once this huge step has been made, it is straightforward to see that GAN training can be understood as alternatively estimating this Kantorovich potential, called the *discriminator*, and the transport map, called the *generator*, both being parameterized as DNN.

The sliced-Wasserstein distance. Just like me, a host of researchers suddenly got interested in optimal transport after reading the Wasserstein GAN paper. Getting into it was very difficult for me, because my initial background rather lies in computer science than mathematics. Still, I wondered whether we could not consider the primal formulation (1.48) for training a generative model. Reading the massive book *Optimal transport: old and new*, by C. Villani, I quickly realized that this would not be easy, because the Wasserstein distance appears as very complicated to derive, unless we are in the scalar case $\Omega = \mathbb{R}$.

In the scalar case, it indeed turns out that $\mathcal{W}_2(\mu,\nu)$ simply boils down to a **distance over quantiles**, as illustrated in Figure 1.32. Let F_μ be the cumulative distribution function for a probability measure μ. Its inverse $F_\mu^{-1}:[0,1]\to\mathbb{R}$ is called the *quantile* function. $F_\mu^{-1}(0)$ is the minimum value of $x\sim\mu$, $F_\mu^{-1}(1)$ its maximum, $F_\mu^{-1}(1/2)$ its median, etc. We have:

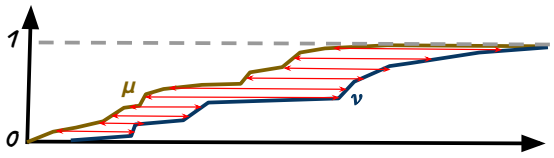


Figure 1.32: In the scalar case, the 2-Wasserstein distance is easily computed.

$$\Omega\subset\mathbb{R}\Rightarrow\forall(\mu,\nu),\mathcal{W}_2(\mu,\nu)^2=\int_{q=0}^{1}\left\|F_\mu^{-1}(q)-F_\nu^{-1}(q)\right\|^2dq. \tag{1.50}$$

Since we have a closed-form expression for $\mathcal{W}_2$, we can compute the optimal transport map as:

$$T(\mu\to\nu)\triangleq F_\nu^{-1}\circ F_\mu, \tag{1.51}$$

also known as the *increasing rearrangement*. It is easy to understand as mapping minimum to minimum, median to median, maximum to maximum, etc.

In practice, this particular scalar case is amenable to very straightforward computations. To compute $W^2(\mu,\nu)^2$, one simply takes samples from μ and ν, sort them, and computes the norm of the difference between the sorted values. I liked the simplicity of the procedure and found out through further reading that N. Bonnotte showed during his Ph.D. that it could be generalized to higher dimensions $\Omega\subset\mathbb{R}^D$, through *random projections*.

Let $\boldsymbol{\theta}\in\mathbb{S}^D$ be a direction in the unit sphere from $\mathbb{R}^D$. We write $\boldsymbol{\theta}^\star\sharp\mu$ as the pushforward measure for the projection along $\boldsymbol{\theta}$: $\mathbf{x}\mapsto\boldsymbol{\theta}^\star\mathbf{x}$. N. Bonnotte shows that the **sliced Wasserstein distance** $\mathcal{SW}(\mu,\nu)$ shares many properties with $W^2(\mu,\nu)^2$. It is defined as follows:

$$\mathcal{SW}(\mu,\nu)=\int_{\boldsymbol{\theta}}\mathcal{W}_2\left(\boldsymbol{\theta}^\star\sharp\mu,\boldsymbol{\theta}^\star\sharp\nu\right)^2d\boldsymbol{\theta} \tag{1.52}$$

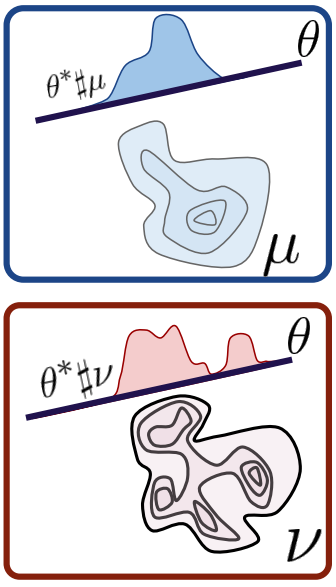


Figure 1.33: The sliced Wasserstein distance compares μ and ν over many projections.

and can be understood as characterizing the discrepancies between distributions μ and ν as the cumulated difference they exhibit over a multitude of projections. In practice, (1.52) is approximated through a Monte Carlo estimate by randomly sampling the projections $\boldsymbol{\theta}$ on the unit sphere.

I was much inspired by the sliced Wasserstein (SW) approach and wondered whether this simpler metric could serve as the basis for an alternative to the GAN dual formulation. It turns out that the topic was already active and that papers came out at that time about SW generative models, while I was busy implementing mine.

Sliced Wasserstein flows: theory In his work, N. Bonnotte proposed the *Iterative Distribution Transfer* (IDT) algorithm, which consists in leveraging the sliced Wasserstein distance iteratively: the data is projected on many random directions, then increasing rearrangement is performed on each one of them independently, and finally the transported data is brought back to the initial domain. In small dimensional cases, IDT is empirically shown to yield effective transportation.

I suspected that the IDT algorithm could serve as the basis for a more general nonparametric transport scheme, in the sense that it starts with some arbitrary measure μ and aims at transporting it iteratively. At each time step t, our current transport map T_t is augmented so that $\mu_t\triangleq T_t\sharp\mu$ gets closer to the target measure ν. At this point, my capabilities in mathematics were challenged to their limits and I started a collaboration about the topic with U. Simsekli, who also called A. Durmus

to the rescue. This all resulted in the proposal of the *sliced Wasserstein flows* (SWF, [LŞM+19]). I am enthusiastic about the resulting theory, that I summarize briefly now.

In [LŞM+19], we frame the approach as a functional optimization problem in the space of probability measures:

$$\min_{\mu}\Big\{\mathcal{F}_\lambda^\nu(\mu) \triangleq \frac{1}{2}\mathcal{SW}(\mu,\nu) + \lambda\mathcal{H}(\mu)\Big\}, \tag{1.53}$$

where the $\mathcal{H}(\mu)$ term is an original contribution in this context and stands for the entropy of μ. It serves as a practical regularization term that avoids the generated samples to collapse to the training data and also allows for an interesting treatment.
Similar to Euclidean spaces, one way to formulate this optimization problem is to construct a gradient flow of the form:

$$\partial_t \mu_t = -\nabla\mathcal{F}(\mu_t), \tag{1.54}$$

where ∇ denotes a notion of gradient in the metric space induced by the SW distance. If such a flow can be constructed, one can utilize it both for practical algorithms and theoretical analysis.

It turns out that this is the case and the core contribution of [LŞM+19] is to show that the partial differential equation (1.54) (PDE) has a stochastic counterpart. More precisely, let us consider a stochastic process $(\mathbf{x}_t)_t$, that is the solution of the following stochastic differential equation starting at $\mathbf{x}_0 \sim \mu_0$:

$$d\mathbf{x}_t = v(\mathbf{x}_t, \mu_t)dt + \sqrt{2\lambda}d\mathbf{w}_t, \tag{1.55}$$

where $(\mathbf{w}_t)_t$ denotes a standard Brownian motion and $v(\mathbf{x}_t, \mu_t)$ is a *drift function* whose details are not important at this point. Then, the probability distribution of $\mathbf{x}_t$ at time t solves the PDE given in (1.53). This means that, if we can simulate (1.55), then the distribution of $\mathbf{x}_t$ converges to the solution of (1.53), therefore, we could use the sample paths $(\mathbf{x}_t)_t$ as samples drawn from $(\mu_t)_t$, so that we could hope for their distribution to become close to ν in the limit $t \to \infty$, due to the choice (1.53) for our cost functional.

Sliced Wasserstein Flow: algorithm The theory comes with several difficulties for which we propose a solution and I will skip all the details here to instead just provide the resulting algorithm, trying to convey what I understand as the meaning of each term. Considering a population $\mathcal{X}_k$ of N samples $\mathcal{X}_k \triangleq \left\{\mathbf{x}_k^n \in \mathbb{R}^D\right\}_n$ and writing $\bar{\mu}_k$ as its empirical distribution, the algorithm iteratively applies the following update equations for all $n \in \mathbb{N}$:

$$\mathbf{x}_0^n \sim \mu_0, \mathbf{x}_{k+1}^n = \mathbf{x}_k^n + h\, v_k(\mathbf{x}_k^n) + \alpha\, \mathbf{z}_{k+1}^n, \tag{1.56}$$

where μ_0 is easy to sample from, like Gaussian i.i.d, $h, \alpha > 0$ are the stepsize and regularization terms, respectively, $\mathbf{z}_{k+1}^n \in \mathbb{R}^D$ is white i.i.d. Gaussian noise, and the drift v_k at iteration k is given by:

$$v_k(\mathbf{x}) = -\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{S}^D}\left[\Delta_{k\boldsymbol{\theta}}\left(\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle\right)\boldsymbol{\theta}\right]. \tag{1.57}$$

It is the main quantity used to update each sample $\mathbf{x}_k^n$. I understand it as a weighted average of the directions $\boldsymbol{\theta} \in \mathbb{S}^D$, where each direction is weighted by the displacement required to match the target distribution along this direction. This displacement is provided by $\Delta_{k\boldsymbol{\theta}}$, defined as:

$$\Delta_{k\boldsymbol{\theta}}(z) = z - \boldsymbol{T}(\boldsymbol{\theta}^\star\sharp\bar{\mu}_k \to \boldsymbol{\theta}^\star\sharp\nu)\left(z\right), \tag{1.58}$$

which is easy to compute as (1.51). As we can notice, if $\bar{\mu}_k$ is already very close to the target ν, then $\boldsymbol{T}(\boldsymbol{\theta}^\star\sharp\bar{\mu}_k \to \boldsymbol{\theta}^\star\sharp\nu)\left(z\right)$ will be very close to z, so that the particles will not move much, as expected.

To summarize, in practice, at each iteration k and for many different directions $\boldsymbol{\theta}$, we compute the quantiles of both the projected current particles $\{\langle\boldsymbol{\theta}, \boldsymbol{x}_k^n\rangle\}_n$ and the projected training data $\{\langle\boldsymbol{\theta}, \boldsymbol{y}^n\rangle\}_n$. This allows to construct the increasing rearrangements $\boldsymbol{T}(\boldsymbol{\theta}^\star\sharp\bar{\mu}_k \to \boldsymbol{\theta}^\star\sharp\nu)$ and thus to compute the drift $v_k(\boldsymbol{x}_k^n)$ for each particle by averaging over the directions through (1.57). This is then used to update the particles with (1.56).

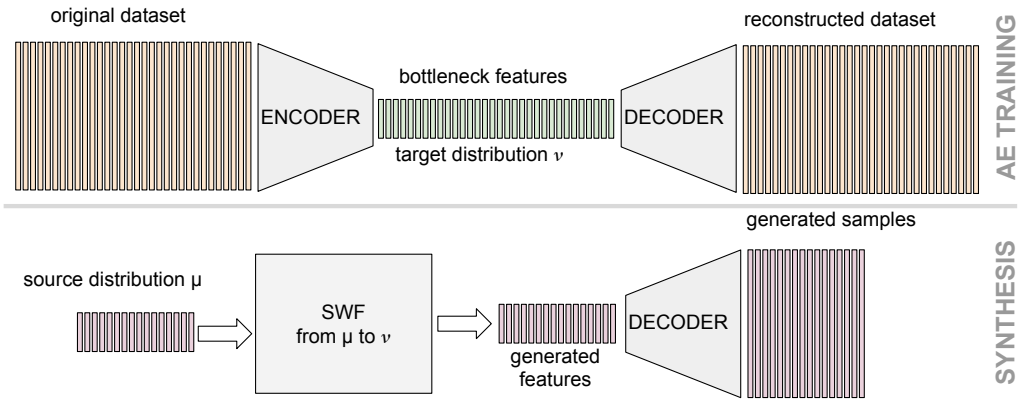


Figure 1.34: In order to use SWF for large-dimensional data, [LŞM+19] considers applying it on the bottleneck features of a pre-trained auto-encoder

SWF: handling large-dimensional data. The main difficulty involved by SWF lies in the need to compute the drift as an average over the D-dimensional sphere in (1.57). When the dimension becomes large as for realistic images, this expectation is plagued with the curse of dimensionality, so that the drift is poorly estimated with a limited number of projections.
To address this problem, we came up with a simple fix in [LŞM+19] that is illustrated in Figure 1.34. We first train an unconstrained auto-encoder on the data, before applying SWF on the resulting bottleneck features. The method proved very effective and allows to produce very realistic samples as shown in Figure 1.35.

Figure 1.35: Examples of samples obtained by SWF on the MNIST, FashionMNIST and CelebA datasets. From [LŞM+19].

Discussion on SWF. We can show that SWF simplifies to IDT when we don't consider regularization ($\alpha = 0$), pick a stepsize $h = 1$ and choose the directions $\boldsymbol{\theta}$ as an orthonormal basis of $\mathbb{R}^D$. It thus stands as a promising direction to understand such algorithms.

A remarkable point about SWF that I wish to highlight is it allows in theory to generate data that mimicks the target distribution ν *without ever requiring the actual observation of samples from it*, but only quantiles for random projections making it a potentially interesting privacy-preserving generative approach. Furthermore, the computation of those projections and quantiles can be made in an embarrassingly parallel fashion, which is exploited by my open-source MIT implementation.

1.3.4 Relative positional encoding for linear Transformers

Since I believe it both illustrates my current interests and is important for the continuation of my research programme, I want to briefly describe my very recent work on Transformer architectures, which is detailed in [LCW+21]. Once again, this research is the outcome of a fruitful collaboration, this time with two other research groups from France and Taiwan, respectively lead by G. Richard and Y. Yang.

The Transformer model proposed by A. Vaswani in the landmark paper *Attention is all you need* is a new kind of neural networks that quickly became state of the art in many application domains, including the processing of natural language, images, audio or bioinformatics. Its core and new component is the *attention layer*. It computes M output values $\mathbf{y}_m$ from N input values $\mathbf{v}_n$, all being vectors of an arbitrary dimension. The way it is done is reminiscent of classical non-parametric regression and consists in a simple weighted sum:

$$\mathbf{y}_m = \frac{\sum_n a_{mn}\mathbf{v}_n}{\sum_n a_{mn}}\,, \tag{1.59}$$

where each *attention* coefficient $a_{mn} \in \mathbb{R}_+$ — gathered in the $M \times N$ matrix $\mathbf{A}$ — indicates how important is the value $\mathbf{v}_n$ in the computation of the output $\mathbf{y}_m$.

One of the main interests of the Transformer is to provide an original method to compute those attentions. D-dimensional feature vectors $\mathbf{k}_n$ and $\mathbf{q}_m$ are attached to all items of the input and output sequences and are called *keys* and *queries*, respectively. Gathering them in the $N \times D$ and $M \times D$ matrices $\mathbf{K}$ and $\mathbf{Q}$, we define *softmax dot-product attention* as:

$$\mathbf{A} = \exp\Big(\mathbf{Q}\mathbf{K}^\top \Big/ \sqrt{D}\Big) \equiv [a_{mn} = \mathcal{K}(\mathbf{q}_m, \mathbf{k}_n)]_{mn}\,, \tag{1.60}$$

where the function exp is applied element-wise. The right hand side in (1.60) is a generalization introduced recently for which $\mathcal{K}$ is a *kernel* function.
Parameters of the model pertain to how keys $\mathbf{k}_n$, values $\mathbf{v}_n$ and queries $\mathbf{q}_m$ are obtained from the raw input sequences $\mathbf{x}_n$, usually by time-distributed fully connected layers, i.e. through $\mathbf{k}_n = \mathbf{W}^k\mathbf{x}_n$, $\mathbf{v}_n = \mathbf{W}^v\mathbf{x}_n$ and $\mathbf{q}_n = \mathbf{W}^q\mathbf{x}_n$, with parameters $\mathbf{W}^k, \mathbf{W}^v, \mathbf{W}^q$.

Motivations: attention and non-local models. From the perspective of the research I presented in Section 1.2.1, I consider the attention mechanism leveraged by the Transformer as a very nice extension to the "non-local regression" framework I was working on with kernel additive modeling. Of course, this should not be understood in anyways as a claim that KAM can be seen as a precursor to the Transformer, but rather the other way around: I see it as an interesting idea to revisit those old models in the light of the recent breakthrough brought in by the Transformer: can we leverage this architecture to train kernels like those in KAM? Can we have additive Transformers? Can we do median attention? etc.

For all these reasons, I was very interested by the Transformer. However, I was frustrated about two particular aspects for it that I thought were relevant topics for investigation. They were scalability and positional encoding and I review them now.

Linear Transformers The original Transformer architecture explicitly computes the attention matrix $\mathbf{A}$, leading to a $\mathcal{O}(MN)$ complexity that prevents scaling to very long sequence lengths. Although this is not necessarily a problem when sequence lengths are barely longer than a few hundreds, as in some language processing tasks, it is prohibitive for very large signals like high-resolution images or audio.

I was focusing on this scalability issue when I realized that several satisfying approaches had already been proposed to allow for long sequences:

- *Attention clustering* schemes were proposed that group items among which dependencies are computed through regular attention. This is either done based on simple temporal proximity within the sequences, leading to chunking strategies, or by clustering the keys and values. Inter-cluster dependencies are either ignored or summarized via fixed-length context vectors that are coined in as *memory*.

- Assuming the attention matrix to be *sparse*. In this case, only a few a_{mn} are nonzero, and they are usually chosen *a priori* as some predefined neighborhood of each location m.
- Assuming $\mathbf{A}$ has a (low-rank) *structure* and can be decomposed as the product of two smaller matrices. A prototypical example is the Linformer, which is limited to fixed-length inputs. Another very recent line of research in this same vein takes:

$$\mathbf{A} \approx \phi(\mathbf{Q})\phi(\mathbf{K})^\top, \tag{1.61}$$

  where $\phi : \mathbb{R}^D \to \mathbb{R}^R$ is a non-linear *feature map* applied to each key $\mathbf{k}_n$ and query $\mathbf{q}_m$, and the feature dimension is small: $R \ll \min(M, N)$.
- When $\mathcal{K}$ in (1.60) is a positive (semi) definite kernel, the Performer proposed recently by K. Choromansky leverages *reproducing kernel Hilbert spaces* to show random ϕ may be used to exploit this convenient decomposition (1.61) *on average*, even when $\mathbf{A}$ is not low rank:

$$\mathcal{K} \succeq 0 \Leftrightarrow \mathbf{A} = \mathbb{E}_\phi \left[\phi(\mathbf{Q})\phi(\mathbf{K})^\top \right], \tag{1.62}$$

  where ϕ is drawn from a distribution that depends on $\mathcal{K}$. A simple example is $\phi_W(\mathbf{k}_n) = \max(0, \mathbf{W}\mathbf{k}_n)$, with a random $\mathbf{W} \in \mathbb{R}^{K \times D}$ for some $K \in \mathbb{N}$.

Whenever an efficient scheme like (1.61) or (1.62) is used, the outputs $\mathbf{Y}$ can be obtained in linear complexity without actually computing the attention coefficients a_{mn}, as:

$$\mathbf{Y} \leftarrow \mathrm{diag}(\mathbf{d})^{-1} \left[\phi(\mathbf{Q}) \left[\phi(\mathbf{K})^\top \mathbf{V} \right] \right], \text{ with } \mathbf{d} = \phi(\mathbf{Q}) \left[\phi(\mathbf{K})^\top \mathbf{1}_N \right] \tag{1.63}$$

Since these approaches looked sufficiently convincing to me, I decided to abandon my investigations on this topic for the time being[20] and use these linear variants instead, to rather focus on another interesting direction I had identified.

Positional encoding. In my work on KAM presented above in Section 1.2.1, I could largely notice that exploiting the *positions* of the samples and not only their *values* could be the key to strong boosts in performance (see e.g. Figure 1.21). Two typical approaches were undertaken in the literature to incorporate the position in Transformer models:

- The original Transformer by A. Vaswani literally *adds* some additional information to the inputs of the network, i.e. before the first attention layer. This can be equivalently understood as augmenting the keys, values and queries:

$$\mathbf{k}_n \leftarrow \mathbf{k}_n + \overline{\mathbf{k}}_n, \mathbf{v}_n \leftarrow \mathbf{v}_n + \overline{\mathbf{v}}_n, \mathbf{q}_m \leftarrow \mathbf{q}_m + \overline{\mathbf{q}}_m, \tag{1.64}$$

  where we write $\overline{\mathbf{k}}_n \in \mathbb{R}^D$ for the *keys positional encoding* (PE) at position $n \in \mathbb{N}$ and analogously for values and queries. They propose a deterministic scheme based on trigonometric functions, which is shown to work just as well as trainable embeddings.
- As an example of positional encoding *in the attention domain*, *Self-Attention with Relative Position Representations* (RPE) was proposed by Shaw *et al.*, building on the idea that time lags $m - n$ are more important than absolute positional encoding (APE) for prediction. It reads:

$$\mathbf{A} = \exp\Big(\Big(\mathbf{Q}\mathbf{K}^\top + \mathbf{\Omega} \Big) \Big/ \sqrt{D} \Big), \text{ with:} \mathbf{\Omega} \equiv \left[\omega_{mn} = \sum_{d=1}^{D} q_{md} \mathcal{P}_d(m - n) \right]_{mn}. \tag{1.65}$$

  The terms $\mathcal{P}_d$ now act as D different encodings for *time lags*, selected based on the queries. This change is advocated as bringing important performance gains in many application areas and enjoyed a widespread use ever since.

Although writing down the positional encoding in the attention domain was shown beneficial for performance by many authors, I was only aware of implementations that either require the computation of $\mathbf{A}$, or clustered attention schemes, which *in fine* decompose $\mathbf{A}$ into smaller attention matrices, and *compute them*. This is in sharp contrast to (1.61) and (1.62), which never compute the attention matrix.

[20] I was working on strategies similar to the *Routing Transformer* when that paper came out...

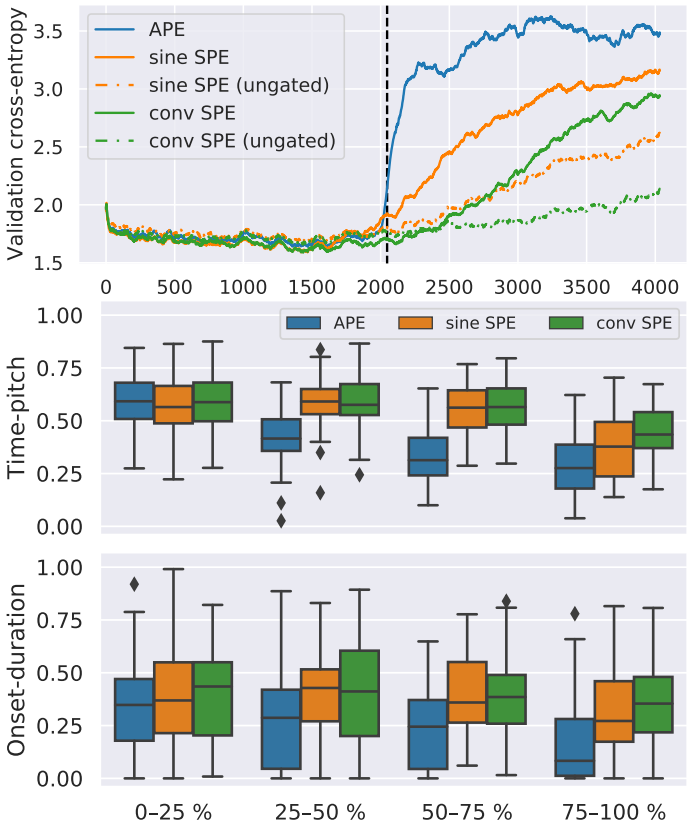


Figure 1.36: Performance of different variants of the proposed SPE scheme. (top) Validation cross-entropy vs. token position on a music generation task. (lower is better) the **black** vertical line indicates the max position to which the models are trained. (bottom) Musical style similarity on a music generation (higher is better) between output and an initial prompt through two musically-motivated metrics, as a function of the time position in the output. From [LCW+21]

We proposed **Stochastic Positional Encoding** (SPE) in [LCW+21], as a general way to do PE *in the keys domain* while enforcing a particular attention pattern devised *in the attention domain*. This notably enables RPE without explicit computation of attentions. To our knowledge, it is the first RPE strategy that is compatible with $\mathcal{O}(N)$ Transformers. The key idea for SPE is to see the attention kernel $\mathcal{P}_d(m,n)$ in (1.65) as a *covariance*:

$$\forall\left(\mathcal{M},\mathcal{N}\right),\forall\left(m,n\right),\mathcal{P}_d(m,n)=\mathbb{E}\left[\overline{\mathbf{Q}}_d(m)\overline{\mathbf{K}}_d(n)\right], \tag{1.66}$$

where $\overline{\mathbf{Q}}_d(m)$ and $\overline{\mathbf{K}}_d(n)$ are two real zero-mean random variables, which are chosen with the single condition that their covariance function matches $\mathcal{P}_d(m,n)$. Semantically, they should be understood as (randomly) encoding position m for queries and position n for keys, respectively. When multiplied together as in dot-product attention, they yield the desired attention template $\mathcal{P}_d(m,n)$ on average. The central intuition is that **the actual positional encodings do not matter as much as their dot-product**.

As a very remarkable special case, it is easy to derive *stationary* positional encodings $\overline{\mathbf{Q}}_d$ and $\overline{\mathbf{K}}_d$ that yield relative encoding $\mathcal{P}_d(m-n)$. This simple fact has a very important consequence: it allows implementing RPE without ever actually computing the attention matrix **A**, thus being compatible with linear Transformers.

Performance . We experimentally report in [LCW+21] that the proposed PE scheme translates the main feature of classical RPE to long sequences, which is to be robust to test sequence lengths that are not observed in the training data. This is illustrated in 1.36.
In my opinion, SPE is a nice example of how (even basic) probabilistic signal processing could be of use in machine learning

2 Research programme: machine learning for signal processing

2.1 The work I did that is mostly obsolete

Considering the progress made by the community, there are many works I did that are no more really relevant among the large body of research I presented in the preceding chapter. I want to shortly discuss the reasons for this, so as to better picture the plans I have for the future.

Predicting filters is obsolete. DNN-based separation models are so powerful now that directly predicting the target source signal from the mixture in an *end-to-end fashion* actually works better than predicting the parameters for constructing filters.
In most of my research on source separation, the sophisticated masking strategies I described in Section 1.1 helped recovering the sources from approximate source spectrograms. Now, their bounded performance — even if it is very good — is standing as a limitation for the quality of the output. We could indeed observe during SiSEC 2018 that the best performing systems already showed *no significant difference in performance to oracle Wiener filtering* [SLI18].
This step forward was notably made possible by two facts. First, DNN mappings between the mixture and sources magnitude spectrograms became so powerful that simply plugging in the mixture phase to the estimates is often the most efficient solution. I interpret this by the fact that it is very inappropriate to use Gaussian or α-harmonizable models that put the highest probability mass on 0 when very accurate magnitude estimates.[1] Second, the most effective systems today integrate the forward and inverse transforms *inside the model* as for the Conv-TasNet, which takes them as 1d-convolutions and estimates their parameters. My guess is that even more powerful models based on Transformers that relax the need to do processing in a ”sliding window” fashion will yield even better performance in a close future.

Kernel spectrogram models are obsolete. Although I long acknowledged this fact, which motivated me into embracing deep learning in the first place, I want to highlight this again here for completeness of this section. Even if I believe they set up a new state-of-the-art for a few years, the kernel methods I proposed and that are described in Section 1.2.1 are largely obsolete now — at least in terms of performance — for several reasons. First, they operate in the time-frequency domain, which is progressively being abandoned as a fixed representation, as mentioned above. Second, they impose fixed regularities for the signals of interest that are vastly too restrictive and were soon strongly outperformed by DNN-based methods.

Source separation into fixed stems is a mature technology. In my view, the decision I took to define a fixed scenario for MUSDB18, where music should be separated into four stems: vocals, bass, percussion and ”other” is probably the one that had the deepest impact on the community. Since this dataset serves as the *de facto* standard in the domain, this choice probably shaped at least part of the research in the domain.
Inspecting the performance attained by its most recent developments, I think we can say that such source separation into fixed stems is no longer so much of a research topic but rather a *mature technology* that is bound to make it to a wide audience very soon and is already being integrated into commercial products.
Although it was useful for advancing research, I hence believe that this *fixed source scenario* as a research topic is obsolete now.

I must say that it is quite an extraordinary thing for me to look over the shoulder and see how far a whole community of researchers can get in 15 years of time. I am not sure I would have believed it back then if I had listened to the separated signals we get today, that are barely different from the originals. Even if the research mentioned above may not be used in current systems, it was certainly useful *on the way*, and I am happy with this.

[1] It may be interesting to check whether the BEADS model presented on page 19 could help in this regard.

2.2 My long-term research strategy

Even if I consider the remaining part of the work I did to still be relevant today, notably the most theoretical investigations on probabilistic models presented in Section 1.1 as well as the recent developments described from Section 1.3.3 onward, experience shows they will probably get obsolete anytime soon. Even in the unlikely case that I get lucky and find some groundbreaking new model that revolutionizes signal processing, the community will rapidly come up with extensions that make it look old-fashioned in a matter of months.
Before switching to more short-term and practical perspectives in the next section and because I understand it as one of the objectives of the exercise I'm doing with writing this thesis, I want to briefly describe what I feel are the guiding principles for my work in the long run and that shouldn't change much anytime soon.

Community service. In terms of long-term impact, it looks to me in retrospect that one of my best moves was to dedicate a significant amount of my time to provide tools, datasets and a benchmark canvas for the music separation community to do better collaborative research. I know that a significant part of my future research effort will continue in this direction. In my view, it comes in two different flavours.

- **Engineering.** In the last ten years, it looks to me that it became fundamental for impactful research to be *reproducible*. In our community, I believe this mostly means it should come with well-designed and open-sourced implementations and datasets. This has obvious drawbacks in the sense it requires a lot of engineering, which takes time and resources, but it is precisely where there is some room for me and the team I can gather to help out, because *I like engineering*.
- **Publishing and social service.** Along the years, I spent a (very) significant part of my time reviewing several hundreds of papers, which I believe is part of my duty and has actually been acknowledge through some awards I received.[2] Hence, I can say I am currently still playing a role in "the old way" of publishing research.
  However, I first have serious ethical issues regarding the interaction of public funding with the publishing institutions managing those publications. Second, I think this publishing workflow is not flexible enough to account for the most recent practice and I am not surprised it is progressively *de facto* abandoned in favor of open alternatives like arxiv.org. I am willing to be increasingly involved in the sustainability of these new approaches, promoting open and continuous peer-review, instant publishing, large-scale community communication, online events, minimal and transparent fees, etc.

The non-measurable impact of theory. Although I am first and foremost an *applied scientist*, I enjoy looking for deeper theoretical insights regarding the methods I develop. This lead me to study several mathematical objects that were not so common in my community, like Gaussian processes, α-stable distributions or sliced Wasserstein distances.
Even if some of those mathematical methods or objects may have lost their initial interest in the context I proposed them for, I believe that delving into theoretically grounding research proves rewarding in the long run. To illustrate this, I know that K. Yoshii from Kyoto University developed very general extensions of the Gaussian models I proposed for source separation, and that U. Simsekli was partially inspired by my work on α-stable distributions to see how they could be used to model gradient noise in stochastic optimization. Both had a significant impact.

Supervision and collaborations. Even if the investigations I personally lead could contribute to some tiny degree in advancing state of the art in signal processing, they are negligible compared to what my students or collaborators have and will achieve.
The reason why I am applying to professorship through writing this thesis is precisely because I believe that *teaching and collaborating* is a fundamental aspect of my work and is the one with the deeper long-term impact. Even if I played a role in their initial training, I am already getting a very strong feedback from my students that largely surpasses the initial effort, when they regularly come to me with exciting new research topics. The same goes for collaborators, except that I was the one getting all the benefits all the way.

[2] I had an IEEE award for outstanding reviewing, I am a member of the "reviewing" group for the IEEE audio technical committee, I am area chair for ICASSP, an "expert reviewer" for ICML and I was granted free registration to NeurIPS as one of "the top 10% high-scoring reviewers" (whatever that means).

2.3 Short-term research questions

In this section, I will detail three theoretical short-term directions that I would like to investigate, because they look both promising to me and are the natural continuation for my past research. Such directions should provide sufficient challenges for my future students to work on in the next few years. Applications will be mentioned in the following section.

Positional encoding and unstructured models In Section 1.3.4, I presented my recent work on positional encoding. In this vein, I see several short-term tracks for research that could be worth investigating:

- **Signal-dependent PE**. From some perspective, fixed schemes for positional encodings, whether they be absolute or relative, all come with the same limitation, which is: they boil down to the assumption that positions should be handled *in a signal independent fashion*. However, I see this as extremely limited and counter-intuitive. On the contrary, experience shows that temporal dependency structure, e.g. in music, highly depends on the actual track considered.
  In my opinion, the fact that most PE methods in the literature are signal-independent is an explanation for the surprising observation that we don't see long-term attention happening in practice: if we are to devise an attention pattern that works *regardless of the signal considered*, the best conservative solution is indeed instinctively to attend only to immediate neighborhoods: even if very long term dependencies may be more appropriate on an instance level, they are not systematic.
- **Unstructured models.** It seems clear to me that Transformers as introduced by A. Vaswani are only once step in the long process of minimizing the impact of *inductive bias* in the design of deep architectures. As I see it, the next natural step would be to leverage the very flexible attention mechanism to dynamically update the actual *structure and connectivity of the network*, rather than limiting it to the signal domain. The recent *vision Transformer* or *Perceiver* models all go in this direction. The particular twist I could be investigating is the role of flexible PE schemes in these dynamic architectures.

Optimal transport with diffusion models When they were proposed a few years ago as a continuous version of the ResNet, neural ordinary differential equations (NODE) striked me as a very refreshing view for deep neural networks. In essence, they relax the notion of *depth* within a DNN, to make it a continuous variable. In this new formulation, a model coincides with an ODE for which the drift term becomes a neural network. Stochastic extensions were also proposed.

- **Better handling the depth**. The different implementations for NODE I could see all incorporate depth in a rather trivial way: it is usually either simply ignored, or concatenated to the input. Even if some research on the topic has been done that models weights along depth through another NODE, I think there is much room for improvement in this domain, that could take inspiration from how weights are related in ResNets in the first place. An intriguing direction would be to exploit attention mechanisms in this setup.
- **Direct feedback alignment and NODE.** Training NODE is notoriously difficult. It turns out I am regularly fascinated by alternative training strategies for deep learning that are proposed in the literature. Among them, I have a particular interest for *direct feedback alignment* (DFA), that offers training through random feedback of the error within a deep network instead of back-propagation and enjoys many desirable properties.[3] I feel that trying to train NODE with DFA may open interesting research directions. Preliminary work on this matter are encouraging and done in collaboration with J. Philibert and H. Glotin.
- **Optimal transport and diffusion models.** I see recently proposed diffusion-based generative models like *wavegrad* as related to NODE in the sense they relax the notion of layers to replace it by some continuous "whitening" processes from the signal to the noise domain. The process is inverted for generation. I believe we could be inspired by both this approach and SWF to provide generative NODE models that would operate along Wasserstein gradient flows for optimal transport.

[3] Apart from the scientific interest I have in DFA, I also have some personal reasons to like it: it can leverage the powerful optical computers developed by lighton.io. It is the company created by the people I worked for on the topic of hardware compressed sensing (see Section 1.3.2).

Stable gradients models. As a more exploratory research direction, I want to highlight my interest in the recent developments that U. Simsekli gave to the theory of stochastic gradient descent, that involve α-stable processes instead of the more classical Brownian-motion based approach. As he shows with his collaborators, the gradients of the millions of parameters of neural networks can be jointly modeled as obeying to stable dynamics along iterations.

- **Joint gradient models.** For now, assumptions about the actual *spectral measure* for these processes — see (1.30) — remained very general and were not the actual topic for the investigations. I want to leverage my past research on source localization [FVLB17a, FVLB17b] to identify *cliques* and dependencies among the parameters, by picking specific models for this spectral measure, like atomic (1.32) or elliptically contoured (1.34). Doing so, the obvious challenge is *scale*, since a model typically involves dozens of millions of parameters.
- **Localization as optimization.** The intriguing *lottery ticket hypothesis* and its follow-ups suggest that within a randomly initialized DNN, there are groups of parameters that already lead to almost-optimal performance when taken is isolation, i.e. when setting all the others to zero. This suggests that a new kind of optimization algorithms could be developed, that aims at identifying such ”cliques”, instead of doing gradient descent over all parameters.
  The bet on this matter would be that the gradient of parameters within the same clique share common probabilistic features, which are yet to be discovered.

2.4 Applications

The research I want to do comprises a significant part of theory, with initial research directions summarized above. However, contributions in this domain have to come with demonstrations of relevance in applications.
As I made it clear in the previous chapter, most of my previous research was applied to music filtering, with some occasional excursions in other domains of signal processing. In the near future, I picture myself as doing mostly the same, except that the audio and non audio applications envisioned will change and may be balanced differently, with probably a little bit less of audio, since my coming to Montpellier lead me to meet researchers working in life sciences with whom I want to spend time collaborating.

2.4.1 Audio and music processing

I love everything about research in audio and music processing. First, it is a very challenging domain whose data is readily available and sufficiently complex to be an ideal playground for many mathematical models and theoretical contributions. Second, as time went by, I came to know many researchers working in the field, with whom I spent a lot of time at the numerous conferences I attended. Young Ph.D. students became team leaders and newcomers dropped in, but I feel this is a lucky community where researchers working on similar subtopics are more likely to become collaborators than competitors.

Music separation. It is clear that I will continue to have regular collaborations and some contributions applied to music separation, that is still far from a solved topic in many cases of interest.

- **Evaluation campaigns.** Although I handed general chairing of SiSEC over to Y. Mitsufuji, it is quite likely that I set up some new task for music separation in the future. My short-term plan on this purpose is to produce a new music dataset. It could notably be a new version for MUSDB where all stems are available, instead of the 4-stems scenario it comprises now.
- **It was attention all the way.** I think it could be interesting to present how KAM could be seen as a particular instance of *additive transformers* with median attention. This would notably be a natural testbed for my theoretical research on signal-aware PE.
- **Waveform models.** My research is still involving new models for audio signals, and music separation is a an easy and convincing way to demonstrate their interest. In particular, I consider using it for evaluating long-range transformers and diffusion-based waveform models.
- **Unsupervised training for source separation** would unlock the main limitation for this research topic, which is scarcity of data.

Audio enhancement. In the different models I proposed for source separation, the observed mixture is always taken as a sum of the desired source with some additive interference, and this kind of approaches gets good performance provided we have a sufficiently large training dataset. However, there are many cases where it is not the actual target source that is found within the mixture, but rather a very degraded version of it, as in very old and noisy recordings from which we would like to reconstruct high-quality signals. The filtering paradigm I considered so far cannot fit in that canvas and generate new parts of a signal that are not in the mixture already: it can only *remove* information. In this setting, generative audio models can be used instead, as recent continuations to the classical *bandwidth extension* problem. It is a rapidly advancing field with many contributions, especially in speech and to some lesser extent in music.

- **Audio heritage enhancement.** As I mentioned already, I am interested in the impact of technology on music creation. At some point, I got convinced that our immaterial audio heritage could also serve as extraordinary raw material to promote trans-generational creation. This lead me to propose the ANR-funded KAMoulox project, to collaborate with owners for large audio heritage archives, and it would be interesting to apply new audio enhancement methods to revive such historical material, and see how they could be used in musical contexts.
- **Music conversion.** As a more general instance of the audio enhancement scenario mentioned above, I would be glad to apply my efforts to unpaired music conversion, notably those on diffuse models and Transformers.

Music generation Another disruptive technology that could be coming in the future may very well be the capability to automatically generate realistic music. Even if the idea may seem somewhat daunting, my guess is that *unconditional music creation* (from scratch) will probably be kind of boring and not really impactful except for some nice scientific papers. Instead, I see *informed* scenarios as more realistic and desirable, where users could guide the generation process, for instance in an examplar-based way.[4] Such technologies could notably lead to new exciting practices for easy DJ-ing and personalized audio streaming.

- **Symbolic music generation** is a domain I recently got into [LCW+21], and it is a good experimental application for long-range Transformer architectures.
  Although our experiments still don't really show an advantage of using models that can account for very long range dependencies, I suspect it is largely due to the limitations of the models we considered rather than to a supposedly intrinsic short-ranged aspect of musical content, that is going against musicological intuition. Actually, seeing it the other way around, I want to investigate the viability of *measuring the attention span on carefully designed music tasks as a way to evaluate PE schemes*.
- **Music generation in the waveform domain** has witnessed astonishing contributions recently, e.g. with the *jukebox: a generative model for music* by P. Dhariwal. I would be curious to contribute to such an exciting topic, and my starting point would be the design of relevant objective metrics. On this point, the initial research direction I want to consider is to exploit the optimal-transport setup presented in Section 1.3.3, where GAN *discriminators* stand for Kantorovitch potentials and could be used to assess quality.

2.4.2 Life sciences

In 2017, I joined the Montpellier University as a move to diversify the topics on which to apply my research. As of today, this lead me to be involved in two new research collaborations.

Time-series models for genomics In the context of a collaborative project initiated by G. Krouk and in collaboration with B. Dumortier, I investigate how sequence models like long range Transformers can be used in genomics applications.

- **Phenotype prediction.** One of the objectives of the project on our side is to see how some phenotypic features concerning the roots of plants can be predicted based on gene regulatory activation measures acquired on their leaves. This inverse problem is notably made difficult by the scarcity of data. We are for now still investigating several classic machine learning algorithms to predict human-labels from the measures. Future plans involve generating fake images of roots, and then to condition the generator based on the measures.

[4] There are already some companies selling such synthesis services for symbolic music.

- **Protein sequences processing.** Proteins are sequences of amino acids, taken among a vocabulary of 22 possibilities. They have a length of up to a few thousands. Several tasks of interest were identified within the project. First, *protein-protein interaction* prediction consists in predicticting whether two given proteins will interact or not. Very recently, Transformers have be tried out for this purpose and yield good performance. The follow-up I want to investigate is *informed generation* of protein sequences conditionally on positive interaction. If successful, this application could have a nice impact in terms of drug design.
  In any case, I see protein sequences as a great application domain for my current research on Transformer models and positional encoding.

Ecological research. In the context of a collaboration between the micro-electronics department and the functional ecology laboratory of Montpellier, S. Chamaille-Jammes gathered large amounts of audio and geolocalisation data from sensors that were attached to several zebra in the wild. On the machine learning side, this work is supervised by my colleague F. Stöter.

- **Large scale audio vizualization**. In order to be of use, this massive dataset requires curation, but the audio content has some very specific features due to the way it was acquired. In particular, many parts of the recordings are very noisy. Then, designing new unsupervised imaging methods would be useful for exploiting the data in practice.
- **Animal movement modeling.** The migration of animals in the wild is of particular interest for the ecologists. Beyond this particular data, www.movebank.org is a worldwide archive for localization data to study animal behaviour. In my opinion, this particular topic is a very nice application area for my investigations on NODE. My initial research direction on this point would be to define a NODE model to predict the velocity field for the migrations as a function of the location, and possibly also of the time in the day/month.

2.4.3 Conclusion

Witnessing the joint evolution of these research fields, my feeling is that the frontiers between signal processing and machine learning are becoming increasingly blurry as time goes by. Researchers from signal processing like me pretend to be doing machine learning, while researchers in machine learning are *in fine* designing new ways to analyze and process signals. On top of both, I see probability theory as *the logic of science*, as E.T. Jaynes would put it, i.e. a tool that is often useful to go from the models to methods exploiting them.

In retrospect, I see the applications for my past research effort as mostly characterized by two ingredients that are of interest to me: (i) **Sequential data**. Of course, audio in general and music in particular have been the main areas I considered, but I like to study other kinds of settings. The fundamental question I am interested in is how to make sense of always varying spatial and temporal observations. (ii) **Data processing**, as opposed to information extraction, has always been my favorite canvas for applications. I like to design systems that transform or generate data, so that I tend to be rather interested by regression than by classification tasks.

Finally, as a last word, I could describe the ideal theoretical contribution I would like to witness in my lifetime and that maybe I can help getting to. It would be a data-integration model that is compatible with the internal consistency and beauty of probability theory and that involves a host of small-scale processing agents — say neurons — to reach its goal. The underlying vision would be to better put words and equations on the deepest mystery: how such a metaphysical object like the human mind can be implemented by one kilo and a half of microscopic particles that are essentially composed of water and fat.

A Publications

Publications: book chapters

[B.Liu12] A. Liutkus. *Processus gaussiens pour la séparation de sources et le codage informé*. PhD thesis, Télécom ParisTech, 2012.

[B.Nug18] A. A. Nugraha, A. Liutkus, and E. Vincent. Deep neural network based multichannel audio source separation. In *Audio Source Separation*, pages 157–195. Springer, Mar. 2018. doi:10.1007/978-3-319-73031-8_7.

[B.Par18] B. Pardo, A. Liutkus, Z. Duan, and G. Richard. Applying source separation to music. In *Audio Source Separation and Speech Enhancement*, volume Chapter 16. Wiley, Aug. 2018. doi:10.1002/9781119279860.ch16.

[B.Raf14] Z. Rafii, A. Liutkus, and B. Pardo. REPET for Background/Foreground Separation in Audio. In G. Naik and W. Wang, editors, *Blind Source Separation*, pages 395–411. Springer Berlin Heidelberg, 2014. doi:10.1007/978-3-642-55016-4_14.

Publications: journals

[J.Can19] E. Cano, D. Fitzgerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter. Musical Source Separation: An Introduction. *IEEE Signal Processing Magazine*, 36(1):31–40, Jan. 2019. doi:10.1109/MSP.2018.2874719.

[J.Car19] B. Caramiaux, F. Lotte, J. Geurts, G. Amato, M. Behrmann, F. Bimbot, F. Falchi, A. Garcia, J. Gibert, G. Gravier, et al. AI in the media and creative industries. 2019.

[J.Dre15] A. Drémeau, A. Liutkus, D. Martina, O. Katz, C. Schülke, F. Krzakala, S. Gigan, and L. Daudet. Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques. *Optics Express*, 29(9):11898–11911, Apr. 2015. doi:10.1364/OE.23.011898.

[J.Fit16] D. Fitzgerald, A. Liutkus, and R. Badeau. Projection-based demixing of spatial audio. *IEEE Transactions on Audio, Speech and Language Processing*, May 2016. doi:10.1109/TASLP.2016.2570945.

[J.Fon20] M. Fontaine, R. Badeau, and A. Liutkus. Separation of Alpha-Stable Random Vectors. *Signal Processing*, page 107465, Jan. 2020. doi:10.1016/j.sigpro.2020.107465.

[J.Liu11] A. Liutkus, R. Badeau, and G. Richard. Gaussian Processes for Underdetermined Source Separation. *IEEE Transactions on Signal Processing*, 59(7):3155 – 3167, Feb. 2011. doi:10.1109/TSP.2011.2119315.

[J.Liu12] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937–1949, 2012. doi:10.1016/j.sigpro.2011.09.016.

[J.Liu14a] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron. Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium. *Scientific Reports*, 4:14, July 2014. doi:10.1038/srep05552.

[J.Liu14b] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel Additive Models for Source Separation. *IEEE Transactions on Signal Processing*, page 14, June 2014. doi:10.1109/TSP.2014.2332434.

[J.Liu15] N. Liu, A. Liutkus, J.-F. Aubry, L. Marsac, M. Tanter, and L. Daudet. Random Calibration for Accelerating MR-ARFI Guided Ultrasonic Focusing in Transcranial Therapy. *Physics in Medicine and Biology*, 60(3):21, Jan. 2015. doi:10.1088/0031-9155/60/3/1069.

[J.Liu16] A. Liutkus and E. Vincent. Démixer la musique. *Interstices*, Jan. 2016. URL https://hal.inria.fr/hal-01350450.

[J.Nug16] 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*, 24(10):1652–1664, June 2016. doi:10.1109/TASLP.2016.2580946.

[J.Oze13] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation: Nonnegative tensor factorization approach. *IEEE Transactions on Audio, Speech and Language Processing*, 21(8):1699–1712, Aug. 2013. doi:10.1109/TASL.2013.2260153.

[J.Raf18] Z. Rafii, A. Liutkus, F.-R. Stöter, S. Ioannis Mimilakis, D. Fitzgerald, and B. Pardo. An Overview of Lead and Accompaniment Separation in Music. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(8):1307–1335, 2018. doi:10.1109/TASLP.2018.2825440.

[J.Sim15] U. Simsekli, A. Liutkus, and T. Cemgil. Alpha-Stable Matrix Factorization. *IEEE Signal Processing Letters*, page 5, Sept. 2015. URL https://hal.inria.fr/hal-01194354.

[J.Sto19] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-Unmix - A Reference Implementation for Music Source Separation. *Journal of Open Source Software*, 4(41):1667, Sept. 2019. doi:10.21105/joss.01667.

Publications: conferences

[BL14] J. Beliao and A. Liutkus. OOPS: une approche orientée objet pour l'interrogation et l'analyse linguistique de l'interface prosodie/syntaxe/discours. In *4e Congrès Mondial de Linguistique Française*, volume 8, pages 2565–2581, Berlin, Germany, July 2014. doi:10.1051/shsconf/20140801273.

[DCDL17] D. Di Carlo, K. Déguernel, and A. Liutkus. Gaussian framework for interference reduction in live recordings. In *AES International Conference on Semantic Audio*, Erlangen, Germany, June 2017. URL https://hal.inria.fr/hal-01515971.

[DCLD18] D. Di Carlo, A. Liutkus, and K. Déguernel. Interference reduction on full-length live recordings. In *ICASSP: International Conference on Acoustics, Speech, and Signal Processing*, pages 736–740, Calgary, Canada, Apr. 2018. IEEE. doi:10.1109/ICASSP.2018.8462621.

[DLGE13a] C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Non-negative matrix factorization for single-channel EEG artifact rejection. In *ICASSP*, Vancouver, Canada, 2013. doi:10.1109/ICASSP.2013.6637836.

[DLGE13b] C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Nonnegative Tensor Factorization for Single-Channel EEG Artifact Rejection. In *IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, United Kingdom, Sept. 2013. URL https://hal.telecom-paris.fr/hal-02288386.

[DLM+15] A. Dremeau, A. Liutkus, D. Martina, O. Katz, C. Schülke, F. Krzakala, S. Gigan, and L. Daudet. Approches Bayésiennes pour la reconstruction de phase Application à l'optique des milieux complexes. In *GRETSI*, Lyon, France, Sept. 2015. URL https://hal.archives-ouvertes.fr/hal-02892511.

[FLB16] D. Fitzgerald, A. Liutkus, and R. Badeau. PROJET - Spatial Audio Separation Using Projections. In *41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016. IEEE. URL https://hal.archives-ouvertes.fr/hal-01248014.

[FLBR12] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard. Probabilistic model for main melody extraction using constant-Q transform. In *37th International Conference on Acoustics, Speech, and Signal Processing ICASSP'12*, pages 5357–5360, Kyoto, Japan, 2012. IEEE. URL https://hal.inria.fr/hal-00945290.

[FLGB17] M. Fontaine, A. Liutkus, L. Girin, and R. Badeau. Explaining the Parameterized Wiener Filter with Alpha-Stable Processes. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, United States, Oct. 2017. URL https://hal.archives-ouvertes.fr/hal-01548508.

[FLR+14] D. Fitzgerald, A. Liutkus, Z. Rafii, B. Pardo, and L. Daudet. Harmonic/Percussive Separation Using Kernel Additive Modelling. In *IET Irish Signals & Systems Conference 2014*, Limerick, Ireland, June 2014. URL https://hal.inria.fr/hal-01000001.

[FNB+19] M. Fontaine, A. A. Nugraha, R. Badeau, K. Yoshii, and A. Liutkus. Cauchy Multichannel Speech Enhancement with a Deep Speech Prior. In *EUSIPCO 2019 - 27th European Signal Processing Conference*, Coruña, Spain, Sept. 2019. URL https://hal.telecom-paris.fr/hal-02288063.

[FRL17] D. Fitzgerald, Z. Rafii, and A. Liutkus. User Assisted Separation of Repeating Patterns in Time and Frequency using Magnitude Projections. In *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, United States, Mar. 2017. URL https://hal.inria.fr/hal-01515956.

[FSL+18] M. Fontaine, F.-R. Stöter, A. Liutkus, U. Simsekli, R. Serizel, and R. Badeau. Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition. In D. Y., G. S., M. R., P. M., and W. D., editors, *LVA/ICA: Latent Variable Analysis and Signal Separation*, volume LNCS, pages 13–23, Surrey, United Kingdom, July 2018. Springer. doi:10.1007/978-3-319-93764-9_2.

[FVLB17a] M. Fontaine, C. Vanwynsberghe, A. Liutkus, and R. Badeau. Scalable Source Localization with Multichannel Alpha-Stable Distributions. In *25th European Signal Processing Conference (EUSIPCO)*, Proc. of 25th European Signal Processing Conference (EUSIPCO), pages 11–15, Kos, Greece, Aug. 2017. URL https://hal.archives-ouvertes.fr/hal-01531252.

[FVLB17b] M. Fontaine, C. Vanwynsberghe, A. Liutkus, and R. Badeau. Sketching for nearfield acoustic imaging of heavy-tailed sources. In *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)*, volume 10169 of *Latent Variable Analysis and Signal Separation 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings*, pages 80–88, Grenoble, France, Feb. 2017. doi:10.1007/978-3-319-53547-0_8.

[KDL18] N. Keriven, A. Deleforge, and A. Liutkus. Blind Source Separation Using Mixtures of Alpha-Stable Distributions. In *ICASSP: International Conference on Acoustics, Speech and Signal Processing*, pages 771–775, Calgary, Canada, Apr. 2018. IEEE. doi:10.1109/ICASSP.2018.8462095.

[KLC15] S. Kırbız, A. Liutkus, and A. T. Cemgil. Dialogue enhancement using kernel additive modelling. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 2242–2245. IEEE, 2015.

[KOLG14] S. Kırbız, A. Ozerov, A. Liutkus, and L. Girin. Perceptual coding-based informed source separation. In *22nd European Signal Processing Conference (EUSIPCO-2014)*, Lisbonne, Portugal, Sept. 2014. URL https://hal.inria.fr/hal-01016314.

[LB15] A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015. IEEE. URL https://hal.archives-ouvertes.fr/hal-01110028.

[LBR10] A. Liutkus, R. Badeau, and G. Richard. Informed Source Separation Using Latent Components. In V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, editors, *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 498–505, Saint Malo, France, 2010. Springer. doi:10.1007/978-3-642-15995-4_62.

[LBR11] A. Liutkus, R. Badeau, and G. Richard. Multidimensional Signal Separation with Gaussian Processes. In *Statistical Signal Processing Workshop*, pages 401–404, Nice, France, June 2011. doi:10.1109/SSP.2011.5967715.

[LBR13] A. Liutkus, R. Badeau, and G. Richard. Low bitrate informed source separation of realistic mixtures. In *ICASSP*, pages 66–70, Vancouver, Canada, 2013. IEEE. doi:10.1109/ICASSP.2013.6637610.

[LCW+21] A. Liutkus, O. Cífka, S.-L. Wu, U. Şimşekli, F.-R. Stöter, Y.-H. Yang, and G. Richard. Relative positional encoding for transformers with linear complexity. In *International Conference on Machine Learning*. PMLR, 2021.

[LDA+12] A. Liutkus, A. Drémeau, D. Alexiadis, S. Essid, and P. Daras. Analysis of dance movements using gaussian processes. In *the 20th ACM international conference*, page 1375, Nara, France, Oct. 2012. ACM Press. doi:10.1145/2393347.2396492.

[LDDR13] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *WIAMIS 2013 - The 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, pages 1–4, Paris, France, July 2013. doi:10.1109/WIAMIS.2013.6616139.

[LFB15] A. Liutkus, D. Fitzgerald, and R. Badeau. Cauchy Nonnegative Matrix Factorization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, Oct. 2015. URL https://hal.inria.fr/hal-01170924.

[LFR15] A. Liutkus, D. Fitzgerald, and Z. Rafii. Scalable audio separation with light kernel additive modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015. IEEE. URL https://hal.inria.fr/hal-01114890.

[LGS+12] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard. Informed Audio Source Separation: A Comparative Study. In *20th European Signal Processing Conference (EUSIPCO-2012)*, page n/c, Bucarest, Romania, Aug. 2012. URL https://hal.archives-ouvertes.fr/hal-00809525.

[LL10] A. Liutkus and P. Leveau. Separation of Music+Effects sound track from several international versions of the same movie. In *AES 128th Convention*, London, United Kingdom, May 2010. URL https://hal.inria.fr/hal-00959108.

[LMGD14] A. Liutkus, D. Martina, S. Gigan, and L. Daudet. Compressed sensing under strong noise. Application to imaging through multiply scattering media. In *European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sept. 2014. URL https://hal.inria.fr/hal-01074786.

[LOBR12] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard. Spatial coding-based informed source separation. In *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, Aug. 2012. URL https://hal.inria.fr/hal-00869618.

[LOMG15] A. Liutkus, T. Olubanjo, E. Moore, and M. Ghovanloo. Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, Oct. 2015. URL https://hal.inria.fr/hal-01174886.

[LRB+12] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *37th International Conference on Acoustics, Speech, and Signal Processing ICASSP'12*, pages 53–56, Kyoto, Japan, 2012. IEEE. doi:10.1109/ICASSP.2012.6287815.

[LRD18] A. Liutkus, C. Rohlfing, and A. Deleforge. Audio source separation with magnitude priors: the BEADS model. In *ICASSP: International Conference on Acoustics, Speech and Signal Processing*, Signal Processing and Artificial Intelligence: Changing the World, pages 56–60, Calgary, Canada, Apr. 2018. IEEE. doi:10.1109/ICASSP.2018.8462515.

[LRP+14] A. Liutkus, Z. Rafii, B. Pardo, D. Fitzgerald, and L. Daudet. Kernel Spectrogram models for source separation. In *HSCMA*, Nancy, France, May 2014. URL https://hal.inria.fr/hal-00959384.

[LŞC15] A. Liutkus, U. Şimşekli, and T. Cemgil. Extraction of Temporal Patterns in Multi-rate and Multi-modal Datasets. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015. URL https://hal.inria.fr/hal-01170932.

[LSL+17] S. Leglaive, U. Simsekli, A. Liutkus, R. Badeau, and G. Richard. Alpha-Stable Multichannel Audio Source Separation. In *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Proc. 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, United States, Mar. 2017. IEEE. URL https://hal.archives-ouvertes.fr/hal-01416366.

[LSL+19] S. Leglaive, U. Simsekli, A. Liutkus, L. Girin, and R. Horaud. Speech enhancement with variational autoencoders and alpha-stable distributions. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 541–545, Brighton, United Kingdom, May 2019. IEEE. doi:10.1109/ICASSP.2019.8682546.

[LŞM+19] A. Liutkus, U. Şimşekli, S. Majewski, A. Durmus, and F.-R. Stöter. Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions. In *36th International Conference on Machine Learning (ICML)*, Long Beach, United States, June 2019. URL https://hal.inria.fr/hal-02191302.

[LSR+17] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave. The 2016 Signal Separation Evaluation Campaign. In P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, editors, *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)*, volume 10169 of *Theoretical Computer Science and General Issues*, pages 323 – 332, Grenoble, France, Feb. 2017. Springer. doi:10.1007/978-3-319-53547-0_31.

[LY17] A. Liutkus and K. Yoshii. A diagonal plus low-rank covariance model for computationally efficient source separation. In *IEEE international workshop on machine learning for signal processing (MLSP)*, Tokyo, Japan, Sept. 2017. URL https://hal.inria.fr/hal-01580733.

[MBB+12] S. Marchand, R. Badeau, C. Baras, L. Daudet, D. Fourer, L. Girin, S. Gorlow, A. Liutkus, J. Pinel, G. Richard, N. Sturmel, and S. Zhang. DReaM: A Novel System for Joint Source Separation and Multi-Track Coding. In *133rd AES Convention*, page CD 133papers, San Francisco, United States, Oct. 2012. URL https://hal.archives-ouvertes.fr/hal-00809503.

[MBL17a] P. Magron, R. Badeau, and A. Liutkus. Lévy NMF : un modèle robuste de séparation de sources non-négatives. In *Colloque GRETSI*, Actes du XXVIème Colloque GRETSI, Juan-Les-Pins, France, Sept. 2017. URL https://hal.archives-ouvertes.fr/hal-01540484.

[MBL17b] P. Magron, R. Badeau, and A. Liutkus. Lévy NMF for Robust Nonnegative Source Separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017)*, New Paltz, NY, United States, Oct. 2017. IEEE. URL https://hal.archives-ouvertes.fr/hal-01548488.

[NLV16] A. A. Nugraha, A. Liutkus, and E. Vincent. Multichannel Music Separation with Deep Neural Networks. In *European Signal Processing Conference (EUSIPCO)*, Proceedings of the 24th European Signal Processing Conference (EUSIPCO), pages 1748–1752, Budapest, Hungary, Aug. 2016. URL https://hal.inria.fr/hal-01334614.

[OLBR11] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, Mohonk, NY, United States, Oct. 2011. URL https://hal.inria.fr/inria-00610526.

[ORK+15] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus. The 2015 Signal Separation Evaluation Campaign. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, volume 9237 of *Latent Variable Analysis and Signal Separation*, pages 387–395, Liberec, France, Aug. 2015. doi:10.1007/978-3-319-22482-4_45.

[PBLM15] T. Prätzlich, R. Bittner, A. Liutkus, and M. Müller. Kernel additive modeling for interference reduction in multi-channel music recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015. URL https://hal.inria.fr/hal-01116686.

[PPL17] F. Pishdadian, B. Pardo, and A. Liutkus. A multi-resolution approach to common fate-based audio separation. In *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, United States, Mar. 2017. URL https://hal.inria.fr/hal-01515951.

[RBL07] Z. Rafii, R. Blouet, and A. Liutkus. Discriminant approach within non-negative matrix factorization for musical components recognition. DMRN+2, 2007.

[RCL17] C. Rohlfing, J. E. Cohen, and A. Liutkus. Very Low Bitrate Spatial Audio Coding with Dimensionality Reduction. In *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, United States, Mar. 2017. URL https://hal.inria.fr/hal-01515954.

[RLB17] C. Rohlfing, A. Liutkus, and J. M. Becker. Quantization-aware Parameter Estimation for Audio Upmixing. In *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, United States, Mar. 2017. URL https://hal.inria.fr/hal-01515955.

[RLP15] Z. Rafii, A. Liutkus, and B. Pardo. A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, France, Apr. 2015. URL https://hal.inria.fr/hal-01116689.

[SEL+18] U. Simsekli, H. Erdogan, S. Leglaive, A. Liutkus, R. Badeau, and G. Richard. Alpha-stable low-rank plus residual decomposition for speech enhancement. In *ICASSP: International Conference on Acoustics, Speech, and Signal Processing*, pages 651–655, Calgary, Canada, Apr. 2018. IEEE. doi:10.1109/ICASSP.2018.8461539.

[SLB+16] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron. Common Fate Model for Unison source Separation. In *41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Proceedings of the 41st International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016. IEEE. URL https://hal.archives-ouvertes.fr/hal-01248012.

[SLI18] F.-R. Stöter, A. Liutkus, and N. Ito. The 2018 Signal Separation Evaluation Campaign. In D. Y., G. S., M. R., P. M., and W. D., editors, *LVA/ICA: Latent Variable Analysis and Signal Separation*, volume LNCS, pages 293–305, Surrey, United Kingdom, July 2018. Springer. doi:10.1007/978-3-319-93764-9_28.

[SLP+12] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet. Linear Mixing Models for Active Listening of Music Productions in Realistic Studio Conditions. In *132nd AES Convention*, page Paper 8594, Budapest, Hungary, Apr. 2012. URL https://hal.archives-ouvertes.fr/hal-00790783.

[SNV+15] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales Cordovilla, S. Dalmia, I. Illina, and A. Liutkus. Robust ASR using neural network based speech enhancement and feature simulation. In *ASRU*, Arizona, United States, Dec. 2015. URL https://hal.inria.fr/hal-01204553.

[SRG+16] A. J. R. Simpson, G. Roma, E. M. Grais, R. D. Mason, C. Hummersone, A. Liutkus, and M. D. Plumbley. Evaluation of Audio Source Separation Models Using Hypothesis-Driven Non-Parametric Statistical Methods. In *European Signal Processing Conference*, Budapest, Hungary, Aug. 2016. EURASIP. URL https://hal.inria.fr/hal-01410176.

[WMK+18] D. Ward, R. D. Mason, C. Kim, F.-R. Stöter, A. Liutkus, and M. D. Plumbley. SiSEC 2018: State of the art in musical audio source separation - subjective selection of the best algorithm. In *WIMP: Workshop on Intelligent Music Production*, Huddersfield, United Kingdom, Sept. 2018. URL https://hal.inria.fr/hal-01945362.

[ZGL13] S. Zhang, L. Girin, and A. Liutkus. Informed Source Separation from compressed mixtures using spatial wiener filter and quantization noise estimation. In *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 61–65, Vancouver, Canada, May 2013. doi:10.1109/ICASSP.2013.6637609.

Publications: patents

[P.Blo10] R. Blouet, S. M. Aziz Sbai, and A. Liutkus. Automatic audio source separation with joint spectral shape, expansion coefficients and musical state estimation, 2010. URL https://patents.google.com/patent/US20100174389A1.

[P.Gir10] L. Girin, A. Liutkus, G. Richard, and R. Badeau. Method and device for forming a digital audio mixed signal, method and device for separating signals, and corresponding signal, Oct. 2010. URL https://hal.telecom-paris.fr/hal-02651076.

[P.Sba10] S. M. Aziz Sbai, R. Blouet, and A. Liutkus. Automatic gathering strategy for unsupervised source separation algorithms , 2010. URL https://patents.google.com/patent/US20100138010A1/.

Publications: software and datasets

[S.Liu14] A. Liutkus. ccmixter - a corpus for vocals-music separation. 2014. URL http://liutkus.net/ccmixter_corpus.zip.

[S.Liu19] A. Liutkus and F.-R. Stöter. Norbert: Multichannel-wiener filtering. URL https://doi.org/10.5281/zenodo.3386463.

[S.Raf17] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. MUSDB18 - a corpus for music separation. URL https://doi.org/10.5281/zenodo.1117372.

[S.Sto19a] F.-R. Stöter and A. Liutkus. Musdb 0.3.1. doi:10.5281/zenodo.3271451.

[S.Sto19b] F.-R. Stöter and A. Liutkus. Museval 0.3.0. doi:10.5281/zenodo.3376621.

[S.Sto19c] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 2019. doi:10.21105/joss.01667.

B Résumé en français

Ce chapitre est un résumé succinct en français, exigé par la procédure, par lequel j'ai choisi de me concentrer sur l'explicitation du fil conducteur de ma recherche passée et future. C'est donc délibérément que j'omets ici les aspects techniques, pour lesquels je renvoie au document principal.

B.1 Recherche effectuée: traitement probabiliste de mélanges

Motivations: la séparation de sources. Le sujet scientifique qui m'a principalement occupé au cours de ma recherche passée a été celui de la séparation de sources appliquée à la musique. Il s'agit d'un problème inverse dont le but est de récupérer les contributions isolées des différents instruments d'un morceau de musique, comme la partie chantée, percussive, la basse, etc. Le mélange musical que l'on cherche à inverser peut être régi par les lois de l'acoustique dans le cas d'un enregistrement live, ou bien être le fruit du travail d'un ingénieur du son, et alors comporter une multitude d'effets complexes et non linéaires. Il s'agit d'un sujet très riche qui a attiré un intense effort international de recherche depuis des décennies [J.Can19, J.Raf18].

Lorsque j'ai commencé à m'y intéresser, les méthodes les plus performantes de l'état de l'art reposaient sur un paradigme *de filtrage*. À partir du spectrogramme du mélange, on estime les spectrogrammes des sources, qui sont alors utilisés pour construire un filtre de Wiener permettant la séparation. Ce formalisme soulevait deux aspects principaux: tout d'abord, le **modèle probabiliste sous-jacent** n'était pas clair pour moi, me conduisant à m'intéresser à la théorie des processus stochastiques. Par ailleurs, la technique nécessitait des **modèles de spectrogrammes** suffisamment expressifs pour rendre compte de signaux musicaux complexes. Ces deux aspects ont sous-tendu une large part de mon travail.

B.1.1 Modèles probabilistes pour la séparation de sources.

Processus Gaussiens pour la séparation de sources. Bien qu'ils reposent sur de solides fondations théoriques datant de l'après guerre, je ne comprenais pas encore les modèles probabilistes sous-jacents aux techniques d'estimation et de filtrage de l'état de l'art. C'est notamment en cherchant à comprendre quelles sont les hypothèses qui les sous-tendent dans le domaine temporel que je me suis rendu compte de l'aspect tout à fait général de l'approche, que je me suis attaché à expliciter.

- **Les processus Gaussiens** sont de puissants modèles probabilistes permettant de modéliser des fonctions aléatoires Y à valeur dans $\mathbb{R}$ ou $\mathbb{C}$ et définies sur un ensemble $\mathbb{L}$ quelconque, comme le temps $\mathbb{L} = \mathbb{R}$ ou l'espace $\mathbb{L} = \mathbb{R}^2$. Leur particularité est d'être des modèles *non-paramétriques*, par opposition à des approches qui résument toute la fonction en un jeu fini de paramètres comme les coefficients d'un modèle polynomial. Un processus Gaussien n'est pas défini globalement, mais plutôt en se *donnant* une fonction de covariance, qui renseigne sur la similarité $k(l, l') = \mathbb{E}[Y(l)Y(l')]$ entre les valeurs prises en deux points $(l, l') \in \mathbb{L}$. Ainsi, il est facile de se donner un modèle pour l'ensemble des fonctions "lisses", quelle que soit leur allure générale, tandis que cela sera compliqué avec un modèle paramétrique.
- **Application à la séparation.** Mon premier travail théorique [J.Liu11, LBR11] a été de proposer une manière assez générale de formuler la séparation de sources, comme une inférence Bayésienne des valeurs prises par des processus Gaussiens étant donné leur mélange et leurs fonctions de covariance. Ce formalisme très général rend notamment possible d'appliquer des méthodes jusque là réservées à l'audio à d'autres cas de figures plus exotiques [DLGE13a, DLGE13b, LDA$^+$12].
- **Dans le cas stationnaire**, c'est à dire lorsque les fonctions de covariance ne dépendent pas de l'origine des temps, et lorsque les signaux sont considérés sur une grille régulière du type $\mathbb{Z}^D$, alors le formalisme se simplifie fortement dans le domaine spectral. On retombe sur les

méthodes classiques de l'état de l'art, mais on peut aussi les appliquer en haute dimension, ce qui était moins connu.

Séparation de sources informée. Mon travail de thèse a été effectué dans le cadre d'un projet financé par l'Agence Nationale de la Recherche et qui visait au développement d'un nouveau type d'interaction entre le grand public et le contenu musical, grâce à la séparation de sources [MBB+12]. L'idée principale était de permettre à l'utilisateur de séparer de manière fiable la musique pour du karaoke, de la respatialisation, etc. Pour ce faire, en studio dans une première phase d'encodage où on dispose à la fois des sources et du mix, on préparait une métadonnée suffisamment petite pour qu'elle puisse être tatouée dans le signal, mais tout de même suffisamment informative pour qu'elle puisse être utilisée pour permettre une séparation de bonne qualité dans la deuxième phase de décodage, effectuée auprès de l'utilisateur, quand les sources originales ne sont plus disponibles.

- **Cas paramétrique.** Mes premiers travaux sur ce sujet de la séparation informée ont consisté à compresser les spectrogrammes des sources par diverses méthodes (matrices non négatives et compression d'images), puis à les transmettre au décodeur, qui effectuait un filtrage de Wiener. Cette idée très simple a conduit à des performances remarquables [LBR10, J.Liu12, LGS+12] et je l'ai ensuite étendue au cas des signaux stéréo, en aboutissant en fin de compte à des débits par source à séparer de l'ordre de 1kbps, ce qui est très faible.
- **Rapprochements avec la théorie de l'information.** Les résultats de nos méthodes étaient très bons, mais ils demeuraient d'une qualité bornée qu'on peut comprendre d'un point de vue pratique comme liée à l'utilisation de la phase du mélange pour la reconstruction et qui n'est pas égale à celle de la source originale. D'un point de vue théorique, cette distorsion intrinsèque est égale à la variance *a posteriori* de l'estimateur de Wiener.
  Une idée forte d'A. Ozerov sur ce sujet a été d'exploiter la théorie du codage de sources, qui permet de compresser de manière optimale une information dont on dispose d'une loi de probabilité. La grande originalité de l'approche a consisté à exploiter non pas la distribution *a priori* des sources comme il est d'usage, mais bien plutôt celle obtenue *a posteriori*, après avoir pris en compte le mélange. D'un point de vue assez général, l'idée revient à concevoir un système de codage de source pour lequel l'encodeur et le décodeur partagent une information commune, ici le mélange. Exploiter cette information ne peut que réduire le débit nécessaire. Les systèmes correspondants ont permis de décupler les performances obtenues, à des coûts en débit sensiblement égaux ou inférieurs [OLBR11, J.Oze13, KOLG14].

Processus scalaires alpha-stables. Fort du formalisme des processus Gaussiens, j'ai pu mener à bien mon travail initial de recherche, mélangeant théorie de l'information et séparation de sources, ainsi que comprendre en profondeur les hypothèses sous-jacentes à de nombreuses pratiques de l'état de l'art. Cependant, il me restait certains points à élucider, qui se sont avérés porteurs de développements très riches.

- **Les limites du modèle Gaussien.** Une méthode de séparation de sources s'accompagnait de certains choix, notamment l'utilisation de spectrogrammes de puissance ou bien d'amplitude pour les traitements. Bien qu'une part importante de la communauté faisait le choix des premiers, cohérent avec le modèle Gaussien, une part tout aussi importante obtenait des performances remarquables avec des spectrogrammes d'amplitude. Dans ce cas, le filtrage se faisait de la même manière, mais en utilisant ces spectrogrammes d'un genre peu orthodoxe.
  Pour ce qui est de l'utilisation de spectrogrammes d'amplitude, il existait bien certains modèles de type Poisson, mais ils ne me convainquaient pas, parce qu'ils ne pouvaient pas être menés jusqu'aux formes d'onde, et ne permettaient en tout cas pas de justifier l'opération de filtrage, qui sont deux caractéristiques particulièrement élégantes du modèle Gaussien à mes yeux. Au contraire, on se retrouvait au fond à devoir ignorer l'existence de la phase et le fait que nos représentations étaient avant tout des transformées de Fourier.
- **Les modèles α-harmonisables.** C'est en recherchant dans de nombreuses directions si je pouvais trouver un modèle compatible avec de telles pratiques que j'ai compris que le cas Gaussien faisait partie de la famille beaucoup plus large des processus α-stables, regroupant l'ensemble des lois stables par addition: si on modélise des sources comme α-stables, leur mélange le sera aussi par définition. Pour faire court, le cas particulier des distributions symétriques et α-stables

complexes ($\mathcal{S}\alpha\mathcal{S}_c$) se comporte exactement comme dans le cas Gaussien, sauf que ce sont les moments d'ordre α qui s'ajoutent au lieu d'être les variances.
La première étape était franchie, il ne restait plus qu'à dérouler le fil: il se trouve qu'on peut définir les processus α-harmonisables, qui sont des processus dont la transformée de Fourier a des coefficients indépendants $\mathcal{S}\alpha\mathcal{S}_c$. On peut montrer qu'ils existent et sont stationnaires. L'additivité des spectrogrammes d'amplitude était acquise.

- **Filtrage α-Wiener.** Outre l'introduction des modèles α-harmonisables dans le domaine de la séparation de sources audio, la principale contribution de [LB15] a été de prouver que ces modèles permettent non seulement de justifier l'additivité des α-spectrogrammes, mais également leur utilisation à des fins de filtrage, d'une manière analogue au filtrage de Wiener. Par exemple, si on suppose des sources Cauchy-harmonisables ($\alpha = 1$), alors le spectrogramme d'amplitude du mélange pourra se décomposer comme la somme de ceux des sources, et le formalisme indique que les utiliser tels quels pour le filtrage est optimal dans le sens où il s'agit de l'espérance *a posteriori* des sources étant donné le mélange.
- **Estimation des paramètres.** Bien qu'il ait l'avantage de donner une base théorique à la manipulation des α-spectrogrammes pour l'analyse et le filtrage de mélanges, le modèle α-harmonisable présentait quelques difficultés inattendues. À part le cas Gaussien ($\alpha = 2$) et le cas Cauchy ($\alpha = 1$), il ne permet notamment pas l'écriture d'une fonction de vraisemblance, étape classique pour estimer les paramètres d'un modèle. Au lieu de cela, c'est la fonction caractéristique qui est facile à écrire, ce qui au fond signifie qu'on dispose d'une expression analytique pour des *moments* des variables considérées. J'ai appliqué ces modèles à quelques cas de séparation de données très impulsives [LOMG15] ou à des données très bruitées [LFB15].

Processus alpha-stables multivariés. Cette recherche sur les processus α-stables s'est très vite montrée riche de pistes nouvelles. En particulier, leur extension au cas multivarié m'a révélé une flexibilité que le modèle Gaussien n'a pas, exploitable en traitement du signal.
Dans le cas général, la distribution d'un vecteur α-stable ne se définit pas avec une matrice de dispersion comme dans le cas Gaussien, mais plutôt par le biais d'une *mesure spatiale*, qui indique l'énergie du signal en provenance de chaque direction de la sphère. Cette représentation s'avère extrêmement adaptée à des problématiques de type filtrage ou localisation acoustique.

- **Modèle ponctuel.** Des travaux de l'état de l'art concernaient l'utilisation des modèles α-stables pour l'analyse en composante indépendante. Leur idée de départ était de considérer des mélanges linéaires instantanés de sources i.i.d., du type $\mathbf{x}_t = \mathbf{A}\mathbf{s}_t$. Dans ce cas, la mesure spectrale est une somme de Diracs, chaque atome étant concentré dans la direction d'une source. Une partie de la recherche conduite par M. Fontaine a consisté à appliquer ce formalisme pour la localisation acoustique, lorsque les sources sont α-harmonisables [FVLB17a, FVLB17b].
  La difficulté principale de l'approche est la méthode d'estimation des paramètres. Nous avons adopté une stratégie de type *sketching*, qui consiste à d'abord calculer à partir des observations un ensemble bien choisi de statistiques (de type moments généralisés), puis à estimer les paramètres du modèle qui collent au mieux à ces statistiques. L'approche a de nombreux avantages, notamment celui d'être très lègère sur le plan calculatoire, et de donner des performances très encourageantes.
  J'ai également collaboré sur l'exploitation de ce formalisme pour la séparation de sources [FLB16, J.Fit16, FRL17]. Dans ce cas, l'approche a consisté de la même manière à apprendre des modèles structurés à partir de différentes projections des observations.
- **Le cas sous-Gaussien**, dit aussi *elliptique* constitue un sous-ensemble des distributions α-stables. On peut comprendre un tel vecteur comme Gaussien dont la matrice de covariance est perturbée aléatoirement par un coefficient d'impulsion qui peut atteindre des valeurs très élevées. L'intérêt principal de ce modèle est qu'il permet de se ramener au cas Gaussien, conditionnellement à une estimation de ces variables d'impulsion.
  Une première idée a été de modéliser la distribution jointe des sources et des mélanges par un tel modèle, ce qui a eu pour conséquence d'introduire une forme de *robustesse* de l'estimation des parmaètres [LSL$^+$17]. Par ailleurs, ce modèle nous a aussi permis d'introduire un algorithme de décomposition matricielle (NMF) dans le cas α-stable général [SEL$^+$18, J.Sim15], dépassant le cas Cauchy que j'avais considéré jusqu'à présent [LFB15].
- **Cas général.** Le tour de force réalisé par M. Fontaine a été de montrer que le cas général d'un

vecteur symétrique α-stable pouvait être considéré pour mettre au point des filtres d'un nouveau genre [J.Fon20], dont les performances dans des contextes fortement bruités dépasse celles de l'estimateur Gaussien.

Modèles hybrides. Mes efforts pour proposer des modèles probabilistes plus adaptés aux séries temporelles très impulsives qu'on observe en audio avaient fait du chemin. Cependant, j'en avais identifié deux limitations supplémentaires que je me suis attelé à lever.

- **Des modèles non unimodaux.** Le modèle α-stable est compatible avec des moments fractionnaires qui s'ajoutent pour des sources indépendantes. Comme on l'a vu, il explique ainsi qu'en moyenne, le module d'une somme de variables stables avec $\alpha \approx 1$ vaille la somme des modules de chacune. Cependant, une telle distribution attribue la plus grande masse de probabilité à 0, ce qui ne reflète que très mal notre *a priori* dans certains cas. En effet, lorsque nous utilisons un puissant modèle profond pour estimer le spectrogramme d'amplitude d'une source à partir de celui du mélange, c'est plutôt une distribution en forme de tore ("de donut") que nous nous attendrions à avoir: nous avons en fait une idée assez précise de l'amplitude, c'est la phase qui est inconnue. Après filtrage, il faudrait que le module obtenu pour cette source soit cohérent avec l'*a priori*, ce qui n'était pas le cas avec des masques de type Wiener. Dans [LRD18], nous proposons le modèle BEADS qui approxime ce tore par un mélange de Gaussiennes, et qui s'est avéré très performant.
- **Modèles multi-alphas.** Une limitation du formalisme tel que je l'ai évoqué jusqu'ici est d'imposer le même modèle pour toutes les sources, ce qui n'est pas intuitif. Dans [FLGB17], M. Fontaine a considéré le cas d'une séparation de sources dont chacune a son exposant caractéristique α_j. Cela l'a conduit à donner une justification théorique à une forme paramétrique du filtre de Wiener, fréquemment utilisée sur des bases heuristiques depuis sont introduction dans les années 1980. Dans [SEL+18, J.Sim15], nous avons exploité la sous-Gaussianité des variables α-stables pour aborder ce problème avec des algorithmes de type Markov-Chain Monte Carlo (MCMC).
- **Mélanges de distribution α-stables.** Pour finir, une idée que nous avons envisagée dans [KDL18] a été de supposer qu'une seule source est majoritairement présente dans chaque point temps fréquence, aboutissant à un modèle de mélange que des stratégies d'optimisation parcimonieuse nous ont permis d'aborder.

B.1.2 Modèles de spectrogrammes et service scientifique

Les modèles de signaux décrits plus haut permettent de fournir un cadre rigoureux pour la séparation de sources. Ils ne sont cependant pas suffisants, car ils n'obtiennent au fond de bonnes performances que si les densités spectrales de puissance (DSP) des sources sont correctement estimées. C'est en effet à ce prix que les sources seront bien séparées.
Pour obtenir de bons résultats de séparation, il est donc nécessaire de se donner une méthode qui permet d'estimer ces DSP — ou leur version "fractionnaire" α-DSP — à partir du mélange seul. À dire vrai, c'est ce point précis qui a fait l'objet de la recherche la plus intense dans la communauté, et il a donc naturellement occupé une part importante de mon travail.
Après avoir été utilisateur pendant plusieurs années de modèles à base de factorisations en matrices non négatives [LBR10, LGS+12, J.Liu12, LBR13], je me suis penché sur deux autres types de modèles.

Modèles nonparamétriques. Partant du constat que les spectrogrammes des sources *harmoniques* et des sources *percussives* ont des allures allongées le long de l'axe du temps et des fréquences, respectivement, D. Fitzgerald a proposé en 2010 de simplement les estimer par un filtrage médian du spectrogramme du mélange, pris selon l'axe correspondant. Cette méthode extrêmement efficace a eu à l'époque l'impact majeur d'une bonne idée simple à implémenter. L'année suivante, Z. Rafii a repris le principe en exploitant cette fois l'idée qu'un accompagnement musical est souvent periodique. Cette idée a dans la foulée été étendue au cas où il est plutôt récurrent, c'est à dire redondant mais pas nécessairement de manière périodique. J'ai moi-même proposé une extension efficace de ces idées qui prend en compte une *pseudo-periodicité* de l'accompagnement plutôt qu'une répétition à l'identique [LRB+12, B.Raf14].

- **La séparation par filtrage médian** s'est donc rapidement imposée comme une technique technique efficace pour séparer des autres le spectrogramme d'une source dont on suppose con-

nues certaines régularités. Il peut s'agir d'une stabilité le long d'un axe, d'une pseudo periodicité, d'une auto-similarité forte, etc. Il me paraissait clair que toutes ces méthodes qui constituaient l'état de l'art de l'époque partageaient des caractéristiques communes, mais le formalisme général qui les réunirait restait à identifier.

- **La régression locale** m'est apparue comme ce cadre théorique. Tel que proposée dans les années 1980, elle consiste à estimer un modèle paramétrique — typiquement polynomial — seulement *localement*, c'est à dire en utilisant de nouveaux coefficients pour chaque position, qui sont estimés en ne considérant que les observations du voisinage. Le type de modèle par lequel approximer localement les observations et la fonction de coût à utiliser pour peuvent être choisis librement.
- **Modèles à noyau pour la séparation.** Une fois ce rapprochement effectué, l'essentiel était fait, mais il restait quelques ingrédients à réunir pour aboutir au modèle présenté en [LRP+14, J.Liu14b]. La première étape a consisté à introduire la notion de noyau de proximité, par laquelle on indique qu'un spectrogramme a une valeur proche en deux points qui peuvent être éloignés dans le temps, comme dans le cas pseudo-périodique. Ensuite, il a fallu montrer comment plusieurs sources pouvaient être estimées conjointement et cela s'est fait en généralisant à ces modèles non locaux les *modèles additifs généralisés* introduits dans les années 80.
- **Extensions.** Le formalisme des modèles additifs à noyau a constitué l'état de l'art en séparation musical pendant deux ans environ. Il a été l'occasion d'un nombre considérable de collaborations, visant à en réduire l'impact mémoire [LFR15], à mettre au point des interfaces utilisateur pertinentes [RLP15], à l'utiliser pour des séparations de gros corpus de musique live [PBLM15, DCLD18], pour une amélioration de la séparation harmonique/percussive [FLR+14], ou encore pour l'isolation des dialogues du fond sonore [KLC15] pour ne citer que quelques exemples.

Réseaux de neurones profonds pour la modélisation audio. En tant que chercheur en traitement probabiliste du signal, le schéma général de mon activité consistait à proposer des caractérisations des signaux qui soient justifiées *instinctivement* et à les traduire par des formalisations mathématiques qui conduisent *in fine* à un lot de paramètres d'une taille assez limitée. Un algorithme permet alors d'estimer ces paramètres et ne nécessite en général pas d'astuce particulière: la méthodologie probabiliste permet de le déduire naturellement du modèle.
Cette méthodologie a été en grande partie balayée — ou tout au moins rendue largement obsolète en matière de performance — par l'arrivée de l'apprentissage profond, dont les premières instances en séparation de la musique datent de 2014. Bien que la qualité initiale obtenue par ces systèmes ait été moins bonne que celle des modèles à noyaux, E. Vincent a bien senti que cela ne durerait pas, et m'a proposé la même année de co-encadrer la thèse de A. Nugraha sur le sujet, me plongeant de fait dans le bain de l'apprentissage profond au moment idéal.

- **Le schéma général** d'une approche de séparation audio exploitant les réseaux de neurones prenait initialement pour acquis le formalisme probabiliste que j'ai exposé plus haut, et sa contribution se bornait à estimer le spectrogramme $\mathbf{V}$ d'une source cible à partir du spectrogramme du mélange $|\ \mathbf{X}\ |$. En pratique, un DNN se présente ainsi comme une fonction paramétrique $\mathbf{V} = f(|\ \mathbf{X}\ |; \Theta)$, où le nombre de paramètres dans Θ est très grand, de l'ordre de la dizaine de millions. Pour utiliser un tel modèle, il suffit de calculer le spectrogramme $|\ \mathbf{X}\ |$ du mélange, puis d'appliquer l'ensemble des réseaux appris, un pour chaque source ou un un modèle conjoint, pour obtenir presqu'instantanément les spectrogrammes des sources, à utiliser pour construire un filtre de Wiener permettant la séparation.
  Avant de pouvoir utiliser le réseau, un bon lot de paramètres doit être identifié, ce qui est appelé *entraînement*. La solution la plus simple dans ce but est de minimiser le risque empirique sur un ensemble d'apprentissage $\{|\ \mathbf{X}\ |_n, \mathbf{V}_n\}_n$, comprenant typiquement un nombre N très grand d'exemples, de l'ordre du million:

$$\Theta^\star \leftarrow \operatorname*{argmin}_\Theta \sum_n \mathcal{L}\left(f(|\ \mathbf{X}\ |_n; \Theta), \mathbf{V}_n\right), \tag{B.1}$$

  où $\mathcal{L}$ est une fonction de coût à minimiser.[n] Tel qu'il est perçu en 2015, ce canevas général soulève un certain nombre d'interrogations sur lesquelles nous nous sommes penchés.
- **Architectures proposées.** On peut sans exagérer dire qu'aujourd'hui une nouvelle architecture est proposée chaque semaine pour la séparation, tandis qu'à l'époque, seuls des réseaux récurrents simples avaient été testés, qui sont similaires à des modèles de Markov. À ce titre, ils souffrent d'une mémoire très limitée. Pour palier à ce problème, nous avons proposé de

simplement prendre comme entrée au modèle des tranches entières de spectrogrammes dont la dimension était réduite par analyse en composantes principales. En sortie, on cherchait à prédire la trame centrale du spectrogramme de la source. Cette approche très simple s'est avérée capable de bien meilleures performances.

- **Combiner DNN et DSP.** Hormis la proposition d'un modèle profond un peu plus performant que l'état de l'art, la véritable contribution de notre travail [J.Nug16] a été d'intégrer l'utilisation des DNN à un formalisme Gaussien rigoureux se présentant comme un algorithme d'espérance maximisation où séparation et réestimation des paramètres sont itérés.
  On peut dire que le modèle résultant et certaines variantes mineures [NLV16, B.Nug18] a véritablement établi un nouvel état de l'art pendant deux ans, supplantant largement le modèle à noyaux.
- **Apprentissage.** Au début de mon travail sur les DNN, je me suis accroché comme beaucoup à l'idée que la fonction de coût $\mathcal{L}$ à utiliser dans (B.1) était importante et pourrait être le point d'accroche d'un éventuel modèle probabiliste. Cela explique que la question occupe autant la partie expérimentale de [J.Nug16].
  Cependant, avec du recul, je dirais que l'importance de la fonction de coût n'est que très secondaire à côté d'autres aspects de l'apprentissage comme la taille du corpus, l'algorithme d'optimisation utilisé et ses hyperparamètres, les différentes stratégies de régularisation, etc.

Pour ce qui est de mes contributions ultérieures sur le sujet des DNN pour la séparation de sources [LSL$^+$19, FNB$^+$19], elles ont été le fruit de collaborations et intègrent assez harmonieusement à mes yeux l'ensemble de mon travail passé : pour effectuer du débruitage, on combine un modèle génératif profond de type auto-encodeur variationnel appris sur du signal de parole propre avec un modèle flexible de type NMF pour le bruit. Au moment du débruitage, on estime à la fois et de manière itérative le code latent d'entrée du modèle génératif et les facteurs de la décomposition, dans un cadre probabiliste α-stable.

Service et médiation scientifiques. Comme je l'ai montré succinctement, la première phase de mon travail sur les modèles profonds s'est traduite par des contributions de modèles et algorithmes. Cependant, l'aspect colossal de ce travail mené par A. Nugraha m'a fait sentir que je ne parviendrai pas à maintenir de cette manière une contribution notable, compte tenu de mes moyens limités et du temps nécessaire à une seule étude d'un niveau irréprochable.
Lorsqu'en 2014, N. Ono me fait l'honneur de me proposer de présider la campagne internationale d'évaluation de séparation de sources (SiSEC), je perçois l'opportunité: il s'agit du cadre idéal pour mener à bien l'ambition de jouer un rôle déterminant dans cette communauté, par le biais d'un *service scientifique* maintenu sur le long terme.

- **Challenges.** SiSEC est une campagne internationale d'évaluation lancée en 2007 et organisée en conjonction avec la conférence LVA-ICA et j'y suis fortement impliqué depuis 2010. Son but est de comparer les performances des algorithmes de séparation de sources sur la base de données communes. Bien qu'elle se soit toujours voulue multidisciplinaire, c'est surtout dans le domaine de la séparation de musique que son succès ne s'est pas démenti, avec une participation croissante. Lorsque j'en prends les rennes en 2015, il y a 20 systèmes différents soumis pour la tâche de démixage musical.
  Cependant, SiSEC est à un tournant. On passe d'une communauté centrée sur des méthodes à base de modèles à une nouvelle ère exploitant largement l'apprentissage automatique. Tout est à faire pour fournir à la communauté du démixage musical les outils nécessaires pour y rentrer de plein pied. Je veux souligner ici l'importance déterminante de F. Stöter dans ce travail.
- **Les données** que nous considérions avant l'avènement de l'apprentissage profond n'étaient cruciales que pour l'évaluation des méthodes, et sur ce point, le corpus utilisé par SiSEC commençait déjà à montrer ses limites : jusqu'en 2013, l'ensemble de test comportait une dizaine d'extraits musicaux, pour un total d'environ 2 mn 30 de musique.
  Lorsque j'ai proposé les modèles additifs à noyau [J.Liu14b, LFR15], j'étais heureux de constater ce que je percevais comme une grande robustesse à la fois au cours d'un morceau, mais d'un morceau à l'autre. Cependant, les données alors disponibles ne permettaient pas de le montrer : il y avait clairement eu un effet de *surapprentissage* sur le corpus SiSEC. C'est pour cela que j'ai préparé le corpus `ccmixter`, composé de 50 morceaux complets, pour lesquels on disposait de la voix et de l'accompagnement séparés [S.Liu14]. Il s'agissait du premier corpus

de morceaux complets pour la séparation, d'un niveau musical semi-professionnel.
Avec l'arrivée de l'apprentissage profond, je n'ai pu que constater l'absence criante d'un corpus commun qui permettrait de faire de l'apprentissage. Il était en effet déjà assez clair que ce ne serait qu'à ensemble d'apprentissage égal que l'on pourrait vraiment comparer les différentes approches. Après des mois d'un travail de mixage acharné, nous étions fiers de proposer DSD100 à la communauté pour SiSEC2016 [LSR+17], que nous avons encore enrichi pour finalement parvenir à MUSDB18 [S.Raf17]. Il s'agit aujourd'hui du standard *de fait* dans le domaine, téléchargé environ 10 000 fois et dont le succès ne se dément pas.

- **Logiciels.** Le besoin de données de qualité n'était pas le seul aspect sur lequel la communauté de la séparation de sources était en demande. Son écosystème logiciel aussi était en mutation. Alors que l'essentiel des chercheurs travaillait en Matlab par le passé, une transition massive et soudaine vers Python s'est opérée en quelques années. Par le biais de la communauté sigsep.github.io, F. Stöter et moi-même avons offert aux chercheurs intéressés une quantité importante de ressources logicielles diverses librement utilisables.
  - Les métriques `BSSeval` d'évaluation de la qualité de la séparation ont été établies il y a une dizaine d'années. Bien qu'elles soient critiquables sous certains aspects, elles constituent le standard que chaque article du domaine a coutume d'utiliser. Le fait qu'aucune implémentation officielle n'était disponible en Python était préjudiciable à la bonne marche de la recherche.
    En tant qu'organisateurs de SiSEC nous avons remédié à ce problème en proposant `musdb`, dont la totale rétrocompatibilité sur la version Matlab est garantie et qui intègre également certaines accélérations importantes [S.Sto19b, SLI18].
  - Le module `musdb` permet un interfaçage natif avec MUSDB18 [S.Sto19a].
  - Le module `norbert` implémente le filtrage de Wiener en stéréo [S.Liu19].
- **open-unmix**. Pour compléter le panel des logiciels qui manquaient, nous nous sommes attelés à produire une implémentation en licence libre d'une méthode de séparation de sources par apprentissage profond, qui soit d'un niveau de maintenance professionnel. Ce travail s'est fait en collaboration avec Sony. Le but était d'éviter toute innovation en terme de modèle (nous avons choisi un réseau de type BLSTM), mais de s'attacher à mener ce modèle standard au bout de ses capacités par une implémentation de qualité. Au bout d'un an d'efforts, nous avons publié `open-unmix` ainsi que les poids des modèles pré-appris sur MUSDB18. Ce logiciel a été salué par un prix au hackathon annuel PyTorch en 2019 [J.Sto19].
- **Campagnes d'évaluation.** Pour finir, nous avons fortement modernisé le déroulement de SiSEC. Jusqu'à 2013, les participants envoyaient aux organisateurs les résultats de leur séparation, sous un format libre. C'était aux organisateurs de procéder aux évaluations, ce qui prenait un temps déraisonnable, compte tenu des particularités de chaque soumission. En 2015-2016, j'ai fourni des scripts que les participants pouvaient utiliser pour générer leurs résultats, dans une démarche d'évaluation reposant sur la confiance [ORK+15, LSR+17]. En 2018, enfin, chaque participant devait créer une *pull request* sur un dépôt consacré, après avoir exécuté une routine d'évaluation intégrée à `museval` [WMK+18].
  Nous avons embarqué tous les scores obtenus par les participants de SiSEC2018 au module d'évaluation (`museval`) lui-même, ce qui permet à quiconque d'immédiatement pouvoir comparer les performances de sa méthode avec celle de tous les participants de SiSEC2018.

L'ensemble de ce travail de service scientifique n'est certes pas de la recherche fondamentale, mais j'aime à croire qu'il a joué un rôle au moins aussi important en permettant à de nombreux chercheurs de travailler dans un cadre de travail plus ouvert et reproductible, ainsi qu'en réduisant la phase d'expérimentation à son minimum, qui est d'évaluer les techniques que l'on propose soi même, tout en fournissant un canevas de confiance et de qualité pour se comparer à l'état de l'art. L'expérience accumulée tout au long de ce travail a été valorisée sous la forme de tutoriaux aux conférences ISMIR'18 et EUSIPCO'19, ainsi que sous la forme d'articles d'overview invités [J.Raf18, J.Can19] et d'un chapitre d'introduction au sujet [B.Par18].

B.1.3 Apprentissage automatique

L'essentiel de la recherche présentée ci dessus a été appliquée au traitement du signal audio et musical. Cependant, mes centres d'intérêt en tant que chercheur ne se sont pas limités à ce domaine. Je résume ici les travaux que j'ai menés en apprentissage automatique et en traitement du signal, dont l'audio n'est pas l'application principale.

Les factorisation en matrices non-négatives sont de puissants modèles explicatifs permettant d'extraire la redondance de données bruitées. Après en avoir été un utilisateur, j'en ai proposé certaines extensions liées aux modèles α-stables que j'ai déjà évoqués.
En collaboration avec plusieurs chercheurs, je me suis tout d'abord intéressé aux cas particuliers Cauchy et Levy [LFB15, MBL17a, MBL17b]. Les premières rendent compte de données réelles ou complexes très bruitées, tandis que les secondes permettent de décomposer des matrices nonnégatives.
Une approche très générale compatible avec n'importe quel exposant caractéristique $\alpha \in (0, 2)$ a été présentée dans [SEL$^+$18, J.Sim15]. Il est intéressant de noter qu'elle permet d'également inférer l'exposant à utiliser en fonction des données.

Parcimonie et acquisition compressée materielle. Après ma thèse, j'ai travaillé pendant une année auprès de chercheurs en optique non-conventionnelle, dont le but était de mettre au point un nouveau système d'imagerie optique, visant à réduire de manière drastique le nombre de capteurs à qualité équivalente. Le système est présenté en détail en [J.Liu14a] : il s'agit d'une implémentation matérielle du principe de *l'acquisition compressée*.

- **Optique non conventionelle en milieu fortement diffusants.** Lorsque la lumière pénètre dans un milieu fortement diffusant tel qu'une fine couche de papier ou de peinture, elle est soumise à une très grande quantité de diffractions, de telle manière que de l'autre côté du milieu, la lumière qui sort en chaque point est un mélange très complexe de la lumière qui rentre. Plus le milieu est *diffusant*, plus le mélange sera intense.
  Exploiter de tels milieux diffusants pour de l'imagerie peut paraître étrange dans la mesure où ce qu'on observe en sortie est un *chatoiement* qui semble aléatoire, quelle que soit l'image en entrée. On s'éloigne en effet fortement de l'imagerie classique, dont l'objectif est au contraire d'être parfaitement transparente et calibrée.
  Quoiqu'il en soit, lorsque j'ai commencé mon travail, il était déjà acquis que de tels milieux fortement diffusant se ramènent à appliquer au signal d'entrée, mettons $\mathbf{x}$, une transformation linéaire pour obtenir en sortie un front d'onde $\mathbf{y}$ donné par :

$$\mathbf{y} = \mathbf{H}\mathbf{x}, \tag{B.2}$$

  dont le chatoiement observé est l'énergie $|\ \mathbf{y}\ |^2$. La *matrice de transmission* $\mathbf{H}$ de tels milieux a la particularité de se comporter comme une matrice aléatoire dont tous les coefficients sont indépendants, pour peu qu'on considère des capteurs suffisamment éloignés.
- **Acquisition compressée matérielle.** La grande idée de l'équipe menée conjointement par L. Daudet et S. Gigan sur ce sujet était qu'un tel scénario a des similitudes frappantes avec le sujet de l'acquisition compressée, développé dans les années précédentes. Son principe fondamental est que si le signal original $\mathbf{x}$ dans (B.2) présente une *structure* qui se traduit comme une parcimonie dans une base connue, alors il est possible de le reconstruire à partir d'un nombre d'échantillons dans $\mathbf{y}$ qui est très faible, dès lors que la matrice de transmission $\mathbf{H}$ est suffisamment "mélangeante". Plus précisément, le nombre de capteurs nécessaire ne dépend pas de la taille initiale du signal, mais plutôt de sa complexité, c'est à dire du nombre de coefficients requis pour le décrire. Au fond, ce que cela signifie est que si chaque mesure contient de l'information d'un peu partout sur l'entrée, alors on peut la reconstruire si elle est suffisamment régulière.
  L'expérience que nous avons réalisée a consisté à contrôler le front d'onde optique $\mathbf{x}$ pénétrant dans un milieu diffusant par le biais d'une grille de 1024 micro-miroirs, puis à en mesurer la sortie (B.2) en combinant plusieurs mesures selon un *modus operandi* qui m'échappe. Quoiqu'il en soit, le but était de reconstruire le front d'onde d'entrée à partir d'un nombre limité de mesures en sortie.
  Après plusieurs mois d'expérimentations à la fois en chambre noire et en calculs, nous avons finalement eu la joie de parvenir à reconstruire les motifs que nous donnions en entrée, avec des nombres de mesures qui correspondent à ce qui était prédit par la théorie. Pour la première fois, on réalisait de l'acquisition compressée à la vitesse de la lumière, malgré un bruit assez important [LMGD14]. Je m'en souviens comme d'un moment assez fort de ma carrière.

Modèles génératifs nonparamétriques. Le paradigme du filtrage que j'avais considéré jusqu'à présent pour traiter les signaux est limité dans le sens où il ne peut que retirer de l'information des signaux bruts et est incapable d'en ajouter, comme cela serait par exemple nécessaire pour reconstruire des signaux dégradés par une acquisition ancienne et bruitée.
Intéressé par les réseaux antagonistes génératifs (GAN) du fait de leur évident potentiel dans tous les articles où je les ai vus utilisés, ce n'est que lorsque j'ai compris leur lien avec la théorie du transport optimal que je me suis véritablement intéressé à la question.

- **Transport optimal et modèles génératifs**. Le problème que se pose cette branche relativement ancienne des mathématiques datant du XVIII$^{\text{ème}}$ siècle est d'étudier dans quelle mesure des échantillons d'une distribution *source* μ peuvent être transformés en échantillons représentatifs d'une autre distribution *cible* ν, toutes deux mesurant un ensemble $\Omega \subset \mathbb{R}^D$. Il s'agit bien d'une manière appropriée de formaliser un problème génératif, puisque μ peut par exemple être une Gaussienne dont on peut facilement tirer des échantillons, tandis que ν est l'ensemble des données que l'on souhaite "imiter", des images par exemple.
- **Le transport optimal en bref.** Dans ce cadre formel, deux types d'objets apparaissent. Le premier est le *plan de transport*. C'est fonction $T : \Omega \to \Omega$ qui indique vers quel point cible transporter chaque point source. On souhaite que si $\mathbf{x} \sim \mu$, alors $T(x) \sim \nu$. Cela s'écrit: $T\sharp\mu = \nu$. Grâce au travail de L. Kantorovich et Y. Brenier, il est acquis qu'une telle fonction existe dès lors que μ est bien conditionnée. Par contre, la preuve n'est pas constructive.
  Le deuxième type d'objet qui apparaît est une fonction de distance d'un nouveau type, appelée distances de Wasserstein, qui permettent de juger de la difficulté du transport entre deux distributions. Si $\mathcal{W}(\mu, \nu)$ est élevé, c'est que le plan de transport entre elles conduira à de nombreux déplacements, tandis que s'il est faible, c'est qu'elles sont déjà presqu'identiques. Leur grand intérêt est d'être mieux conditionnées que d'autres distances comme Kullback-Leibler.
- **Dans le cas scalaire** où μ et ν s'appliquent donc sur $\mathbb{R}$, $\mathcal{W}(\mu, \nu)$ se calcule simplement comme la norme des écarts entre leurs quantiles (minimum vs minimum, median vs median, maximum vs maximum, etc). Le plan de transport lui aussi s'exprime simplement dans ce cas comme associant chaque quantile de μ au quantile équivalent de ν. Dans le cas général, à la fois le plan de transport et la distance de Wasserstein sont difficiles à identifier, ce qui motive un large effort de recherche.
- **La distance de Wasserstein projetée** permet de se ramener à ce cas scalaire même pour des espaces de haute dimension. L'idée qui la sous-tend est de se donner un grand nombre de vecteurs unitaires $\boldsymbol{\theta}$, donc disposés sur la sphère $\mathbb{S}^D$, et de simplement projeter les distributions μ et ν par simple produit scalaire, pour obtenir des distributions dans $\mathbb{R}$: $\boldsymbol{\theta}^\star\sharp\mu$ et $\boldsymbol{\theta}^\star\sharp\nu$ que l'on peut alors facilement comparer par le calcul de leur distance de Wasserstein $\mathcal{W}(\boldsymbol{\theta}^\star\sharp\mu, \boldsymbol{\theta}^\star\sharp\nu)$. N. Bonnotte prouve que si on somme ces distances scalaires sur l'ensemble de la sphère, on obtient une distance équivalente à la distance de Wasserstein dans l'espace original. Elle est notée $\mathcal{SW}(\mu, \nu)$, comme *sliced-Wasserstein*:

$$\mathcal{SW}(\mu, \nu) = \int_{\boldsymbol{\theta}} \mathcal{W}(\boldsymbol{\theta}^\star\sharp\mu, \boldsymbol{\theta}^\star\sharp\nu) d\boldsymbol{\theta} \tag{B.3}$$

- **Sliced Wasserstein flows.** Après cette introduction, je peux décrire la contribution principale de notre article récent [LŞM$^+$19], qui consiste à exploiter la distance de Wasserstein projetée pour itérativement déplacer un ensemble de points de μ vers ν. L'algorithme, appelé *Sliced Wasserstein Flow* est inspiré de *l'Iterative Distribution Transfert* de N. Bonnotte, mais le généralise en l'intégrant dans un cadre mathématique plus rigoureux.
  L'essence de l'algorithme est la suivante: on commence par tirer une population de quelques centaines de points dans μ, puis on se donne un ensemble de vecteurs unitaires $\boldsymbol{\theta}$. Pour chacun, on calcul les quantiles empiriques des distributions projetées $\boldsymbol{\theta}^\star\sharp\mu$ et $\boldsymbol{\theta}^\star\sharp\nu$. Enfin, on met à jour les particules par un pas calculé comme une moyenne de tous les vecteurs $\boldsymbol{\theta}$, chacun pondéré par le déplacement nécessaire pour faire correspondre $\boldsymbol{\theta}^\star\sharp\mu$ et $\boldsymbol{\theta}^\star\sharp\nu$. L'algorithme est très simple à implémenter, massivement parallélisable, et donne des résultats très intéressants dès lors que la dimension des signaux n'est pas trop élevée.

Sans rentrer plus avant dans les détails techniques, je dirais que ce travail a été le fruit d'une riche collaboration. J'ai initialement testé l'algorithme et mené les expériences qui s'avéraient concluantes, mais le cadre théorique dépassait mes compétences mathématiques. C'est grâce au travail de mes collaborateurs que nous sommes arrivés à la formulation qu'on trouve dans [LŞM$^+$19].

Intégration de la position dans les modèles Transformer. La dernière contribution dont je veux parler illustre assez bien à mes yeux comment le traitement probabiliste du signal peut apporter des contributions intéressantes à l'apprentissage automatique.

- **Le Transformer** a été introduit il y a quelques années comme un nouveau type d'architecture profonde visant à traiter des séquences. Son nouvel ingrédient principal est la couche *d'attention*, qui calcule chaque élément $\mathbf{y}_m$ de la sortie comme une combinaison linéaire de toutes les entrées $\mathbf{v}_n$, pondérées par un coefficient dit *d'attention* a_{mn}:

$$\mathbf{y}_m = \sum_n a_{mn} \mathbf{v}_n. \tag{B.4}$$

  La particularité de l'approche est de calculer cette attention a_{mn} comme un produit scalaire entre des features $\mathbf{q}_m$ et $\mathbf{k}_n$. Tous ces vecteurs dits de *valeurs* $\mathbf{v}_n$, de *clés* $\mathbf{k}_n$ et de *requêtes* $\mathbf{q}_m$ sont typiquement obtenus à partir de la séquence d'entrée $\mathbf{x}_n$ par simple transformation linéaire: $\mathbf{v}_n = \mathbf{V}\mathbf{x}_n$, $\mathbf{k}_n = \mathbf{K}\mathbf{x}_n$ et $\mathbf{q}_m = \mathbf{Q}\mathbf{x}_m$. Ce sont les coefficients de ces matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ qui constituent les paramètres. En empilant ce mécanisme d'attention plusieurs fois, les Transformers ont démontré des performances révolutionnaires qui marquent un nouvel état de l'art.
  En guise d'explication du procédé, je dirais que les réseaux convolutifs et récurrents classiques choisissent *a priori* les échantillons de la séquence d'entrée qu'il faut utiliser pour calculer la sortie: le voisinage immédiat et les symboles précédents dans le temps, respectivement. Ce qui y est appris est alors le traitement, plus ou moins sophistiqué, qu'il faut appliquer à ces voisins. Le cas du Transformer est différent: le traitement est trivial et se ramène à une simple moyenne pondérée. Par contre, les points à considérer pour ce calcul sont identifiés d'une manière non-locale comme une proximité entre features.
  Dans sa version initiale, le mécanisme d'attention a une complexité quadratique, puisqu'il nécessite le calcul d'une large matrice $\mathbf{A} \equiv [a_{mn}]_{mn}$. Des variantes récentes ont permis de tomber sur une complexité linéaire en la factorisant comme $\mathbf{A} \approx \phi(\mathbf{Q})^T \phi(\mathbf{K})$ où ϕ est une fonction non-linéaire.
- **L'encodage de la position.** Sans une subtilité supplémentaire, on voit que le calcul (B.4) de la sortie d'un Transformer ne prend pas en compte les *positions* n/m des entrées/sorties. Pourtant, ces positions revêtent une grande importance, au même titre que le signal proprement dit. C'est pour cette raison qu'a été introduit *l'encodage de la position*, comme une information supplémentaire ajoutée aux clés $\mathbf{k}_n$ et requêtes $\mathbf{q}_m$ avant le calcul de l'attention et qui dépend des positions n et m, respectivement.
  Une variante plus récente et efficace, dite d'**encodage relatif de la position** rajoute ce terme supplémentaire dépendant de la position directement à la matrice d'attention:

$$\mathbf{A} \leftarrow \mathbf{Q}^\top \mathbf{K} + [\mathcal{P}(m,n)]_{mn}\,, \tag{B.5}$$

  ce qui permet notamment d'exploiter les positions relatives par un schéma d'encodage de type $\mathcal{P}(m-n)$, souvent plus pertinent qu'une exploitation des positions absolues.
- **Encodage stochastique de la position.** Une piste d'amélioration que nous avons identifiée est que les Transformers à complexité linéaire mentionnés plus haut étaient encore incompatibles avec un encodage relatif de la position. Dans [LCW+21], nous montrons qu'une approche d'encodage *stochastique* de la position est possible, qui est équivalente à l'approche (B.5).
  L'idée principale parait assez naturelle du point de vue du traitement du signal: le calcul $\mathbf{A} = \mathbf{Q}^\top \mathbf{K}$ de la matrice d'attention peut au fond être vu comme le calcul d'une covariance empirique entre clés et requêtes. Il suffit donc de faire en sorte que ces signaux soient *stationnaires* pour que la matrice d'attention correspondante ait la forme de Toeplitz désirée. On peut ainsi garantir une attention relative, sans jamais calculer explicitement cette matrice $\mathbf{A}$.

B.2 Programme de recherche

B.2.1 Introduction

Une carrière de chercheur s'inscrit dans la dynamique d'une communauté scientifique. Il y a des pans de ma recherche passée qui ne sont plus autant d'actualité qu'ils l'ont été. Par exemple, les traitements profonds de bout en bout se sont imposés, qui consistent à apprendre non seulement le module de traitement non linéaire du signal, mais aussi la représentation dans laquelle ce filtrage doit se faire, ainsi que la procédure complémentaire de reconstruction. Dans ce contexte, les différents travaux que j'ai pu mener sur les modèles temps-fréquence ont perdu de leur actualité. De la même manière, les modèles à noyaux pour la séparation à base de régression non locale ont été largement dépassés par les réseaux de neurones. D'une manière générale, je peux aujourd'hui dire que le démixage musical, qui m'a tant occupé, est devenu une technologie mature. C'est d'ailleurs pour cette raison que j'ai également passé du temps ces dernières années à conduire à leur terme un certain nombre de projets de transfert industriel sur le sujet.
Cependant, j'estime qu'une large part de cette activité passée est encore d'actualité ou tout au moins peut être prolongée de manière intéressante, c'est ce qui constitue le cœur de mon projet de recherche, que je vais brièvement résumer ici.

Pour ce qui est du non technique, je suis convaincu avec du recul que mon travail en service scientifique à la communauté est probablement celui qui aura le plus d'impact à long terme. De la même manière, les collaborations et l'enseignement ont une portée non-mesurable que j'estime supérieure à toutes les contributions individuelles que j'ai pu ou que je pourrai proposer. C'est ce qui me pousse bien-sûr à vouloir obtenir l'habilitation qui me permettra de diriger des recherches.

B.2.2 Questions et pistes théoriques

Sur le plan des recherches théoriques, je m'intéresse à la modélisation probabiliste des signaux, qui intègre pleinement les dernières avancées en modélisation profonde, mais qui bénéficie d'une interprétation et fondation théorique solides. J'identifie plusieurs directions porteuses en ce sens.

- **Encodage de la position et modèles non structurés.** L'arrivée des Transformer a régénéré l'intérêt que je porte à la régression non-paramétrique, qui constitue le cœur de mon travail sur les modèles additifs à noyaux. À mes yeux, le formalisme des Transformers apporte un regard nouveau sur ces approches.
  Je vois deux pistes à explorer à court terme sur ce point. Tout d'abord, il serait intéressant de considérer des encodages de position qui soient dépendants du signal proprement dit. En effet, l'approche actuelle revient au fond à supposer qu'un intervalle de temps ou de position donné aura la même valeur quel que soit le signal d'entrée, ce qui est intuitivement un non-sens pour moi, et explique à mes yeux que seuls des contextes assez courts soient au fond exploités par ces modèles parce que ce sont les seuls qu'il est sûr de considérer quel que soit le signal à traiter. Ensuite, je pense qu'on peut généraliser le principe de l'attention pour l'appliquer au niveau structurel, et non pas au niveau du signal. Cela signifie que les connexions entre les différents "neurones" composant le système de traitement pourraient elle-même dépendre de leur position et du signal qu'ils portent.
- **Modèles de diffusion et équations différentielles neuronales.** Considérant que c'est la suite logique de mon travail sur les SWF, je m'intéresse de près aux modèles de diffusion, qu'il s'agisse des équation différentielles neuronales ou des modèles de diffusion de type *wavegrad* proposés récemment, qui modélisent des signaux comme le processus inverse d'une opération progressive de "blanchiment".
  Compte tenu du fait qu'une grande difficulté de tels modèles est leur apprentissage, je veux étudier dans quelle mesure des approches alternatives de type "boucle directe aléatoire de l'erreur" (direct feedback alignment) permettrait de passer à l'échelle en terme de complexité des modèles.
- **Modèles alpha-stables pour l'optimisation.** Plusieurs travaux récents ont montré que lors d'une optimisation par descente de gradient stochastique, la trajectoire suivie par les millions de paramètres d'un réseau de neurones est plus fidèlement modélisée comme une marche stable que par le modèle classique d'un mouvement Brownien.
  Dans ce contexte, je propose deux pistes de recherche. La première est d'exploiter mon travail passé en modélisation vectorielle α-stable pour étudier dans quelle mesure il permettrait d'identifier des "cliques" de paramètres évoluant de manière conjointe. La deuxième est inspirée

de l'intrigante "théorie du ticket gagnant" (lottery ticket hypothesis) qui montre qu'un réseau de neurones initialisé aléatoirement comprend des sous-réseaux optimaux. L'idée est alors de tenter d'identifier ces sous réseaux par des techniques inspirées de mon travail en localisation de sources. Dans les deux cas, un des enjeux principaux est *l'échelle*, puisqu'un réseau de neurones classique comprends des dizaines, voire des centaines de millions de paramètres.

B.2.3 Applications

Bien que la recherche qui m'intéresse soit avant tout axée sur les contributions théoriques, c'est toujours avec grand intérêt que je m'intéresse aux applications.

Applications en audio. Le traitement du signal audio est l'application principale qui m'a intéressée par le passé. Bien qu'il est possible que son importance dans ma recherche diminue avec les évolutions des différentes contributions que je compte avoir, il reste un sujet qui m'intéresse beaucoup, à la fois parce que je suis passioné de musique et parce qu'il s'agit de signaux extrêmement complexes qui permettent d'illustrer de nombreuses contributions théoriques.

- **La séparation de musique** reste un de mes thèmes d'expertise. Elle m'apparaît en particulier comme un terrain de jeu intéressant pour illustrer mon travail sur l'encodage de la position, et celui sur de nouveaux modèles de formes d'onde.
- **L'amélioration audio** me semble un thème de recherche porteur qui prolonge mon travail passé sur la théorie du filtrage. Appliquer de nouvelles méthodes génératives pour la restauration de vieux enregistrements ethnomusicologiques et la conversion de musique me paraissent des pistes d'application prometteuses pour un travail sur les modèles génératifs.
- **La génération de musique**, enfin, apparaît comme un objectif très prometteur. Je ne suis pas spécifiquement intéressé par la génération inconditionnelle de contenu entièrement nouveau, que j'estime assez loin des cas d'usage, mais plutôt par des stratégies *informées*, qui s'apparenteraient à de "l'agrégation" musicale automatique.

Applications en sciences du vivant. J'ai rejoint l'université de Montpellier en 2017 dans le but de diversifier mes domaines d'application. Le contexte universitaire dans les domaines des sciences du vivant y est en effet réputé et un de mes objectifs affichés est de développer mon activité dans ces domaines. Les points suivants sont la suite logique de deux collaborations principales déjà formalisées par des projets financés par l'université de Montpellier.

- **Modèles probabilistes pour la génétique.** En collaboration avec G. Krouk, je m'intéresse à l'utilisation de méthodes d'analyse de données pour deux applications à fort impact. La première est la prédiction de traits phénotypiques de plantes à partir de mesures d'activité génique effectuées au niveau des feuilles. Il s'agit d'un problème inverse qui permettrait de prédire de manière non invasive dans quel état est le système racinaire d'une plante. La réussite d'un tel projet aurait des conséquences intéressantes en agronomie.
  Par ailleurs, je m'intéresse à la question de la prédiction des interactions entre protéines, qui est un des mécanisme fondamental pour la mise au point de nouveaux médicaments, par exemple. Le but est de construire des modèles profonds de type Transformer pour manipuler les séquences d'acides aminés par lesquelles sont définies les protéines.
- Des chercheurs montpellierrains en **écologie fonctionnelle** ont embarqué des capteurs audio et GPS sur des animaux lors de campagnes terrain, pour mieux étudier leur comportement à l'état sauvage. Ces données soulèvent de nombreux challenges intéressants. Tout d'abord, leur manipulation non-supervisée à des fins exploratoires, visant à identifier des éléments rares ou au contraire communs d'un intérêt scientifique soulève des questions intéressantes en terme de visualisation de données massives. Par ailleurs, les positions des animaux au cours du temps me semblent des données parfaites pour des modélisations de type différentielles, où l'objectif pourrait d'apprendre des champs de déplacement en fonction des coordonnées GPS.

C Collaborations and supervision

C.1 Academic productive collaborations

I wrote around 90 peer-reviewed papers with approximately 100 co-authors in total. These numbers show that I am deeply convinced of the numerous advantages of collaborative research and it would be a bit lengthy to describe all the productive collaborations I had over the years. To provide some illustration still, Figure C.1 summarizes with which individuals, institutions and countries I collaborated the most in terms of numbers of papers. I quickly detail these main collaborations wrt my different main research directions.

Probabilistic models for audio. I have collaborated a lot on this topic with my former Ph.D. supervisors Roland Badeau and Gaël Richard (Telecom Paris), even long after my defense. Mathieu Fontaine, that I supervised on this topic during his Ph.D. is of course also a strong collaborator. My work with Alexey Ozerov (now Interdigital) mostly concerned the connections between source coding and source separation.

Occasional and regular collaborations on probabilistic models have also been done with Laurent Daudet (now lighton.io) as well as Laurent Girin (GIPSA lab).

Deep learning research has mostly been done with Fabian Robert Stöter, Aditya Arie Nugraha and Emmanuel Vincent.

International evaluation campaigns have lead me to worldwide collaborations, notably with researchers from Japan (Y. Mitsufuji, N. Ono), Germany (F. Stöter) and USA (Z. Rafii), excluding all those dozens of participants that I had close interaction with, but that were not author of the final papers.

Physics research regarding non-classical optics has been mostly done with Laurent Daudet, David Martina and Angélique Drémeau (all at Institut Langevin back then).

Machine Learning research I am doing is mostly in collaboration with Umut Simsekli (now Inria), Taylan Cemgil (Bogazici Univ.) and Gaël Richard.

C.2 Stays abroad

I was invited for doing academic stays abroad by several institutions.

Bogazici University (2014, 2 months) I was lucky to be invited by T. Cemgil at Bogazici Univ. (Istambul) to work on multimodal processing [LŞC15]. The longer-term outcome of this stay is the deep and sustained collaboration I still have today with U. Simsekli [SEL+18, J.Sim15, LCW+21, LŞM+19].

Northwestern University (2015, 2 weeks). B. Pardo invited me to work on audio source separation in his team at Northwestern University. The outcome was a lot of subsequent collaborations on kernel additive modelling, overview papers [B.Par18] and modulation-based representations [PPL17].

Kyoto University (2017, 2 months). K. Yoshii invited me to work with him on probabilistic signal processing at the Kyoto University. This notably lead to a new faster method for full-rank modeling [LY17], but also initiated a long-term and sustained collaboration. The two Ph.D. students I directily supervised joined him for their postdoc.

RWTH Aachen (2018, 1 week.) C. Rohlfing invited me to work on non-Gaussian audio models for applications to audio coding. The BEADS model came out of this stay [LRD18].

C.2.1 Invited seminars

I gave official invited talks at research labs in the following cities: Paris (France), Kyoto (Japan), Istambul (Turkey), Erlangen (Germany), Wadern (Germany), Nancy (France), Montpellier (France), Madrid (Spain), San Francisco (USA).

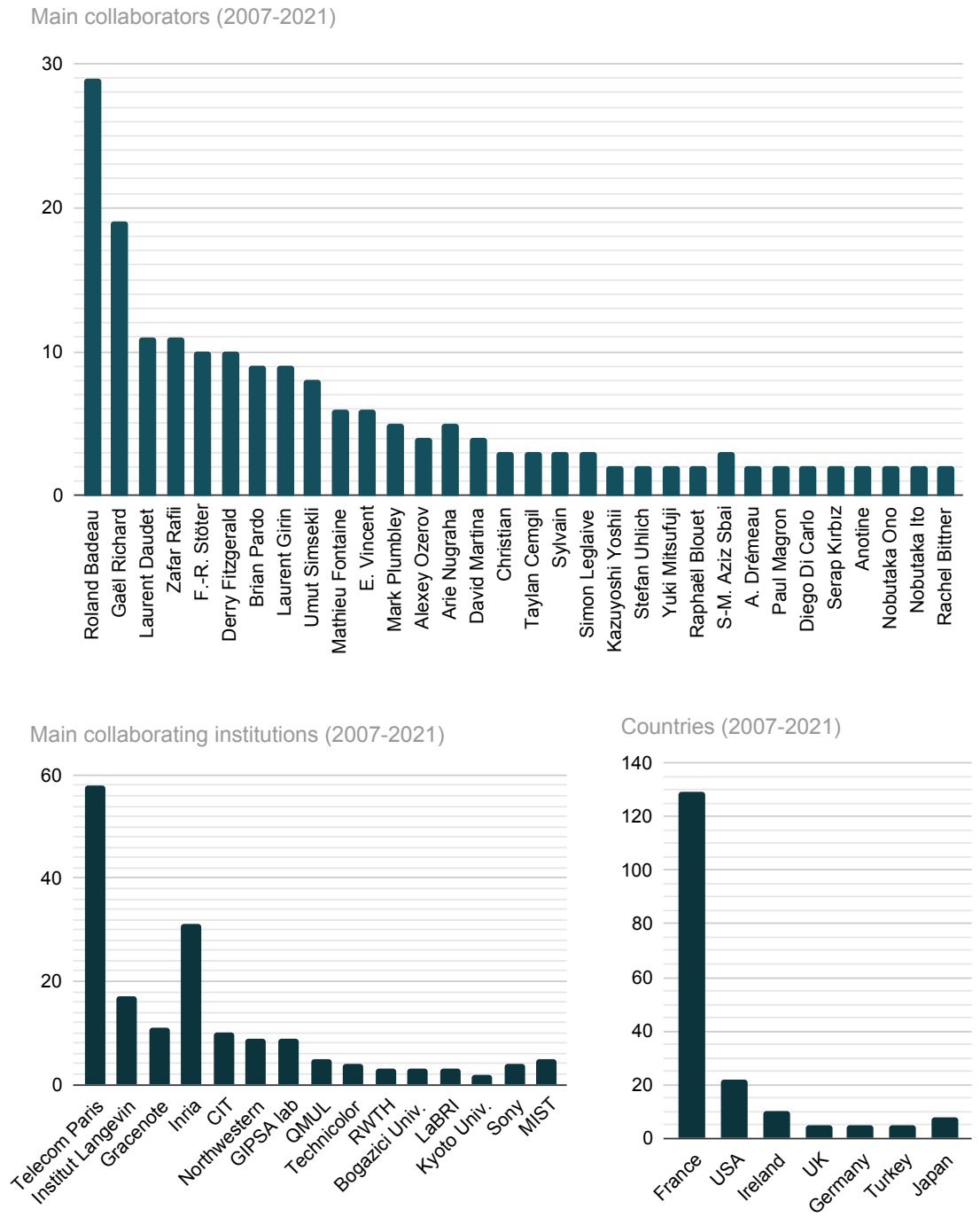


Figure C.1: Summary of my collaborations. I have a total of approximately 100 collaborators. I only display those with more than 2 papers. Top: number of papers I have co-authored with my main collaborators. Below: corresponding institutions and countries.

C.3 Supervision

I supervised the following master students during long internships:

- **Si-Mohamed Aziz Sbai** (2007. ENSEIRB Bordeaux). Source separation, wavelet analyzis/synthesis [P.Sba10, P.Blo10, RBL07]. He is now senior researcher at Apple, California.
- **Diego Di Carlo** (2017, Master computer engineering, Padova, IT). Gaussian processes, source separation for massive datasets [DCDL17, DCLD18]. He is now postdoc at Univ. Rennes 2.
- **Juliette Philibert** (2021, Informatique, spécialité IA+ML, Aix-Marseille). Neural ordinary differential equations.

C.3.1 Ph.D students

I have been so far the official co-supervisor of two Ph.D students.

- **Aditya Arie Nugraha (2015-2017)** on the topic of deep learning applied to source separation [J.Nug16, NLV16, B.Nug18, SNV$^+$15]. He is now researcher at Kyoto University.
- **Mathieu Fontaine (2016-2019)** on the topic of α-stable models for signal processing [FVLB17b, FVLB17a, FLGB17, J.Fon20]. He is now postdoc at Kyoto University.

I am fortunate to have regular and ongoing collaborations with both [FNB$^+$19]. Appart from these Ph.D. students that I officially supervised, I engaged in strong collaborations with several Ph.D. students, that served as a *de facto* co-supervision.

- **Christian Rohlfing (2017)** on the topic of informed source separation [RLB17, RCL17].
- **Paul Magron (2017)** on the topic of nonnegative stochastic modeling [MBL17b, MBL17a].
- **Simon Leglaive (2018)** on the topic of α-stable multichannel models [LSL$^+$17, SEL$^+$18].

C.3.2 Other teaching activities

- **Teaching**. I gave approximately 400 h of courses at the master level in different topics: analog signal processing (60 h), digital signal processing (200 h), computing (100 h).
- **Short supervisions.** I supervised many short projects for students in engineering schools, on the topic of audio, music and signal processing.
- **Tutorials.** I gave tutorials at ICASSP 2014, ISMIR 2018, EUSIPCO 2019.
- **Continuous training.** I regularly teach deep learning in continuous training programmes.

C.4 Research grants

I participated in many funded research projects, but also directly took part in the funding of some:

- **ANR JCJC Kamoulox. PI.** (2016-2020). 300k€
- **Co-financement région Lorraine. PI.** (2016). For the Ph.D of M. Fontaine. 50k€.
- **MUSE AI3P. Part of consortium.** (2021-). University of Montpellier. 260k€
- **MUSE REPOS. Part of consortium.** (2021-). University of Montpellier. 390k€.

C.4.1 Industrial transfer

I was involved in the transfer of source separation technology to companies in Europe and in the USA, for a total of several hundreds of thousands euros. Everything about these is confidential and ruled by non-disclosure agreements.

D Selection of papers

In this section, I selected some papers that illustrate my research curriculum.

The invited overview journal paper [J.Raf18] is reproduced on page 104. It is a very thorough presentation of the research on lead-accompaniment separation in music, which was published when research on the topic was on the verge to become fully DNN-based. With its 364 references, I somehow see it as a legacy of what was done up to 2017 on the topic.

The journal paper [J.Liu11] is given on page 117, in which I fully detail the Gaussian process generalization I proposed for source separation. I was particularly happy about it because it summarized what I understood Gaussian modeling for source separation was about.

The short conference paper [LB15] is reproduced on page 122. This paper corresponds to a strong shift in my research curriculum: it is the first time we present the α-stable model as explaining soft-masking strategies with fractional spectrograms.

The journal paper [J.Liu14b] reproduced on page 136 initiated my work on kernel methods for spectrogram modeling. I was particularly happy to have the main inventors of median-based separation on board.

The journal paper [J.Nug16] reproduced on page 149 is my most cited paper to date. It presents our effort to combine DNN with probabilistic signal processing in a principled way.

The journal paper [J.Sto19] reproduced on page 155 is the official reference for the release of `open-unmix`. It is the outcome of a long collaboration with Sony, and illustrates our strong community service effort towards reproducible research.

The journal paper [J.Liu14a] reproduced on page 162 presents our work on hardware compressed sensing. I want to highlight here that although it was rejected at several places before eventual acceptance at Nature scientific reports, it immediately reached a large audience.

The conference paper [LŞM$^+$19] reproduced on page 172 is my first publication at a pure ML conference. I was particularly happy about it because it illustrates a strong and successful collaboration towards the mathematical understanding of a new generative algorithm.

An Overview of Lead and Accompaniment Separation in Music

Zafar Rafii, *Member, IEEE,* Antoine Liutkus, *Member, IEEE,* Fabian-Robert Stöter,
Stylianos Ioannis Mimilakis, *Student Member, IEEE,* Derry FitzGerald, and Bryan Pardo, *Member, IEEE*

***Abstract*—Popular music is often composed of an accompaniment and a *lead* component, the latter typically consisting of vocals. Filtering such mixtures to extract one or both components has many applications, such as automatic karaoke and remixing. This particular case of *source separation* yields very specific challenges and opportunities, including the particular complexity of musical structures, but also relevant prior knowledge coming from acoustics, musicology or sound engineering. Due to both its importance in applications and its challenging difficulty, lead and accompaniment separation has been a popular topic in signal processing for decades. In this article, we provide a comprehensive review of this research topic, organizing the different approaches according to whether they are model-based or data-centered. For model-based methods, we organize them according to whether they concentrate on the lead signal, the accompaniment, or both. For data-centered approaches, we discuss the particular difficulty of obtaining data for learning lead separation systems, and then review recent approaches, notably those based on deep learning. Finally, we discuss the delicate problem of evaluating the quality of music separation through adequate metrics and present the results of the largest evaluation, to-date, of lead and accompaniment separation systems. In conjunction with the above, a comprehensive list of references is provided, along with relevant pointers to available implementations and repositories.**

***Index Terms*—Source separation, music, accompaniment, lead, overview.**

I. Introduction

MUSIC is a major form of artistic expression and plays a central role in the entertainment industry. While digitization and the Internet led to a revolution in the way music reaches its audience [1], [2], there is still much room to improve on how one interacts with musical content, beyond simply controlling the master volume and equalization. The ability to interact with the individual audio objects (e.g., the lead vocals) in a music recording would enable diverse applications such as music upmixing and remixing, automatic karaoke, object-wise equalization, etc.

Most publicly available music recordings (e.g., CDs, YouTube, iTunes, Spotify) are distributed as mono or stereo mixtures with multiple sound objects sharing a track. Therefore, manipulation of individual sound objects requires separation of the stereo audio mixture into several tracks, one for each different sound sources. This process is called *audio source separation* and this overview paper is concerned with an important particular case: isolating the *lead* source — typically, the vocals — from the musical accompaniment (all the rest of the signal).

As a general problem in applied mathematics, source separation has enjoyed tremendous research activity for roughly 50 years and has applications in various fields such as bioinformatics, telecommunications, and audio. Early research focused on so-called *blind* source separation, which typically builds on very weak assumptions about the signals that comprise the mixture in conjunction with very strong assumptions on the way they are mixed. The reader is referred to [3], [4] for a comprehensive review on blind source separation. Typical blind algorithms, e.g., independent component analysis (ICA) [5], [6], depend on assumptions such as: source signals are independent, there are more mixture channels than there are signals, and mixtures are well modeled as a linear combination of signals. While such assumptions are appropriate for some signals like electroencephalograms, they are often violated in audio.

Much research in audio-specific source separation [7], [8] has been motivated by the *speech enhancement* problem [9], which aims to recover clean speech from noisy recordings and can be seen as a particular instance of source separation. In this respect, many algorithms assume the audio background can be modeled as stationary. However, the musical sources are characterized by a very rich, non-stationary spectro-temporal structure. This prohibits the use of such methods. Musical sounds often exhibit highly synchronous evolution over both time and frequency, making overlap in both time and frequency very common. Furthermore, a typical commercial music mixture violates all the classical assumptions of ICA. Instruments are correlated (e.g., a chorus of singers), there are more instruments than channels in the mixture, and there are nonlinearities in the mixing process (e.g., dynamic range compression). This all has required the development of music-specific algorithms, exploiting available prior information about source structure or mixing parameters [10], [11].

This article provides an overview of nearly 50 years of research on lead and accompaniment separation in music. Due to space constraints and the large variability of the paradigms involved, we cannot delve into detailed mathematical description of each method. Instead, we will convey core ideas and

Z. Rafii is with Gracenote, Emeryville, CA, USA (zafar.rafii@nielsen.com). A. Liutkus and F.-R. Stöter are with Inria and LIRMM, University of Montpellier, France (firstname.lastname@inria.fr). S.I. Mimilakis is with Fraunhofer IDMT, Ilmenau, Germany (mis@idmt.fraunhofer.de). D. FitzGerald is with Cork School of Music, Cork Institute of Technology, Cork, Ireland (Derry.Fitzgerald@cit.ie). B. Pardo is with Northwestern University, Evanston, IL, USA (pardo@northwestern.edu).

This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01) funded by ANR, the French State agency for research. S.I. Mimilakis is supported by the European Unions H2020 Framework Programme (H2020- MSCA-ITN-2014) under grant agreement no 642685 MacSeNet

methodologies, grouping approaches according to common features. As with any attempt to impose an *a posteriori* taxonomy on such a large body of research, the resulting classification is arguable. However, we believe it is useful as a roadmap of the relevant literature.

Our objective is not to advocate one methodology over another. While the most recent methods — in particular those based on deep learning — currently show the best performance, we believe that ideas underlying earlier methods may also be inspiring and stimulate new research. This point of view leads us to focus more on the strengths of the methods rather than on their weaknesses.

The rest of the article is organized as follows. In Section II, we present the basic concepts needed to understand the discussion. We then present sections on *model-based* methods that exploit specific knowledge about the lead and/or the accompaniment signals in music to achieve separation. We show in Section III how one body of research is focused on modeling the lead signal as harmonic, exploiting this central assumption for separation. Then, Section IV describes many methods achieving separation using a model that takes the musical accompaniment as *redundant*. In Section V, we show how these two ideas were combined in other studies to achieve separation. Then, we present data-driven approaches in Section VI, which exploit large databases of audio examples where both the isolated lead and accompaniment signals are available. This enables the use of machine learning methods to learn how to separate. In Section VII, we show how the widespread availability of stereo signals may be leveraged to design algorithms that assume centered-panned vocals, but also to improve separation of most methods. Finally, Section VIII is concerned with the problem of how to evaluate the quality of the separation, and provides the results for the largest evaluation campaign to date on this topic.

II. Fundamental concepts

We now very briefly describe the basic ideas required to understand this paper, classified into three main categories: signal processing, audio modeling and probability theory. The interested reader is strongly encouraged to delve into the many online courses or textbooks available for a more detailed presentation of these topics, such as [12], [13] for signal processing, [9] for speech modeling, and [14], [15] for probability theory.

A. Signal processing

Sound is a series of pressure waves in the air. It is recorded as a *waveform*, a time-series of measurements of the displacement of the microphone diaphragm in response to these pressure waves. Sound is reproduced if a loudspeaker diaphragm is moved according to the recorded waveform. Multichannel signals simply consist of several waveforms, captured by more than one microphone. Typically, music signals are stereophonic, containing two waveforms.

Microphone displacement is typically measured at a fixed *sampling frequency*. In music processing, it is common to have sampling frequencies of 44.1 kHz (the sample frequency on a compact disc) or 48 kHz, which are higher than the typical sampling rates of 16 kHz or 8 kHz used for speech in telephony. This is because musical signals contain much higher frequency content than speech and the goal is aesthetic beauty in addition to basic intelligibility.

A time-frequency (TF) representation of sound is a matrix that encodes the time-varying *spectrum* of the waveform. Its entries are called TF *bins* and encode the varying spectrum of the waveform for all time frames and frequency channels. The most commonly-used TF representation is the short time Fourier transform (STFT) [16], which has complex entries: the angle accounts for the phase, i.e., the actual shift of the corresponding sinusoid at that time bin and frequency bin, and the magnitude accounts for the amplitude of that sinusoid in the signal. The magnitude (or power) of the STFT is called *spectrogram*. When the mixture is multichannel, the TF representation for each channel is computed, leading to a three-dimensional array: frequency, time and channel.

A TF representation is typically used as a first step in processing the audio because sources tend to be less overlapped in the TF representation than in the waveform [17]. This makes it easier to select portions of a mixture that correspond to only a single source. An STFT is typically used because it can be inverted back to the original waveform. Therefore, modifications made to the STFT can be used to create a modified waveform. Generally, a linear mixing process is considered, i.e., the mixture signal is equal to the sum of the source signals. Since the Fourier transform is a linear operation, this equality holds for the STFT. While that is not the case for the magnitude (or power) of the STFT, it is commonly assumed that the spectrograms of the sources sum to the spectrogram of the mixture.

In many methods, the separated sources are obtained by *filtering* the mixture. This can be understood as performing some equalization on the mixture, where each frequency is attenuated or kept intact. Since both the lead and the accompaniment signals change over time, the filter also changes. This is typically done using a TF *mask*, which, in its simplest form, is defined as the gain between 0 and 1 to apply on each element of the TF representation of the mixture (e.g., an STFT) in order to estimate the desired signal. Loosely speaking, it can be understood as an equalizer whose setting changes every few milliseconds. After multiplication of the mixture by a mask, the separated signal is recovered through an inverse TF transform. In the multichannel setting, more sophisticated filters may be designed that incorporate some delay and combine different channels; this is usually called *beamforming*. In the frequency domain, this is often equivalent to using complex matrices to multiply the mixture TF representation with, instead of just scalars between 0 and 1.

In practice, masks can be designed to filter the mixture in several ways. One may estimate the spectrogram for a single source or component, e.g., the accompaniment, and subtract it from the mixture spectrogram, e.g., in order to estimate the lead [18]. Another way would be to estimate separate spectrograms for both lead and accompaniment and combine them to yield a mask. For instance, a TF mask for the lead can be taken as the proportion of the lead spectrogram over the sum of

both spectrograms, at each TF bin. Such filters are often called *Wiener filters* [19] or *ratio masks*. How they are calculated may involve some additional techniques like exponentiation and may be understood according to assumptions regarding the underlying statistics of the sources. For recent work in this area, and many useful pointers in designing such masks, the reader is referred to [20].

B. Audio and speech modeling

It is typical in audio processing to describe audio waveforms as belonging to one of two different categories, which are *sinusoidal signals* — or pure tones — and *noise*. Actually, both are just the two extremes in a continuum of varying *predictability*: on the one hand, the shape of a sinusoidal wave in the future can reliably be guessed from previous samples. On the other hand, white noise is *defined* as an unpredictable signal and its spectrogram has constant energy everywhere. Different noise profiles may then be obtained by attenuating the energy of some frequency regions. This in turn induces some predictability in the signal, and in the extreme case where all the energy content is concentrated in one frequency, a pure tone is obtained.

A waveform may always be modeled as some *filter* applied on some *excitation signal*. Usually, the filter is assumed to vary smoothly across frequencies, hence modifying only what is called *the spectral envelope* of the signal, while the excitation signal comprises the rest. This is the basis for the *source-filter* model [21], which is of great importance in speech modeling, and thus also in vocal separation. As for speech, the filter is created by the shape of the vocal tract. The excitation signal is made of the glottal pulses generated by the vibration of the vocal folds. This results into *voiced* speech sounds made of time-varying harmonic/sinusoidal components. The excitation signal can also be the air flow passing through some constriction of the vocal tract. This results into *unvoiced*, noise-like, speech sounds. In this context, vowels are said to be voiced and tend to feature many sinusoids, while some phonemes such as fricatives are unvoiced and noisier.

A classical tool for dissociating the envelope from the excitation is the *cepstrum* [22]. It has applications for estimating the fundamental frequency [23], [24], for deriving the Mel-frequency cepstral coefficients (MFCC) [25], or for filtering signals through a so-called *liftering* operation [26] that enables modifications of either the excitation or the envelope parts through the source-filter paradigm.

An advantage of the source-filter model approach is indeed that one can dissociate the pitched content of the signal, embodied by the position of its harmonics, from its TF envelope which describes where the energy of the sound lies. In the case of vocals, it yields the ability to distinguish between the actual note being sung (pitch content) and the phoneme being uttered (mouth and vocal tract configuration), respectively. One key feature of vocals is they typically exhibit great variability in fundamental frequency over time. They can also exhibit larger *vibratos* (fundamental frequency modulations) and *tremolos* (amplitude modulations) in comparison to other instruments, as seen in the top spectrogram in Figure 1.

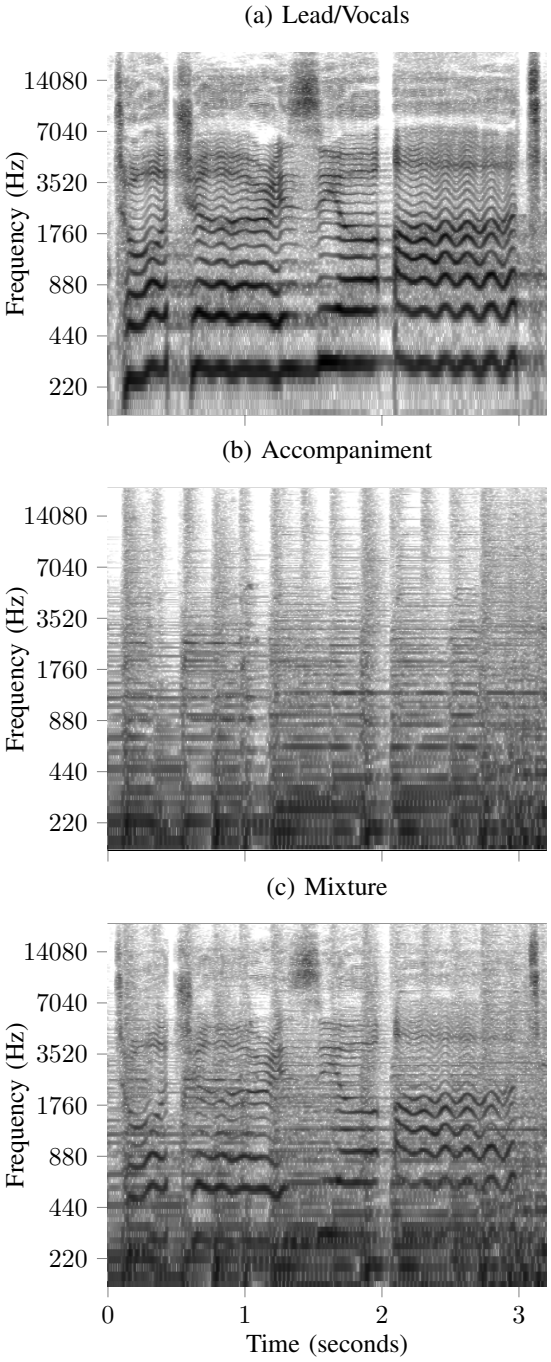


Fig. 1: Examples of spectrograms from an excerpt of the track "The Wrong'Uns - Rothko" from MUSDB18 dataset. The two sources to be separated are depicted in (a) and (b), and its mixture in (c). The vocals (a) are mostly harmonic and often well described by a source-filter model in which an excitation signal is filtered according to the vocal tract configuration. The accompaniment signal (b) features more diversity, but usually does not feature as much vibrato as for the vocals, and most importantly is seen to be *denser* and also *more redundant*. All spectrograms have log-compressed amplitudes as well as log-scaled frequency axis.

A particularity of musical signals is that they typically consist of sequences of pitched notes. A sound gives the perception of having a pitch if the majority of the energy in the audio signal is at frequencies located at integer multiples of some fundamental frequency. These integer multiples are called *harmonics*. When the fundamental frequency changes, the frequencies of these harmonics also change, yielding the typical comb spectrograms of harmonic signals, as depicted in the top spectrogram in Figure 1. Another noteworthy feature of sung melodies over simple speech is that their fundamental frequencies are, in general, located at precise frequency values corresponding to the musical key of the song. These very peculiar features are often exploited in separation methods. For simplicity reasons, we use the terms *pitch* and *fundamental frequency* interchangeably throughout the paper.

C. Probability theory

Probability theory [14], [27] is an important framework for designing many data analysis and processing methods. Many of the methods described in this article use it and it is far beyond the scope of this paper to present it rigorously. For our purpose, it will suffice to say that the *observations* consist of the mixture signals. On the other hand, the *parameters* are any relevant feature about the source signal (such as pitch or time-varying envelope) or how the signals are mixed (e.g., the panning position). These parameters can be used to derive estimates about the target lead and accompaniment signals.

We understand a probabilistic *model* as a function of both the observations and the parameters: it describes how likely the observations are, given the parameters. For instance, a flat spectrum is likely under the noise model, and a mixture of comb spectrograms is likely under a harmonic model with the appropriate pitch parameters for the sources. When the observations are given, variation in the model depends only on the parameters. For some parameter value, it tells how likely the observations are. Under a harmonic model for instance, pitch may be estimated by finding the pitch parameter that makes the observed waveform as likely as possible. Alternatively, we may want to choose between several possible models such as voiced or unvoiced. In such cases, *model selection* methods are available, such as the Bayesian information criterion (BIC) [28].

Given these basic ideas, we briefly mention two models that are of particular importance. Firstly, the hidden Markov model (HMM) [15], [29] is relevant for time-varying observations. It basically defines several *states*, each one related to a specific model and with some probabilities for transitions between them. For instance, we could define as many states as possible notes played by the lead guitar, each one associated with a typical spectrum. The *Viterbi algorithm* is a dynamic programming method which actually estimates the most likely sequence of states given a sequence of observations [30]. Secondly, the Gaussian mixture model (GMM) [31] is a way to approximate any distribution as a weighted sum of Gaussians. It is widely used in clustering, because it works well with the celebrated Expectation-Maximization (EM) algorithm [32] to assign one particular cluster to each data point, while automatically estimating the clusters parameters. As we will see later, many methods work by assigning each TF bin to a given source in a similar way.

III. Modeling the lead signal: harmonicity

As mentioned in Section II-B, one particularity of vocals is their production by the vibration of the vocal folds, further filtered by the vocal tract. As a consequence, sung melodies are *mostly* harmonic, as depicted in Figure 1, and therefore have a fundamental frequency. If one can track the pitch of the vocals, one can then estimate the energy at the harmonics of the fundamental frequency and reconstruct the voice. This is the basis of the oldest methods (as well as some more recent methods) we are aware of for separating the lead signal from a musical mixture.

Such methods are summarized in Figure 2. In a first step, the objective is to get estimates of the time-varying fundamental frequency for the lead at each time frame. A second step in this respect is then to track this fundamental frequency over time, in other words, to find the best sequence of estimates, in order to identify the melody line. This can done either by a suitable pitch detection method, or by exploiting the availability of the score. Such algorithms typically assume that the lead corresponds to the harmonic signal with strongest amplitude. For a review on the particular topic of melody extraction, the reader is referred to [33].

From this starting point, we can distinguish between two kinds of approaches, depending on how they exploit the pitch information.

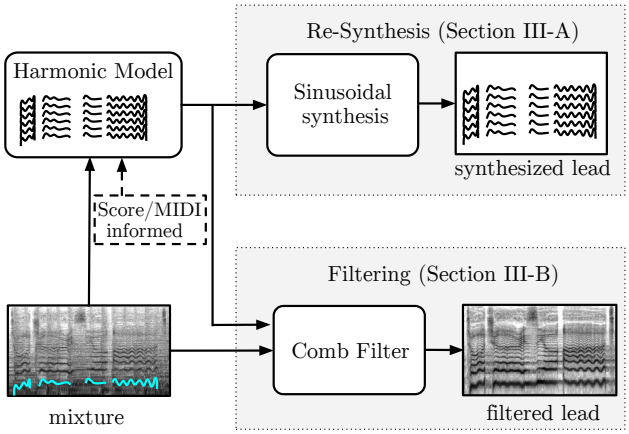


Fig. 2: The approaches based on a *harmonic assumption* for vocals. In a first analysis step, the fundamental frequency of the lead signal is extracted. From it, a separation is obtained either by resynthesis (Section III-A), or by filtering the mixture (Section III-B).

A. Analysis-synthesis approaches

The first option to obtain the separated lead signal is to resynthesize it using a sinusoidal model. A sinusoidal model decomposes the sound with a set of sine waves of varying frequency and amplitude. If one knows the fundamental frequency of a pitched sound (like a singing voice), as well as the

spectral envelope of the recording, then one can reconstruct the sound by making a set of sine waves whose frequencies are those of the harmonics of the fundamental frequency, and whose amplitudes are estimated from the spectral envelope of the audio. While the spectral envelope of the recording is generally not exactly the same as the spectral envelope of the target source, it can be a reasonable approximation, especially assuming that different sources do not overlap too much with each other in the TF representation of the mixture.

This idea allows for time-domain processing and was used in the earliest methods we are aware of. In 1973, Miller proposed in [34] to use the homomorphic vocoder [35] to separate the excitation function and impulse response of the vocal tract. Further refinements include segmenting parts of the signal as voiced, unvoiced, or silences using a heuristic program and manual interaction. Finally, cepstral liftering [26] was exploited to compensate for the noise or accompaniment.

Similarly, Maher used an analysis-synthesis approach in [36], assuming the mixtures are composed of only two harmonic sources. In his case, pitch detection was performed on the STFT and included heuristics to account for possibly colliding harmonics. He finally resynthesized each musical voice with a sinusoidal model.

Wang proposed instantaneous and frequency-warped techniques for signal parameterization and source separation, with application to voice separation in music [37], [38]. He introduced a frequency-locked loop algorithm which uses multiple harmonically constrained trackers. He computed the estimated fundamental frequency from a maximum-likelihood weighting of the tracking estimates. He was then able to estimate harmonic signals such as voices from complex mixtures.

Meron and Hirose proposed to separate singing voice and piano accompaniment [39]. In their case, prior knowledge consisting of musical scores was considered. Sinusoidal modeling as described in [40] was used.

Ben-Shalom and Dubnov proposed to filter an instrument or a singing voice out in such a way [41]. They first used a score alignment algorithm [42], assuming a known score. Then, they used the estimated pitch information to design a filter based on a harmonic model [43] and performed the filtering using the linear constraint minimum variance approach [44]. They additionally used a heuristic to deal with the unvoiced parts of the singing voice.

Zhang and Zhang proposed an approach based on harmonic structure modeling [45], [46]. They first extracted harmonic structures for singing voice and background music signals using a sinusoidal model [43], by extending the pitch estimation algorithm in [47]. Then, they used the clustering algorithm in [48] to learn harmonic structure models for the background music signals. Finally, they extracted the harmonic structures for all the instruments to reconstruct the background music signals and subtract them from the mixture, leaving only the singing voice signal.

More recently, Fujihara et al. proposed an accompaniment reduction method for singer identification [49], [50]. After fundamental frequency estimation using [51], they extracted the harmonic structure of the melody, i.e., the power and phase of the sinusoidal components at fundamental frequency and harmonics. Finally, they resynthesized the audio signal of the melody using the sinusoidal model in [52].

Similarly, Mesaros et al. proposed a vocal separation method to help with singer identification [53]. They first applied a melody transcription system [54] which estimates the melody line with the corresponding MIDI note numbers. Then, they performed sinusoidal resynthesis, estimating amplitudes and phases from the polyphonic signal.

In a similar manner, Duan et al. proposed to separate harmonic sources, including singing voices, by using harmonic structure models [55]. They first defined an average harmonic structure model for an instrument. Then, they learned a model for each source by detecting the spectral peaks using a cross-correlation method [56] and quadratic interpolation [57]. Then, they extracted the harmonic structures using BIC and a clustering algorithm [48]. Finally, they separated the sources by re-estimating the fundamental frequencies, re-extracting the harmonics, and reconstructing the signals using a phase generation method [58].

Lagrange et al. proposed to formulate lead separation as a graph partition problem [59], [60]. They first identified peaks in the spectrogram and grouped the peaks into clusters by using a similarity measure which accounts for harmonically related peaks, and the normalized cut criterion [61] which is used for segmenting graphs in computer vision. They finally selected the cluster of peaks which corresponds to a predominant harmonic source and resynthesized it using a bank of sinusoidal oscillators.

Ryynänen et al. proposed to separate accompaniment from polyphonic music using melody transcription for karaoke applications [62]. They first transcribed the melody into a MIDI note sequence and a fundamental frequency trajectory, using the method in [63], an improved version of the earlier method [54]. Then, they used sinusoidal modeling to estimate, resynthesize, and remove the lead vocals from the musical mixture, using the quadratic polynomial-phase model in [64].

B. *Comb-filtering approaches*

Using sinusoidal synthesis to generate the lead signal suffers from a typical *metallic* sound quality, which is mostly due to discrepancies between the estimated excitation signals of the lead signal compared to the ground truth. To address this issue, an alternative approach is to exploit harmonicity in another way, by filtering out everything from the mixture that is not located close to the detected harmonics.

Li and Wang proposed to use a vocal/non-vocal classifier and a predominant pitch detection algorithm [65], [66]. They first detected the singing voice by using a spectral change detector [67] to partition the mixture into homogeneous portions, and GMMs on MFCCs to classify the portions as vocal or non-vocal. Then, they used the predominant pitch detection algorithm in [68] to detect the pitch contours from the vocal portions, extending the multi-pitch tracking algorithm in [69]. Finally, they extracted the singing voice by decomposing the vocal portions into TF units and labeling them as singing or accompaniment dominant, extending the speech separation algorithm in [70].

Han and Raphael proposed an approach for desoloing a recording of a soloist with an accompaniment given a musical score and its time alignment with the recording [71]. They derived a mask [72] to remove the solo part after using an EM algorithm to estimate its melody, that exploits the score as side information.

Hsu et al. proposed an approach which also identifies and separates the unvoiced singing voice [73], [74]. Instead of processing in the STFT domain, they use the perceptually motivated gammatone filter-bank as in [66], [70]. They first detected accompaniment, unvoiced, and voiced segments using an HMM and identified voice-dominant TF units in the voiced frames by using the singing voice separation method in [66], using the predominant pitch detection algorithm in [75]. Unvoiced-dominant TF units were identified using a GMM classifier with MFCC features learned from training data. Finally, filtering was achieved with spectral subtraction [76].

Raphael and Han then proposed a classifier-based approach to separate a soloist from accompanying instruments using a time-aligned symbolic musical score [77]. They built a tree-structured classifier [78] learned from labeled training data to classify TF points in the STFT as belonging to solo or accompaniment. They additionally constrained their classifier to estimate masks having a connected structure.

Cano et al. proposed various approaches for solo and accompaniment separation. In [79], they separated saxophone melodies from mixtures with piano and/or orchestra by using a melody line detection algorithm, incorporating information about typical saxophone melody lines. In [80]–[82], they proposed to use the pitch detection algorithm in [83]. Then, they refined the fundamental frequency and the harmonics, and created a binary mask for the solo and accompaniment. They finally used a post-processing stage to refine the separation. In [84], they included a noise spectrum in the harmonic refinement stage to also capture noise-like sounds in vocals. In [85], they additionally included common amplitude modulation characteristics in the separation scheme.

Bosch et al. proposed to separate the lead instrument using a musical score [86]. After a preliminary alignment of the score to the mixture, they estimated a score confidence measure to deal with local misalignments and used it to guide the predominant pitch tracking. Finally, they performed low-latency separation based on the method in [87], by combining harmonic masks derived from the estimated pitch and additionally exploiting stereo information as presented later in Section VII.

Vaneph et al. proposed a framework for vocal isolation to help spectral editing [88]. They first used a voice activity detection process based on a deep learning technique [89]. Then, they used pitch tracking to detect the melodic line of the vocal and used it to separate the vocal and background, allowing a user to provide manual annotations when necessary.

C. Shortcomings

As can be seen, explicitly assuming that the lead signal is harmonic led to an important body of research. While the aforementioned methods show excellent performance when their assumptions are valid, their performance can drop significantly in adverse, but common situations.

Firstly, vocals are not always purely harmonic as they contain unvoiced phonemes that are not harmonic. As seen above, some methods already handle this situation. However, vocals can also be whispered or saturated, both of which are difficult to handle with a harmonic model.

Secondly, methods based on the harmonic model depend on the quality of the pitch detection method. If the pitch detector switches from following the pitch of the lead (e.g., the voice) to another instrument, the wrong sound will be isolated from the mix. Often, pitch detectors assume the lead signal is the *loudest* harmonic sound in the mix. Unfortunately, this is not always the case. Another instrument may be louder or the lead may be silent for a passage. The tendency to follow the pitch of the wrong instrument can be mitigated by applying constraints on the pitch range to estimate and by using a perceptually relevant weighting filter before performing pitch tracking. Of course, these approaches do not help when the lead signal is silent.

IV. Modeling the accompaniment: redundancy

In the previous section, we presented methods whose main focus was the modeling of a harmonic lead melody. Most of these studies did not make modeling the accompaniment a core focus. On the contrary, it was often dealt with as adverse noise to which the harmonic processing method should be robust to.

In this section, we present another line of research which concentrates on modeling the accompaniment under the assumption it is somehow more *redundant* than the lead signal. This assumption stems from the fact that musical accompaniments are often highly structured, with elements being repeated many times. Such repetitions can occur at the note level, in terms of rhythmic structure, or even from a harmonic point of view: instrumental notes are often constrained to have their pitch lie in a small set of frequencies. Therefore, modeling and removing the redundant elements of the signal are assumed to result in removal of the accompaniment.

In this paper, we identify three families of methods that exploit the redundancy of the accompaniment for separation.

A. Grouping low-rank components

The first set of approaches we consider is the identification of redundancy in the accompaniment through the assumption that its spectrogram may be well represented by only a few components. Techniques exploiting this idea then focus on algebraic methods that decompose the mixture spectrogram into the product of a few template spectra activated over time. One way to do so is via non-negative matrix factorization (NMF) [90], [91], which incorporates non-negative constraints. In Figure 3, we picture methods exploiting such techniques. After factorization, we obtain several spectra, along with their activations over time. A subsequent step is the clustering of these spectra (and activations) into the lead or the accompaniment. Separation is finally performed by deriving Wiener filters to estimate the lead and the accompaniment from the mixture. For related applications of NMF in music analysis, the reader is referred to [92]–[94].

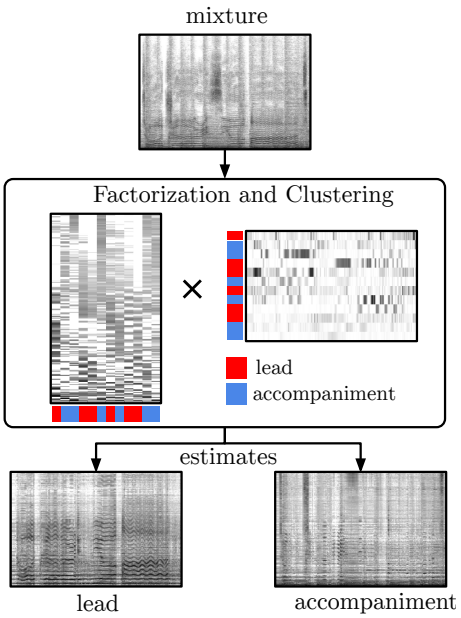


Fig. 3: The approaches based on a *low-rank* assumption. Non-negative matrix factorization (NMF) is used to identify *components* from the mixture, that are subsequently clustered into lead or accompaniment. Additional constraints may be incorporated.

Vembu and Baumann proposed to use NMF (and also ICA [95]) to separate vocals from mixtures [96]. They first discriminated between vocal and non-vocal sections in a mixture by using different combinations of features, such as MFCCs [25], perceptual linear predictive (PLP) coefficients [97], and log frequency power coefficients (LFPC) [98], and training two classifiers, namely neural networks and support vector machines (SVM). They then applied redundancy reduction techniques on the TF representation of the mixture to separate the sources [99], by using NMF (or ICA). The components were then grouped as vocal and non-vocal by reusing a vocal/non-vocal classifier with MFCC, LFPC, and PLP coefficients.

Chanrungutai and Ratanamahatana proposed to use NMF with automatic component selection [100], [101]. They first decomposed the mixture spectrogram using NMF with a fixed number of basis components. They then removed the components with brief rhythmic and long-lasting continuous events, assuming that they correspond to instrumental sounds. They finally used the remaining components to reconstruct the singing voice, after refining them using a high-pass filter.

Marxer and Janer proposed an approach based on a Tikhonov regularization [102] as an alternative to NMF, for singing voice separation [103]. Their method sacrificed the non-negativity constraints of the NMF in exchange for a computationally less expensive solution for spectrum decomposition, making it more interesting in low-latency scenarios.

Yang et al. proposed a Bayesian NMF approach [104], [105]. Following the approaches in [106] and [107], they used a Poisson distribution for the likelihood function and exponential distributions for the model parameters in the NMF algorithm, and derived a variational Bayesian EM algorithm [32] to solve the NMF problem. They also adaptively determined the number of bases from the mixture. They finally grouped the bases into singing voice and background music by using a k-means clustering algorithm [108] or an NMF-based clustering algorithm.

In a different manner, Smaragdis and Mysore proposed a user-guided approach for removing sounds from mixtures by humming the target sound to be removed, for example a vocal track [109]. They modeled the mixture using probabilistic latent component analysis (PLCA) [110], another equivalent formulation of NMF. One key feature of exploiting user input was to facilitate the grouping of components into vocals and accompaniment, as humming helped to identify some of the parameters for modeling the vocals.

Nakamuray and Kameoka proposed an L_p-norm NMF [111], with p controlling the sparsity of the error. They developed an algorithm for solving this NMF problem based on the auxiliary function principle [112], [113]. Setting an adequate number of bases and p taken as small enough allowed them to estimate the accompaniment as the low-rank decomposition, and the singing voice as the error of the approximation, respectively. Note that, in this case, the singing voice was not explicitly modeled as a sparse component but rather corresponded to the error which happened to be constrained as sparse. The next subsection will actually deal with approaches that explicitly model the vocals as the sparse component.

B. Low-rank accompaniment, sparse vocals

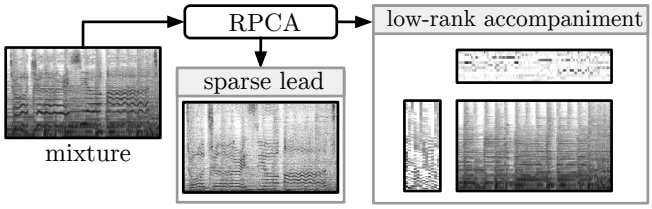


Fig. 4: The approaches based on a *low-rank accompaniment, sparse vocals* assumption. As opposed to methods based on NMF, methods based on robust principal component analysis (RPCA) assume the lead signal has a sparse and non-structured spectrogram.

The methods presented in the previous section first compute a decomposition of the mixture into many components that are sorted *a posteriori* as accompaniment or lead. As can be seen, this means they make a low-rank assumption for the accompaniment, but typically *also for the vocals*. However, as can for instance be seen on Figure 1, the spectrogram for the vocals do exhibit much more freedom than accompaniment, and experience shows they are not adequately described by a small number of spectral bases. For this reason, another track of research depicted in Figure 4 focused on using a low-rank assumption on the accompaniment *only*, while assuming the vocals are *sparse and not structured*. This loose

assumption means that only a few coefficients from their spectrogram should have significant magnitude, and that they should not feature significant redundancy. Those ideas are in line with robust principal component analysis (RPCA) [114], which is the mathematical tool used by this body of methods, initiated by Huang et al. for singing voice separation [115] . It decomposes a matrix into a sparse and low-rank component.

Sprechmann et al. proposed an approach based on RPCA for online singing voice separation [116]. They used ideas from convex optimization [117], [118] and multi-layer neural networks [119]. They presented two extensions of RPCA and robust NMF models [120]. They then used these extensions in a multi-layer neural network framework which, after an initial training stage, allows online source separation.

Jeong and Lee proposed two extensions of the RPCA model to improve the estimation of vocals and accompaniment from the sparse and low-rank components [121]. Their first extension included the Schatten p and ℓ_p norms as generalized nuclear norm optimizations [122]. They also suggested a pre-processing stage based on logarithmic scaling of the mixture TF representation to enhance the RPCA.

Yang also proposed an approach based on RPCA with dictionary learning for recovering low-rank components [123]. He introduced a multiple low-rank representation following the observation that elements of the singing voice can also be recovered by the low-rank component. He first incorporated online dictionary learning methods [124] in his methodology to obtain prior information about the structure of the sources and then incorporated them into the RPCA model.

Chan and Yang then extended RPCA to complex and quaternionic cases with application to singing voice separation [125]. They extended the principal component pursuit (PCP) [114] for solving the RPCA problem by presenting complex and quaternionic proximity operators for the ℓ_1 and trace-norm regularizations to account for the missing phase information.

C. Repetitions within the accompaniment

While the rationale behind low-rank methods for lead-accompaniment separation is to exploit the idea that the musical background should be redundant, adopting a low-rank model is not the only way to do it. An alternate way to proceed is to exploit the musical *structure* of songs, to find *repetitions* that can be utilized to perform separation. Just like in RPCA-based methods, the accompaniment is then assumed to be the only source for which repetitions will be found. The unique feature of the methods described here is they combine music structure analysis [126]–[128] with particular ways to exploit the identification of repeated parts of the accompaniment.

Rafii et al. proposed the REpeating Pattern Extraction Technique (REPET) to separate the accompaniment by assuming it is repeating [129]–[131], which is often the case in popular music. This approach, which is representative of this line of research, is represented on Figure 5. First, a repeating period is extracted by a music information retrieval system, such as a beat spectrum [132] in this case. Then, this extracted information is used to estimate the spectrogram of the accompaniment through an averaging of the identified repetitions. From this, a filter is derived.

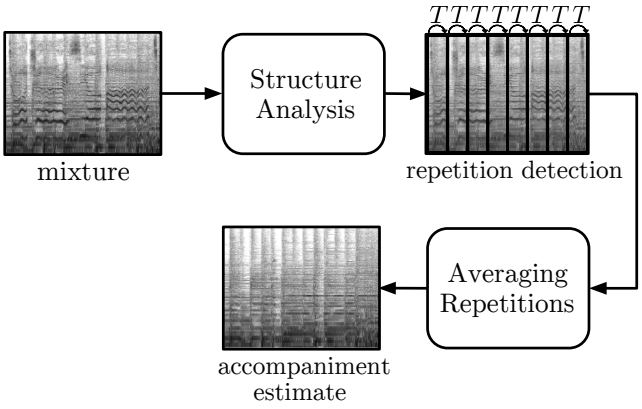


Fig. 5: The approaches based on a *repetition* assumption for accompaniment. In a first analysis step, repetitions are identified. Then, they are used to build an estimate for the accompaniment spectrogram and proceed to separation.

Seetharaman et al. [133] leveraged the two dimensional Fourier transform (2DFT) of the spectrogram to create an algorithm very similar to REPET. The properties of the 2DFT let them separate the periodic background from the non-periodic vocal melody by deleting peaks in the 2DFT. This eliminated the need to create an explicit model of the periodic audio and without the need to find the period of repetition, both of which are required in REPET.

Liutkus et al. adapted the REPET approach in [129], [130] to handle repeating structures varying along time by modeling the repeating patterns only locally [131], [134]. They first identified a repeating period for every time frame by computing a beat spectrogram as in [132]. Then they estimated the spectrogram of the accompaniment by averaging the time frames in the mixture spectrogram at their local period rate, for every TF bin. From this, they finally extracted the repeating structure by deriving a TF mask.

Rafii et al. further extended the REPET approaches in [129], [130] and [134] to handle repeating structures that are not periodic. To do this, they proposed the REPET-SIM method in [131], [135] to identify repeating frames for every time frame by computing a self-similarity matrix, as in [136]. Then, they estimated the accompaniment spectrogram at every TF bin by averaging the neighbors identified thanks to that similarity matrix. An extension for real-time processing was presented in [137] and a version exploiting user interaction was proposed in [138]. A method close to REPET-SIM was also proposed by FitzGerald in [139].

Liutkus et al. proposed the Kernel Additive modeling (KAM) [140], [141] as a framework which generalizes the REPET approaches in [129]–[131], [134], [135]. They assumed that a source at a TF location can be modeled using its values at other locations through a specified kernel which can account for features such as periodicity, self-similarity, stability over time or frequency, etc. This notably enabled modeling of the accompaniment using more than one repeating pattern. Liutkus et al. also proposed a light version using a fast compression algorithm to make the approach more scalable

[142]. The approach was also used for interference reduction in music recordings [143], [144].

With the same idea of exploiting intra-song redundancies for singing voice separation, but through a very different methodology, Moussallam et al. assumed in [145] that all the sources can be decomposed sparsely in the same dictionary and used a matching pursuit greedy algorithm [146] to solve the problem. They integrated the separation process in the algorithm by modifying the atom selection criterion and adding a decision to assign a chosen atom to the repeated source or to the lead signal.

Deif et al. proposed to use multiple median filters to separate vocals from music recordings [147]. They augmented the approach in [148] with diagonal median filters to improve the separation of the vocal component. They also investigated different filter lengths to further improve the separation.

Lee et al. also proposed to use the KAM approach [149]–[152]. They applied the β-order minimum mean square error (MMSE) estimation [153] to the back-fitting algorithm in KAM to improve the separation. They adaptively calculated a perceptually weighting factor α and the singular value decomposition (SVD)-based factorized spectral amplitude exponent β for each kernel component.

D. Shortcomings

While methods focusing on harmonic models for the lead often fall short in their expressive power for the accompaniment, the methods we reviewed in this section are often observed to suffer exactly from the converse weakness, namely they do not provide an adequate model for the lead signal. Hence, the separated vocals often will feature interference from unpredictable parts from the accompaniment, such as some percussion or effects which occur infrequently.

Furthermore, even if the musical accompaniment will exhibit more redundancy, the vocals part will also be redundant to some extent, which is poorly handled by these methods. When the lead signal is not vocals but played by some lead instrument, its redundancy is even more pronounced, because the notes it plays lie in a reduced set of fundamental frequencies. Consequently, such methods would include the redundant parts of the lead within the accompaniment estimate, for example, a steady humming by a vocalist.

V. Joint models for lead and accompaniment

In the previous sections, we reviewed two important bodies of literature, focused on modeling either the lead or the accompaniment parts of music recordings, respectively. While each approach showed its own advantages, it also featured its own drawbacks. For this reason, some researchers devised methods combining ideas for modeling both the lead and the accompaniment sources, and thus benefiting from both approaches. We now review this line of research.

A. Using music structure analysis to drive learning

The first idea we find in the literature is to augment methods for accompaniment modeling with the prior identification of sections where the vocals are present or absent. In the case of the low rank models discussed in Sections IV-A and IV-B, such a strategy indeed dramatically improves performance.

Raj et al. proposed an approach in [154] that is based on the PLCA formulation of NMF [155], and extends their prior work [156]. The parameters for the frequency distribution of the background music are estimated from the background music-only segments, and the rest of the parameters from the singing voice+background music segments, assuming a priori identified vocal regions.

Han and Chen also proposed a similar approach for melody extraction based on PLCA [157], which includes a further estimate of the melody from the vocals signal by an autocorrelation technique similar to [158].

Gómez et al. proposed to separate the singing voice from the guitar accompaniment in flamenco music to help with melody transcription [159]. They first manually segmented the mixture into vocal and non-vocal regions. They then learned percussive and harmonic bases from the non-vocal regions by using an unsupervised NMF percussive/harmonic separation approach [93], [160]. The vocal spectrogram was estimated by keeping the learned percussive and harmonic bases fixed.

Papadopoulos and Ellis proposed a signal-adaptive formulation of RPCA which incorporates music content information to guide the recovery of the sparse and low-rank components [161]. Prior musical knowledge, such as predominant melody, is used to regularize the selection of active coefficients during the optimization procedure.

In a similar manner, Chan et al. proposed to use RPCA with vocal activity information [162]. They modified the RPCA algorithm to constraint parts of the input spectrogram to be non-sparse to account for the non-vocal parts of the singing voice.

A related method was proposed by Jeong and Lee in [163], using RPCA with a weighted l_1-norm. They replaced the uniform weighting between the low-rank and sparse components in the RPCA algorithm by an adaptive weighting based on the variance ratio between the singing voice and the accompaniment. One key element of the method is to incorporate vocal activation information in the weighting.

B. Factorization with a known melody

While using only the knowledge of vocal activity as described above already yields an increase of performance over methods operating blindly, many authors went further to also incorporate the fact that vocals often have a strong melody line. Some redundant model is then assumed for the accompaniment, while also enforcing a harmonic model for the vocals.

An early method to achieve this is depicted in Figure 6 and was proposed by Virtanen et al. in [164]. They estimated the pitch of the vocals in the mixture by using a melody transcription algorithm [63] and derived a binary TF mask to identify where vocals are not present. They then applied NMF on the remaining non-vocal segments to learn a model for the background.

Wang and Ou also proposed an approach which combines melody extraction and NMF-based soft masking [165]. They

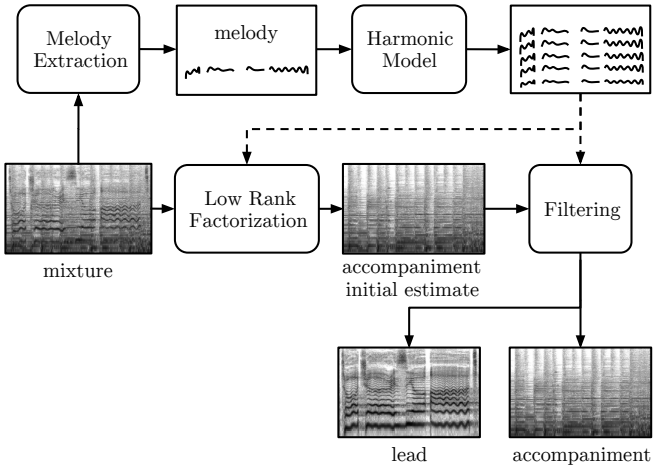


Fig. 6: Factorization informed with the melody. First, melody extraction is performed on the mixture. Then, this information is used to drive the estimation of the accompaniment: TF bins pertaining to the lead should not be taken into account for estimating the accompaniment model.

identified accompaniment, unvoiced, and voiced segments in the mixture using an HMM model with MFCCs and GMMs. They then estimated the pitch of the vocals from the voiced segments using the method in [166] and an HMM with the Viterbi algorithm as in [167]. They finally applied a soft mask to separate voice and accompaniment.

Rafii et al. investigated the combination of an approach for modeling the background and an approach for modeling the melody [168]. They modeled the background by deriving a rhythmic mask using the REPET-SIM algorithm [135] and the melody by deriving a harmonic mask using a pitch-based algorithm [169]. They proposed a parallel and a sequential combination of those algorithms.

Venkataramani et al. proposed an approach combining sinusoidal modeling and matrix decomposition, which incorporates prior knowledge about singer and phoneme identity [170]. They applied a predominant pitch algorithm on annotated sung regions [171] and performed harmonic sinusoidal modeling [172]. Then, they estimated the spectral envelope of the vocal component from the spectral envelope of the mixture using a phoneme dictionary. After that, a spectral envelope dictionary representing sung vowels from song segments of a given singer was learned using an extension of NMF [173], [174]. They finally estimated a soft mask using the singer-vowel dictionary to refine and extract the vocal component.

Ikemiya et al. proposed to combine RPCA with pitch estimation [175], [176]. They derived a mask using RPCA [115] to separate the mixture spectrogram into singing voice and accompaniment components. They then estimated the fundamental frequency contour from the singing voice component based on [177] and derived a harmonic mask. They integrated the two masks and resynthesized the singing voice and accompaniment signals. Dobashi et al. then proposed to use that singing voice separation approach in a music performance assistance system [178].

Hu and Liu proposed to combine approaches based on matrix decomposition and pitch information for singer identification [179]. They used non-negative matrix partial co-factorization [173], [180] which integrates prior knowledge about the singing voice and the accompaniment, to separate the mixture into singing voice and accompaniment portions. They then identified the singing pitch from the singing voice portions using [181] and derived a harmonic mask as in [182], and finally reconstructed the singing voice using a missing feature method [183]. They also proposed to add temporal and sparsity criteria to their algorithm [184].

That methodology was also adopted by Zhang et al. in [185], that followed the framework of the pitch-based approach in [66], by performing singing voice detection using an HMM classifier, singing pitch detection using the algorithm in [186], and singing voice separation using a binary mask. Additionally, they augmented that approach by analyzing the latent components of the TF matrix using NMF in order to refine the singing voice and accompaniment.

Zhu et al. [187] proposed an approach which is also representative of this body of literature, with the pitch detection algorithm being the one in [181] and binary TF masks used for separation after NMF.

C. Joint factorization and melody estimation

The methods presented above put together the ideas of modeling the lead (typically the vocals) as featuring a melodic harmonic line and the accompaniment as redundant. As such, they already exhibit significant improvement over approaches only applying one of these ideas as presented in Sections III and IV, respectively. However, these methods above are still restricted in the sense that the analysis performed on each side cannot help improve the other one. In other words, the estimation of the models for the lead and the accompaniment are done sequentially. Another idea is to proceed *jointly*.

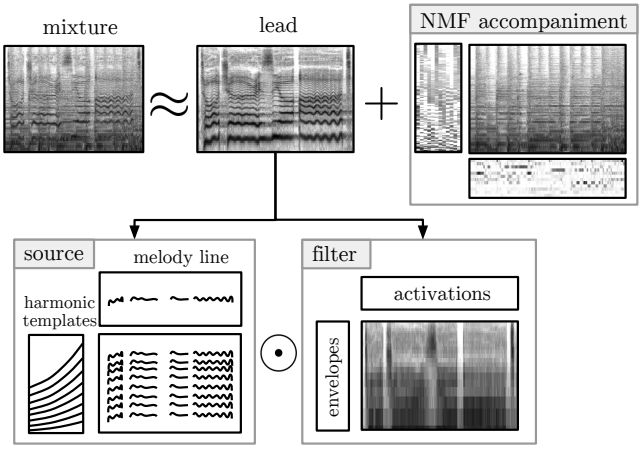


Fig. 7: Joint estimation of the lead and accompaniment, the former one as a source-filter model and the latter one as an NMF model.

A seminal work in this respect was done by Durrieu et al. using a source-filter and NMF model [188]–[190], depicted

in Figure 7. Its core idea is to decompose the mixture spectrogram as the sum of two terms. The first term accounts for the lead and is inspired by the source-filter model described in Section II: it is the element-wise product of an *excitation* spectrogram with a *filter* spectrogram. The former one can be understood as harmonic combs activated by the melodic line, while the latter one modulates the envelope and is assumed low-rank because few phonemes are used. The second term accounts for the accompaniment and is modeled with a standard NMF. In [188]–[190], they modeled the lead by using a GMM-based model [191] and a glottal source model [192], and the accompaniment by using an instantaneous mixture model [193] leading to an NMF problem [94]. They jointly estimated the parameters of their models by maximum likelihood estimation using an iterative algorithm inspired by [194] with multiplicative update rules developed in [91]. They also extracted the melody by using an algorithm comparable to the Viterbi algorithm, before re-estimating the parameters and finally performing source separation using Wiener filters [195]. In [196], they proposed to adapt their model for user-guided source separation.

The joint modeling of the lead and accompaniment parts of a music signal was also considered by Fuentes et al. in [197], that introduced the idea of using a log-frequency TF representation called the constant-Q transform (CQT) [198]–[200]. The advantage of such a representation is that a change in pitch corresponds to a simple translation in the TF plane, instead of a scaling as in the STFT. This idea was used along the creation of a user interface to guide the decomposition, in line with what was done in [196].

Joder and Schuller used the source-filter NMF model in [201], additionally exploiting MIDI scores [202]. They synchronized the MIDI scores to the audio using the alignment algorithm in [203]. They proposed to exploit the score information through two types of constraints applied in the model. In a first approach, they only made use of the information regarding whether the leading voice is present or not in each frame. In a second approach, they took advantage of both time and pitch information on the aligned score.

Zhao et al. proposed a score-informed leading voice separation system with a weighting scheme [204]. They extended the system in [202], which is based on the source-filter NMF model in [201], by using a Laplacian or a Gaussian-based mask on the NMF activation matrix to enhance the likelihood of the score-informed pitch candidates.

Jointly estimating accompaniment and lead allowed for some research in correctly estimating the unvoiced parts of the lead, which is the main issue with purely harmonic models, as highlighted in Section III-C. In [201], [205], Durrieu et al. extended their model to account for the unvoiced parts by adding white noise components to the voice model.

In the same direction, Janer and Marxer proposed to separate unvoiced fricative consonants using a semi-supervised NMF [206]. They extended the source-filter NMF model in [201] using a low-latency method with timbre classification to estimate the predominant pitch [87]. They approximated the fricative consonants as an additive wideband component, training a model of NMF bases. They also used the transient quality to differentiate between fricatives and drums, after extracting transient time points using the method in [207].

Similarly, Marxer and Janer then proposed to separately model the singing voice breathiness [208]. They estimated the breathiness component by approximating the voice spectrum as a filtered composition of a glottal excitation and a wide-band component. They modeled the magnitude of the voice spectrum using the model in [209] and the envelope of the voice excitation using the model in [192]. They estimated the pitch using the method in [87]. This was all integrated into the source-filter NMF model.

The body of research initiated by Durrieu et al. in [188] consists of using algebraic models more sophisticated than one simple matrix product, but rather inspired by musicological knowledge. Ozerov et al. formalized this idea through a general framework and showed its application for singing voice separation [210]–[212].

Finally, Hennequin and Rigaud augmented their model to account for long-term reverberation, with application to singing voice separation [213]. They extended the model in [214] which allows extraction of the reverberation of a specific source with its dry signal. They combined this model with the source-filter NMF model in [189].

D. Different constraints for different sources

Algebraic methods that decompose the mixture spectrogram as the sum of the lead and accompaniment spectrograms are based on the minimization of a *cost* or *loss function* which measures the error between the approximation and the observation. While the methods presented above for lead and accompaniment separation did propose more sophisticated models with parameters explicitly pertaining to the lead or the accompaniment, another option that is also popular in the dedicated literature is to modify the cost function of an optimization algorithm for an existing algorithm (e.g., RPCA), so that one part of the resulting components would preferentially account for one source or another.

This approach can be exemplified by the harmonic-percussive source separation method (HPSS), presented in [160], [215], [216]. It consists in filtering a mixture spectrogram so that horizontal lines go in a so-called *harmonic* source, while its vertical lines go into a *percussive* source. Separation is then done with TF masking. Of course, such a method is not adequate for lead and accompaniment separation *per se*, because all the harmonic content of the accompaniment is classified as harmonic. However, it shows that *nonparametric* approaches are also an option, provided the cost function itself is well chosen for each source.

This idea was followed by Yang in [217] who proposed an approach based on RPCA with the incorporation of harmonicity priors and a back-end drum removal procedure to improve the decomposition. He added a regularization term in the algorithm to account for harmonic sounds in the low-rank component and used an NMF-based model trained for drum separation [211] to eliminate percussive sounds in the sparse component.

Jeong and Lee proposed to separate a vocal signal from a music signal [218], extending the HPSS approach in [160],

[215]. Assuming that the spectrogram of the signal can be represented as the sum of harmonic, percussive, and vocal components, they derived an objective function which enforces the temporal and spectral continuity of the harmonic and percussive components, respectively, similarly to [160], but also the sparsity of the vocal component. Assuming non-negativity of the components, they then derived iterative update rules to minimize the objective function. Ochiai et al. extended this work in [219], notably by imposing harmonic constraints for the lead.

Watanabe et al. extended RPCA for singing voice separation [220]. They added a harmonicity constraint in the objective function to account for harmonic structures, such as in vocal signals, and regularization terms to enforce the non-negativity of the solution. They used the generalized forward-backward splitting algorithm [221] to solve the optimization problem. They also applied post-processing to remove the low frequencies in the vocal spectrogram and built a TF mask to remove time frames with low energy.

Going beyond smoothness and harmonicity, Hayashi et al. proposed an NMF with a constraint to help separate periodic components, such as a repeating accompaniment [222]. They defined a periodicity constraint which they incorporated in the objective function of the NMF algorithm to enforce the periodicity of the bases.

E. Cascaded and iterated methods

In their effort to propose separation methods for the lead and accompaniment in music, some authors discovered that very different methods often have complementary strengths. This motivated the *combination* of methods. In practice, there are several ways to follow this line of research.

One potential route to achieve better separation is to *cascade* several methods. This is what FitzGerald and Gainza proposed in [216] with multiple median filters [148]. They used a median-filter based HPSS approach at different frequency resolutions to separate a mixture into harmonic, percussive, and vocal components. They also investigated the use of STFT or CQT as the TF representation and proposed a post-processing step to improve the separation with tensor factorization techniques [223] and non-negative partial co-factorization [180].

The two-stage HPSS system proposed by Tachibana et al. in [224] proceeds the same way. It is an extension of the melody extraction approach in [225] and was applied for karaoke in [226]. It consists in using the optimization-based HPSS algorithm from [160], [215], [227], [228] at different frequency resolutions to separate the mixture into harmonic, percussive, and vocal components.

HPSS was not the only separation module considered as the building block of combined lead and accompaniment separation approaches. Deif et al. also proposed a multi-stage NMF-based algorithm [229], based on the approach in [230]. They used a local spectral discontinuity measure to refine the non-pitched components obtained from the factorization of the long window spectrogram and a local temporal discontinuity measure to refine the non-percussive components obtained from factorization of the short window spectrogram.

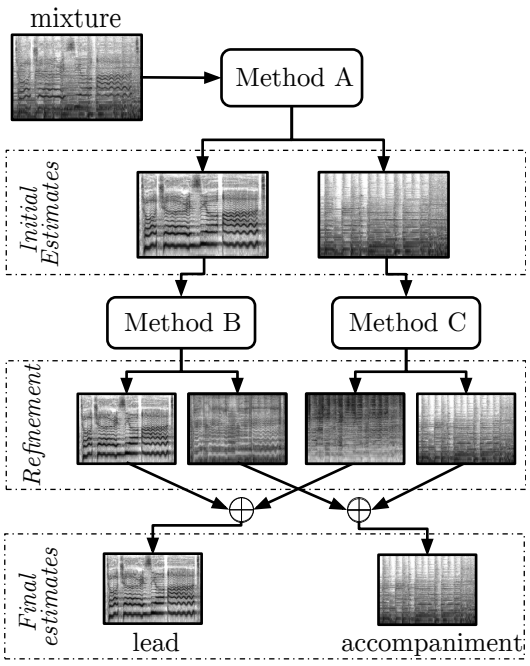


Fig. 8: Cascading source separation methods. The results from method A is improved by applying methods B and C on its output, which are specialized in reducing interferences from undesired sources in each signal.

Finally, this cascading concept was considered again by Driedger and Müller in [231], that introduces a processing pipeline for the outputs of different methods [115], [164], [232], [233] to obtain an improved separation quality. Their core idea is depicted in Figure 8 and combines the output of different methods in a specific order to improve separation.

Another approach for improving the quality of separation when using several separation procedures is not to restrict the number of such iterations from one method to another, but rather to iterate them many times until satisfactory results are obtained. This is what is proposed in Hsu et al. in [234], extending the algorithm in [235]. They first estimated the pitch range of the singing voice by using the HPSS method in [160], [225]. They separated the voice given the estimated pitch using a binary mask obtained by training a multilayer perceptron [236] and re-estimated the pitch given the separated voice. Voice separation and pitch estimation are then iterated until convergence.

As another iterative method, Zhu et al. proposed a multi-stage NMF [230], using harmonic and percussive separation at different frequency resolutions similar to [225] and [216]. The main originality of their contribution was to iterate the refinements instead of applying it only once.

An issue with such iterated methods lies in how to decide whether convergence is obtained, and it is not clear whether the quality of the separated signals will necessarily improve. For this reason, Bryan and Mysore proposed a user-guided approach based on PLCA, which can be applied for the separation of the vocals [237]–[239]. They allowed a user to make annotations on the spectrogram of a mixture, incorporated the

feedback as constraints in a PLCA model [110], [156], and used a posterior regularization technique [240] to refine the estimates, repeating the process until the user is satisfied with the results. This is similar to the way Ozerov et al. proposed to take user input into account in [241].

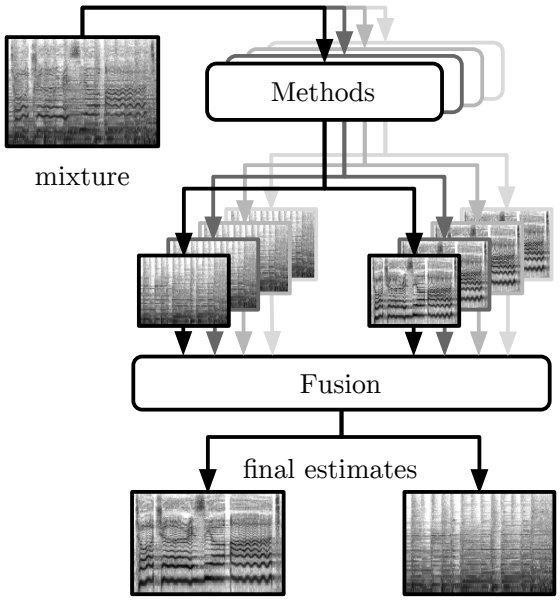


Fig. 9: Fusion of separation methods. The output of many separation methods is fed into a fusion system that combines them to produce a single estimate.

A principled way to aggregate the result of many source separation systems to obtain one single estimate that is consistently better than all of them was presented by Jaureguiberry et al. in their *fusion framework*, depicted in Figure 9. It takes advantage of multiple existing approaches, and demonstrated its application to singing voice separation [242]–[244]. They investigated fusion methods based on non-linear optimization, Bayesian model averaging [245], and deep neural networks (DNN).

As another attempt to design an efficient fusion method, McVicar et al. proposed in [246] to combine the outputs of RPCA [115], HPSS [216], Gabor filtered spectrograms [247], REPET [130] and an approach based on deep learning [248]. To do this, they used different classification techniques to build the aggregated TF mask, such as a logistic regression model or a conditional random field (CRF) trained using the method in [249] with time and/or frequency dependencies.

Manilow et al. trained a neural network to predict quality of source separation for three source separation algorithms, each leveraging a different cue - repetition, spatialization, and harmonicity/pitch proximity [250]. The method estimates separation quality of the lead vocals for each algorithm, using only the original audio mixture and separated source output. These estimates were used to guide switching between algorithms along time.

F. Source-dependent representations

In the previous section, we stated that some authors considered iterating separation at different frequency resolutions, i.e., using different TF representations [216], [224], [229]. This can be seen as a combination of different methods. However, this can also be seen from another perspective as based on picking specific *representations*.

Wolf et al. proposed an approach using rigid motion segmentation, with application to singing voice separation [251], [252]. They introduced harmonic template models with amplitude and pitch modulations defined by a velocity vector. They applied a wavelet transform [253] on the harmonic template models to build an audio image where the amplitude and pitch dynamics can be separated through the velocity vector. They then derived a velocity equation, similar to the optical flow velocity equation used in images [254], to segment velocity components. Finally, they identified the harmonic templates which model different sources in the mixture and separated them by approximating the velocity field over the corresponding harmonic template models.

Yen et al. proposed an approach using spectro-temporal modulation features [255], [256]. They decomposed a mixture using a two-stage auditory model which consists of a cochlear module [257] and cortical module [258]. They then extracted spectro-temporal modulation features from the TF units and clustered the TF units into harmonic, percussive, and vocal components using the EM algorithm and resynthesized the estimated signals.

Chan and Yang proposed an approach using an informed group sparse representation [259]. They introduced a representation built using a learned dictionary based on a chord sequence which exhibits group sparsity [260] and which can incorporate melody annotations. They derived a formulation of the problem in a manner similar to RPCA and solved it using the alternating direction method of multipliers [261]. They also showed a relation between their representation and the low-rank representation in [123], [262].

G. Shortcomings

The large body of literature we reviewed in the preceding sections is concentrated on choosing adequate models for the lead and accompaniment parts of music signals in order to devise effective signal processing methods to achieve separation. From a higher perspective, their common feature is to guide the separation process in a *model-based way*: first, the scientist has some idea regarding characteristics of the lead signal and/or the accompaniment, and then an algorithm is designed to exploit this knowledge for separation.

Model-based methods for lead and accompaniment separation are faced with a common risk that their core assumptions will be violated for the signal under study. For instance, the lead to be separated may not be harmonic but saturated vocals or the accompaniment may not be repetitive or redundant, but rather always changing. In such cases, model-based methods are prone to large errors and poor performance.

VI. Data-driven approaches

A way to address the potential caveats of model-based separation behaving badly in case of violated assumptions is to avoid making assumptions altogether, but rather to let the model be learned from a large and representative database of examples. This line of research leads to *data-driven* methods, for which researchers are concerned about directly estimating a mapping between the mixture and either the TF mask for separating the sources, or their spectrograms to be used for designing a filter.

As may be foreseen, this strategy based on machine learning comes with several challenges of its own. First, it requires considerable amounts of data. Second, it typically requires a high-capacity learner (many tunable parameters) that can be prone to over-fitting the training data and therefore not working well on the audio it faces when deployed.

A. Datasets

Building a good data-driven method for source separation relies heavily on a training dataset to learn the separation model. In our case, this not only means obtaining a set of musical songs, but also their constitutive accompaniment and lead sources, summing up to the mixtures. For professionally-produced or recorded music, the separated sources are often either unavailable or private. Indeed, they are considered amongst the most precious assets of right holders, and it is very difficult to find isolated vocals and accompaniment of professional bands that are freely available for the research community to work on without copyright infringements.

Another difficulty arises when considering that the different sources in a musical content do share some common orchestration and are not superimposed in a random way, prohibiting simply summing isolated random notes from instrumental databases to produce mixtures. This contrasts with the speech community which routinely generates mixtures by summing noise data [263] and clean speech [264].

Furthermore, the temporal structures in music signals typically spread over long periods of time and can be exploited to achieve better separation. Additionally, short excerpts do not often comprise parts where the lead signal is absent, although a method should learn to deal with that situation. This all suggests that including full songs in the training data is preferable over short excerpts.

Finally, professional recordings typically undergo sophisticated sound processing where panning, reverberation, and other sound effects are applied to each source separately, and also to the mixture. To date, simulated data sets have poorly mimicked these effects [265]. Many separation methods make assumptions about the mixing model of the sources, e.g., assuming it is linear (i.e., does not comprise effects such as dynamic range compression). It is quite common that methods giving extremely good performance for linear mixtures completely break down when processing published musical recordings. Training and test data should thus feature realistic audio engineering to be useful for actual applications.

In this context, the development of datasets for lead and accompaniment separation was a long process. In early times, it was common for researchers to test their methods on some private data. To the best of our knowledge, the first attempt at releasing a public dataset for evaluating vocals and accompaniment separation was the Music Audio Signal Separation (MASS) dataset [266]. It strongly boosted research in the area, even if it only featured 2.5 minutes of data. The breakthrough was made possible by some artists which made their mixed-down audio, as well as its constitutive stems (unmixed tracks), available under open licenses such as Creative Commons, or authorized scientists to use their material for research.

The MASS dataset then formed the core content of the early Signal Separation Evaluation Campaigns (SiSEC) [267], which evaluate the quality of various music separation methods [268]–[272]. SiSEC always had a strong focus on vocals and accompaniment separation. For a long time, vocals separation methods were very demanding computationally and it was already considered extremely challenging to separate excerpts of only a few seconds.

In the following years, new datasets were proposed that improved over the MASS dataset in many directions. We briefly describe the most important ones, summarized in Table I.

- The QUASI dataset was proposed to study the impact of different mixing scenarios on the separation quality. It consists of the same tracks as in the MASS dataset, but kept full length and mixed by professional sound engineers.
- The MIR-1K and iKala datasets were the first attempts to scale vocals separation up. They feature a higher number of samples than the previously available datasets. However, they consist of mono signals of very short and amateur karaoke recordings.
- The ccMixter dataset was proposed as the first dataset to feature many full-length stereo tracks. Each one comes with a vocals and an accompaniment source. Although it is stereo, it often suffers from simplistic mixing of sources, making it unrealistic in some aspects.
- MedleyDB has been developed as a dataset to serve many purposes in music information retrieval. It consists of more than 100 full-length recordings, with all their constitutive sources. It is the first dataset to provide such a large amount of data to be used for audio separation research (more than 7 hours). Among all the material present in that dataset, 63 tracks feature singing voice.
- DSD100 was presented for SiSEC 2016. It features 100 full-length tracks originating from the 'Mixing Secret' Free Multitrack Download Library[1] of the Cambridge Music Technology, which is freely usable for research and educational purposes.

Finally, we present here the MUSDB18 dataset, putting together tracks from MedleyDB, DSD100, and other new musical material. It features 150 full-length tracks, and has been constructed by the authors of this paper so as to address all the limitations we identified above:

- It only features full-length tracks, so that the handling of long-term musical structures, and of silent regions in the lead/vocal signal, can be evaluated.

[1] http://www.cambridge-mt.com/ms-mtk.htm

- It only features stereo signals which were mixed using professional digital audio workstations. This results in quality stereo mixes which are representative of real application scenarios.
- As with DSD100, a design choice of MUSDB18 was to split the signals into 4 predefined categories: bass, drums, vocals, and other. This contrasts with the enhanced granularity of MedleyDB that offers more types of sources, but it strongly promotes automation of the algorithms.
- Many musical genres are represented in MUSDB18, for example, jazz, electro, metal, etc.
- It is split into a development (100 tracks, 6.5 h) and a test dataset (50 tracks, 3.5 h), for the design of data-driven separation methods.

All details about this freely available dataset and its accompanying software tools may be found in its dedicated website[2].

In any case, it can be seen that datasets of sufficient duration to build data-driven separation methods were only created recently.

B. Algebraic approaches

A natural way to exploit a training database was to learn some parts of the model to guide the estimation process into better solutions. Work on this topic may be traced back to the suggestion of Ozerov et al. in [276] to learn spectral template models based on a database of isolated sources, and then to adapt this dictionary of templates on the mixture using the method in [277].

The exploitation of training data was formalized by Smaragdis et al. in [110] in the context of source separation within the supervised and semi-supervised PLCA framework. The core idea of this probabilistic formulation, equivalent to NMF, is to learn some spectral bases from the training set which are then kept fixed at separation time.

In the same line, Ozerov et al. proposed an approach using Bayesian models [191]. They first segmented a song into vocal and non-vocal parts using GMMs with MFCCs. Then, they adapted a general music model on the non-vocal parts of a particular song by using the maximum a posteriori (MAP) adaptation approach in [278]

Ozerov et al. later proposed a framework for source separation which generalizes several approaches given prior information about the problem and showed its application for singing voice separation [210]–[212]. They chose the local Gaussian model [279] as the core of the framework and allowed the prior knowledge about each source and its mixing characteristics using user-specified constraints. Estimation was performed through a generalized EM algorithm [32].

Rafii et al. proposed in [280] to address the main drawback of the repetition-based methods described in Section IV-C, which is the weakness of the model for vocals. For this purpose, they combined the REPET-SIM model [135] for the accompaniment with a NMF-based model for singing voice learned from a voice dataset.

As yet another example of using training data for NMF, Boulanger-Lewandowski et al. proposed in [281] to exploit long-term temporal dependencies in NMF, embodied using recurrent neural networks (RNN) [236]. They incorporated RNN regularization into the NMF framework to temporally constrain the activity matrix during the decomposition, which can be seen as a generalization of the non-negative HMM in [282]. Furthermore, they used supervised and semi-supervised NMF algorithms on isolated sources to train the models, as in [110].

[2]https://sigsep.github.io/musdb

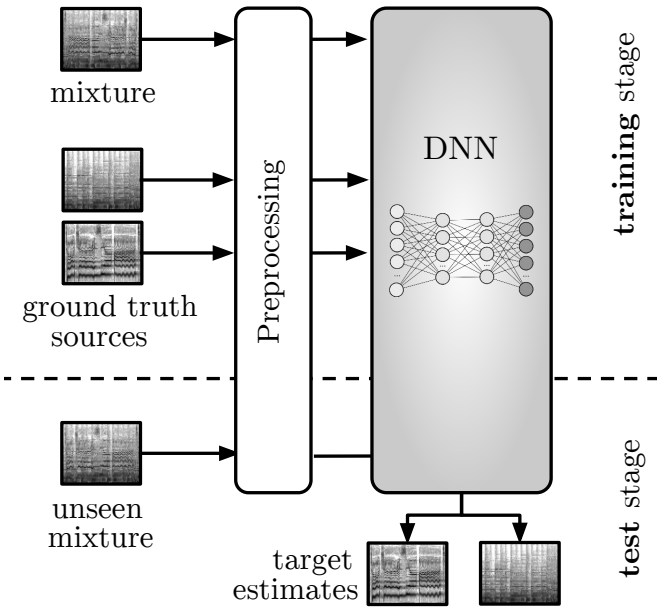


Fig. 10: General architecture for methods exploiting deep learning. The network inputs the mixture and outputs either the sources spectrograms or a TF mask. Methods usually differ in their choice for a network architecture and the way it is learned using the training data.

C. Deep neural networks

Taking advantage of the recent availability of sufficiently large databases of isolated vocals along with their accompaniment, several researchers investigated the use of machine learning methods to directly estimate a mapping between the mixture and the sources. Although end-to-end systems inputting and outputting the waveforms have already been proposed in the speech community [283], they are not yet available for music source separation. This may be due to the relative small size of music separation databases, at most 10 h today. Instead, most systems feature pre and post-processing steps that consist in computing classical TF representations and building TF masks, respectively. Although such end-to-end systems will inevitably be proposed in the near future, the common structure of deep learning methods for lead and accompaniment separation usually corresponds for now to the one depicted in Figure 10. From a general perspective, we may say that most current methods mainly differ in the structure picked for the network, as well as in the way it is learned.

Providing a thorough introduction to deep neural networks is out of the scope of this paper. For our purpose, it suffices

TABLE I: Summary of datasets available for lead and accompaniment separation. Tracks without vocals were omitted in the statistics.

Dataset	Year	Reference(s)	URL	Tracks	Track duration (s)	Full/stereo?
MASS	2008	[266]	http://www.mtg.upf.edu/download/datasets/mass	9	16 ± 7	no / yes
MIR-1K	2010	[74]	https://sites.google.com/site/unvoicedsoundseparation/mir-1k	1,000	8 ± 8	no / no
QUASI	2011	[270], [273]	http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/	5	206 ± 21	yes / yes
ccMixter	2014	[141]	http://www.loria.fr/~aliutkus/kam/	50	231 ± 77	yes / yes
MedleyDB	2014	[274]	http://medleydb.weebly.com/	63	206 ± 121	yes / yes
iKala	2015	[162]	http://mac.citi.sinica.edu.tw/ikala/	206	30	no / no
DSD100	2015	[271]	sisec17.audiolabs-erlangen.de	100	251 ± 60	yes / yes
MUSDB18	2017	[275]	https://sigsep.github.io/musdb	150	236 ± 95	yes / yes

to mention that they consist of a cascade of several possibly non-linear transformations of the input, which are learned during a training stage. They were shown to effectively learn representations and mappings, provided enough data is available for estimating their parameters [284]–[286]. Different architectures for neural networks may be combined/cascaded together, and many architectures were proposed in the past, such as feedforward fully-connected neural networks (FNN), convolutional neural networks (CNN), or RNN and variants such as the long short-term memory (LSTM) and the gated-recurrent units (GRU). Training of such functions is achieved by stochastic gradient descent [287] and associated algorithms, such as backpropagation [288] or backpropagation through time [236] for the case of RNNs.

To the best of our knowledge, Huang et al. were the first to propose deep neural networks, RNNs here [289], [290], for singing voice separation in [248], [291]. They adapted their framework from [292] to model all sources simultaneously through masking. Input and target functions were the mixture magnitude and a joint representation of the individual sources. The objective was to estimate jointly either singing voice and accompaniment music, or speech and background noise from the corresponding mixtures.

Modeling the temporal structures of both the lead and the accompaniment is a considerable challenge, even when using DNN methods. As an alternative to the RNN approach proposed by Huang et al. in [248], Uhlich et al. proposed the usage of FNNs [293] whose input consists of *supervectors* of a few consecutive frames from the mixture spectrogram. Later in [294], the same authors considered the use of bi-directional LSTMs for the same task.

In an effort to make the resulting system less computationally demanding at separation time but still incorporating dynamic modeling of audio, Simpson et al. proposed in [295] to predict binary TF masks using deep CNNs, which typically utilize fewer parameters than the FNNs. Similarly, Schlueter proposed a method trained to detect singing voice using CNNs [296]. In that case, the trained network was used to compute *saliency maps* from which TF masks can be computed for singing voice separation. Chandna et al. also considered CNNs for lead separation in [297], with a particular focus on low-latency.

The classical FNN, LSTM and CNN structures above served as baseline structures over which some others tried to improve. As a first example, Mimilakis et al. proposed to use a hybrid structure of FNNs with skip connections to separate the lead instrument for purposes of remixing jazz recordings [298]. Such skip connections allow to propagate the input spectrogram to intermediate representations within the network, and mask it similarly to the operation of TF masks. As advocated, this enforces the networks to approximate a TF masking process. Extensions to temporal data for singing voice separation were presented in [299], [300]. Similarly, Jansson et al. proposed to propagate the spectral information computed by convolutional layers to intermediate representations [301]. This propagation aggregates intermediate outputs to proceeding layer(s). The output of the last layer is responsible for masking the input mixture spectrogram. In the same vein, Takahashi et al. proposed to use skip connections via element-wise addition through representations computed by CNNs [302].

Apart from the structure of the network, the way it is trained, comprising how the targets are computed, has a tremendous impact on performance. As we saw, most methods operate on defining TF masks or estimating magnitude spectrograms. However, other methods were proposed based on deep clustering [303], [304], where TF mask estimation is seen as a clustering problem. Luo et al. investigated both approaches in [305] by proposing deep bidirectional LSTM networks capable of outputting both TF masks or features to use as in deep clustering. Kim and Smaragdis proposed in [306] another way to learn the model, in a denoising auto-encoding fashion [307], again utilizing short segments of the mixture spectrogram as an input to the network, as in [293].

As the best network structure may vary from one track to another, some authors considered a fusion of methods, in a manner similar to the method [242] presented above. Grais et. al [308], [309] proposed to aggregate the results from an ensemble of feedforward DNNs to predict TF masks for separation. An improvement was presented in [310], [311] where the inputs to the fusion network were separated signals, instead of TF masks, aiming at enhancing the reconstruction of the separated sources.

As can be seen the use of deep learning methods for the design of lead and accompaniment separation has already stimulated a lot of research, although it is still in its infancy. Interestingly, we also note that using audio and music specific knowledge appears to be fundamental in designing effective systems. As an example of this, the contribution from Nie et al. in [312] was to include the construction of the TF mask as an extra non-linearity included in a recurrent network. This is an exemplar of where signal processing elements, such as filtering through masking, are incorporated as a building block of the machine learning method.

The network structure is not the only thing that can benefit from audio knowledge for better separation. The design of appropriate features is another. While we saw that supervectors of spectrogram patches offered the ability to effectively model time-context information in FNNs [293], Sebastian and Murthy [313] proposed the use of the modified group delay feature representation [314] in their deep RNN architecture. They applied their approach for both singing voice and vocal-violin separation.

Finally, as with other methods, DNN-based separation techniques can also be combined with others to yield improved performance. As an example, Fan et al. proposed to use DNNs to separate the singing voice and to also exploit vocal pitch estimation [315]. They first extracted the singing voice using feedforward DNNs with sigmoid activation functions. They then estimated the vocal pitch from the extracted singing voice using dynamic programming.

D. Shortcomings

Data-driven methods are nowadays the topic of important research efforts, particularly those based on DNNs. This is notably due to their impressive performance in terms of separation quality, as can, for instance, be noticed below in Section VIII. However, they also come with some limitations.

First, we highlighted that lead and accompaniment separation in music has the very specific problem of scarce data. Since it is very hard to gather large amounts of training data for that application, it is hard to fully exploit learning methods that require large training sets. This raises very specific challenges in terms of machine learning.

Second, the lack of interpretability of model parameters is often mentioned as a significant shortcoming when it comes to applications. Indeed, music engineering systems are characterized by a strong importance of human-computer interactions, because they are used in an artistic context that may require specific needs or results. As of today, it is unclear how to provide user interaction for controlling the millions of parameters of DNN-based systems.

VII. Including multichannel information

In describing the above methods, we have not discussed the fact that music signals are typically stereophonic. On the contrary, the bulk of methods we discussed focused on designing good spectrogram models for the purpose of filtering mixtures that may be *monophonic*. Such a strategy is called *single-channel* source separation and is usually presented as more challenging than multichannel source separation. Indeed, only TF structure may then be used to discriminate the accompaniment from the lead. In stereo recordings, one further so-called *spatial* dimension is introduced, which is sometimes referred to as *pan*, that corresponds to the perceived *position* of a source in the stereo field. Devising methods to exploit this spatial diversity for source separation has also been the topic of an important body of research that we review now.

A. Extracting the lead based on panning

In the case of popular music signals, a fact of paramount practical importance is that the lead signal — such as vocals — is very often mixed *in the center*, which means that its energy is approximately the same in left and right channels. On the contrary, other instruments are often mixed at positions to the left or right of the stereo field.

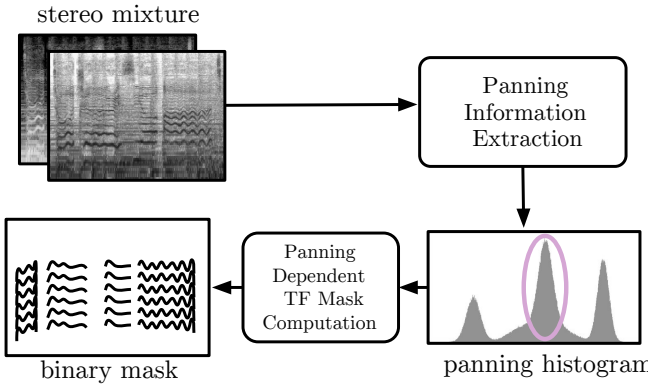


Fig. 11: Separation of the lead based on panning information. A stereo cue called panning allows to design a TF mask.

The general structure of methods extracting the lead based on stereo cues is displayed on Figure 11, introduced by Avendano, who proposed to separate sources in stereo mixtures by using a panning index [316]. He derived a two-dimensional map by comparing left and right channels in the TF domain to identify the different sources based on their panning position [317]. The same methodology was considered by Barry et al. in [318] in his Azimuth Discrimination and Resynthesis (ADRess) approach, with panning indexes computed with differences instead of ratios.

Vinyes et al. also proposed to unmix commercially produced music recordings thanks to stereo cues [319]. They designed an interface similar to [318] where a user can set some parameters to generate different TF filters in real time. They showed applications for extracting various instruments, including vocals.

Cobos and López proposed to separate sources in stereo mixtures by using TF masking and multilevel thresholding [320]. They based their approach on the Degenerate Unmixing Estimation Technique (DUET) [321]. They first derived histograms by measuring the amplitude relationship between TF points in left and right channels. Then, they obtained several thresholds using the multilevel extension of Otsu's method [322]. Finally, TF points were assigned to their related sources to produce TF masks.

Sofianos et al. proposed to separate the singing voice from a stereo mixture using ICA [323]–[325]. They assumed that most commercial songs have the vocals panned to the center and that they dominate the other sources in amplitude. In [323], they proposed to combine a modified version of ADRess with ICA to filter out the other instruments. In [324], they proposed a modified version without ADRess.

Kim et al. proposed to separate centered singing voice in stereo music by exploiting binaural cues, such as inter-channel level and inter-channel phase difference [326]. To this end,

they build the pan-based TF mask through an EM algorithm, exploiting a GMM model on these cues.

B. Augmenting models with stereo

As with using only a harmonic model for the lead signal, using stereo cues in isolation is not always sufficient for good separation, as there can often be multiple sources at the same spatial location. Combining stereo cues with other methods improves performance in these cases.

Cobos and López proposed to extract singing voice by combining panning information and pitch tracking [327]. They first obtained an estimate for the lead thanks to a pan-based method such as [316], and refined the singing voice by using a TF binary mask based on comb-filtering method as in Section III-B. The same combination was proposed by Marxer et al. in [87] in a low-latency context, with different methods used for the binaural cues and pitch tracking blocks.

FitzGerald proposed to combine approaches based on repetition and panning to extract stereo vocals [328]. He first used his nearest neighbors median filtering algorithm [139] to separate vocals and accompaniment from a stereo mixture. He then used the ADRess algorithm [318] and a high-pass filter to refine the vocals and improve the accompaniment. In a somewhat different manner, FitzGerald and Jaiswal also proposed to combine approaches based on repetition and panning to improve stereo accompaniment recovery [329]. They presented an audio inpainting scheme [330] based on the nearest neighbors and median filtering algorithm [139] to recover TF regions of the accompaniment assigned to the vocals after using a source separation algorithm based on panning information.

In a more theoretically grounded manner, several methods based on a probabilistic model were generalized to the multichannel case. For instance, Durrieu et al. extended their source-filter model in [201], [205] to handle stereo signals, by incorporating the panning coefficients as model parameters to be estimated.

Ozerov and Févotte proposed a multichannel NMF framework with application to source separation, including vocals and music [331], [332]. They adopted a statistical model where each source is represented as a sum of Gaussian components [193], and where maximum likelihood estimation of the parameters is equivalent to NMF with the Itakura-Saito divergence [94]. They proposed two methods for estimating the parameters of their model, one that maximized the likelihood of the multichannel data using EM, and one that maximized the sum of the likelihoods of all channels using a multiplicative update algorithm inspired by NMF [90].

Ozerov et al. then proposed a multichannel non-negative tensor factorization (NTF) model with application to user-guided source separation [333]. They modeled the sources jointly by a 3-valence tensor (time/frequency/source) as in [334] which extends the multichannel NMF model in [332]. They used a generalized EM algorithm based on multiplicative updates [335] to minimize the objective function. They incorporated information about the temporal segmentation of the tracks and the number of components per track. Ozerov et al. later proposed weighted variants of NMF and NTF with application to user-guided source separation, including separation of vocals and music [241], [336].

Sawada et al. also proposed multichannel extensions of NMF, tested for separating stereo mixtures of multiple sources, including vocals and accompaniment [337]–[339]. They first defined multichannel extensions of the cost function, namely, Euclidean distance and Itakura-Saito divergence, and derived multiplicative update rules accordingly. They then proposed two techniques for clustering the bases, one built into the NMF model and one performing sequential pair-wise merges.

Finally, multichannel information was also used with DNN models. Nugraha et al. addressed the problem of multichannel source separation for speech enhancement [340], [341] and music separation [342], [343]. In this framework, DNNs are still used for the spectrograms, while more classical EM algorithms [344], [345] are used for estimating the spatial parameters.

C. Shortcomings

When compared to simply processing the different channels independently, incorporating spatial information in the separation method often comes at the cost of additional computational complexity. The resulting methods are indeed usually more demanding in terms of computing power, because they involve the design of beamforming filters and inversion of covariance matrices. While this is not really an issue for stereophonic music, this may become prohibiting in configurations with higher numbers of channels.

VIII. Evaluation

A. Background

The problem of evaluating the quality of audio signals is a research topic of its own, which is deeply connected to psychoacoustics [346] and has many applications in engineering because it provides an objective function to optimize when designing processing methods. While mean squared error (MSE) is often used for mathematical convenience whenever an error is to be computed, it is a very established fact that MSE is not representative of audio perception [347], [348]. For example, inaudible phase shifts would dramatically increase the MSE. Moreover, it should be acknowledged that the concept of quality is rather application-dependent.

In the case of signal separation or enhancement, processing is often only a part of a whole architecture and a relevant methodology for evaluation is to study the positive or negative impact of this module on the overall performance of the system, rather than to consider it independently from the rest. For example, when embedded in an automatic speech recognition (ASR) system, performance of speech denoising can be assessed by checking whether it decreases word error rate [349].

When it comes to music processing, and more particularly to lead and accompaniment separation, the evaluation of separation quality has traditionally been inspired by work in the audio coding community [347], [350] in the sense that it aims at comparing ground truth vocals and accompaniment with

their estimates, just like audio coding compares the original with the compressed signal.

B. Metrics

As noted previously, MSE-based error measures are not perceptually relevant. For this reason, a natural approach is to have humans do the comparison. The gold-standard for human perceptual studies is the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) methodology, that is commonly used for evaluating audio coding [350].

However, it quickly became clear that the specific evaluation of separation quality cannot easily be reduced to a single number, even when achieved through actual perceptual campaigns, but that quality rather depends on the application considered. For instance, karaoke or vocal extraction come with opposing trade-offs between isolation and distortion. For this reason, it has been standard practice to provide different and complementary metrics for evaluating separation that measure the amount of distortion, artifacts, and interference in the results.

While human-based perceptual evaluation is definitely the best way to assess separation quality [351], [352], having computable objective metrics is desirable for several reasons. First, it allows researchers to evaluate performance without setting up costly and lengthy perceptual evaluation campaigns. Second, it permits large-scale training for the fine-tuning of parameters. In this respect, the Blind Source Separation Evaluation (BSS Eval) toolbox [353], [354] provides quality metrics in decibel to account for distortion (SDR), artifacts (SAR), and interferences (SIR). Since it was made available quite early and provides somewhat reasonable correlation with human perception in certain cases [355], [356] it is still widely used to this day.

Even if BSS Eval was considered sufficient for evaluation purposes for a long time, it is based on squared error criteria. Following early work in the area [357], the Perceptual Evaluation of Audio Source Separation (PEASS) toolkit [358]–[360] was introduced as a way to predict perceptual ratings. While the methodology is very relevant, PEASS however was not widely accepted in practice. We believe this is for two reasons. First, the proposed implementation is quite computationally demanding. Second, the perceptual scores it was designed with are more related to speech separation than to music.

Improving perceptual evaluation often requires a large amount of experiments, which is both costly and requires many expert listeners. One way to increase the number of participants is to conduct web-based experiments. In [361], the authors report they were able to aggregate 530 participants in only 8.2 hours and obtained perceptual evaluation scores comparable to those estimated in the controlled lab environment.

Finally, we highlight here that the development of new perceptually relevant objective metrics for singing voice separation evaluation remains an open issue [362]. It is also a highly crucial one for future research in the domain.

C. Performance (SiSEC 2016)

In this section, we will discuss the performance of 23 source separation methods evaluated on the DSD100, as part of the task for separating professionally-produced music recordings at SiSEC 2016. The methods are listed in Table II, along with the acronyms we use for them, their main references, a very brief summary, and a link to the section where they are described in the text. To date, this stands as the largest evaluation campaign ever achieved on lead and accompaniment separation. The results we discuss here are a more detailed report for SiSEC 2016 [272], presented in line with the taxonomy proposed in this paper.

TABLE II: Methods evaluated during SiSEC 2016.

Acronym	Ref.	Summary	Section
HUA	[115]	RPCA standard version	IV-B
RAF1	[130]	REPET standard version	IV-C
RAF2	[134]	REPET with time-varying period	
RAF3	[135]	REPET with similarity matrix	
KAM1-2	[142]	KAM with different configurations	
CHA	[162]	RPCA with vocal activation information	V-A
JEO1-2	[163]	l_1-RPCA with vocal activation information	
DUR	[201]	Source-filter NMF	V-C
OZE	[212]	Structured NMF with learned dictionaries	VI-B
KON	[291]	RNN	VI-C
GRA2-3	[308]	DNN ensemble	
STO1-2	[363]	FNN on *common fate* TF representation	
UHL1	[293]	FNN with context	
NUG1-4	[343]	FNN with multichannel information	VII
UHL2-3	[294]	LSTM with multichannel information	
IBM		ideal binary mask	

The objective scores for these methods were obtained using BSS Eval and are given in Figure 12. For more details about the results and for listening to the estimates, we refer the reader to the dedicated interactive website[3].

As we first notice in Figure 12, the HUA method, corresponding to the standard RPCA as discussed in Section IV-B, showed rather disappointing performance in this evaluation. After inspection of the results, it appears that processing full-length tracks is the issue there: at such scales, vocals also exhibit redundancy, which is captured by the low-rank model associated with the accompaniment. On the other hand, the RAF1-3 and KAM1-3 methods that exploit redundancy through repetitions, as presented in Section IV-C, behave much better for full-length tracks: even if somewhat redundant, vocals are rarely as repetitive as the accompaniment. When those methods are evaluated on datasets with very short excerpts (e.g., MIR-1K), such severe practical drawbacks are not apparent.

Likewise, the DUR method that jointly models the vocals as harmonic and the accompaniment as redundant, as discussed in Section V-C, does show rather disappointing performance, considering that it was long the state-of-the-art in earlier SiSECs [270]. After inspection, we may propose two reasons for this performance drop. First, using full-length excerpts also clearly revealed a shortcoming of the approach: it poorly handles silences in the lead, which were rare in the short-length excerpts tested so far. Second, using a much larger evaluation set revealed that vocals are not necessarily well modeled by a harmonic source-filter model; breathy or saturated voices appear to greatly challenge such a model.

[3]http://www.sisec17.audiolabs-erlangen.de

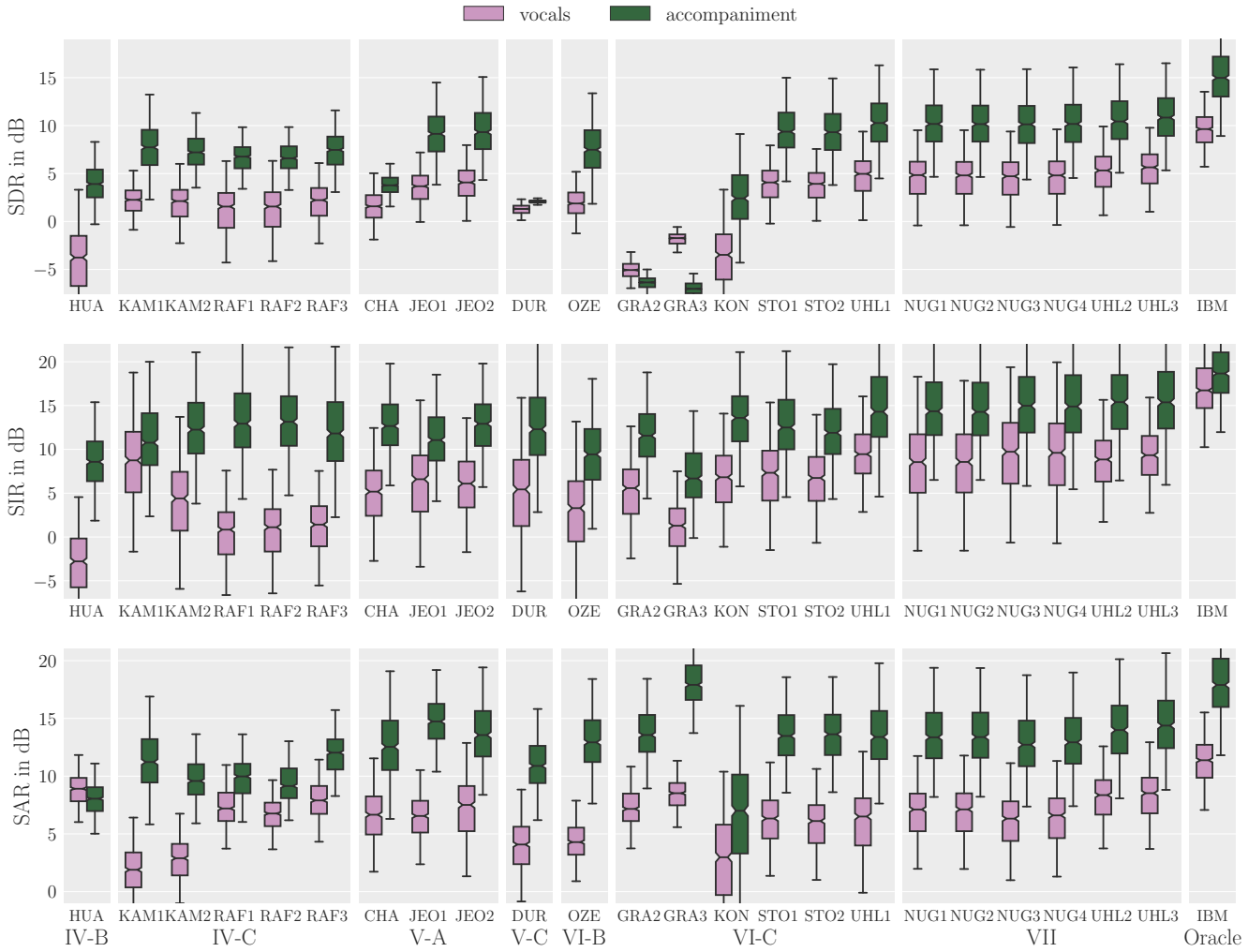


Fig. 12: BSS Eval scores for the vocals and accompaniment estimates for SiSEC 2016 on the DSD100 dataset. Results are shown for the *test* set only. Scores are grouped as in Table II according to the section they are described in the text, indicated below each group.

While processing full-length tracks comes as a challenge, it can also be an opportunity. It is indeed worth noticing that whenever RPCA is helped through vocal activity detection, its performance is significantly boosted, as highlighted by the relatively good results obtained by CHAN and JEO.

As discussed in Section VI, the availability of learning data made it possible to build data-driven approaches, like the NMF-based OZE method which is available through the Flexible Audio Source Separation Toolbox (FASST) [211], [212]. Although it was long state-of-the-art, it has been strongly outperformed recently by other data-driven approaches, namely DNNs. One first reason clearly appears as the superior expressive power of DNNs over NMF, but one second reason could very simply be that OZE should be trained anew with the same large amount of data.

As mentioned above, a striking fact we see in Figure 12 is that the overall performance of data-driven DNN methods is the highest. This shows that exploiting learning data does help separation greatly compared to only relying on *a priori* assumptions such as the harmonicity or redundancy. Additionally, dynamic models such as CNN or LSTM appear more adapted to music than FNN. These good performances in audio source separation go in line with the recent success of DNNs in fields as varied as computer vision, speech recognition, and natural language processing [285].

However, the picture may be seen to be more subtle than simply black-box DNN systems beating all other approaches. For instance, exploiting multichannel probabilistic models, as discussed in Section VII, leads to the NUG and UHL2-3 methods, that significantly outperform the DNN methods ignoring stereo information. In the same vein, we expect other specific assumptions and musicological ideas to be exploited for further improving the quality of the separation.

One particular feature of this evaluation is that it also shows obvious weaknesses in the objective metrics. For instance, the GRA method behaves significantly worse than any other methods. However, when listening to the separated signals, this does not seem deserved. All in all, designing new and

convenient metrics that better match perception and that are specifically built for music on large datasets clearly appears as a desirable milestone.

In any case, the performance achieved by a totally informed filtering method such as IBM is significantly higher than that of any submitted method in this evaluation. This means that lead and accompaniment separation has room for much improvement, and that the topic is bound to witness many breakthroughs still. This is even more true considering that IBM is not the best upper bound for separation performance: other filtering methods such as *ideal ratio mask* [20] or multichannel Wiener filter [344] may be considered as references.

Regardless of the above, we would also like to highlight that good algorithms and models can suffer from slight errors in their low-level audio processing routines. Such routines may include the STFT representation, the overlap-add procedure, energy normalization, and so on. Considerable improvements may also be obtained by using simple tricks and, depending on the method, large impacts can occur in the results by only changing low-level parameters. These include the overlap ratio for the STFT, specific ways to regularize matrix inverses in multichannel models, etc. Further tricks such as the exponentiation of the TF mask by some positive value can often boost performance significantly more than using more sophisticated models. However, such tricks are often lost when publishing research focused on the higher-level algorithms. We believe this is an important reason why sharing source code is highly desirable in this particular application. Some online repositories containing implementations of lead and accompaniment separation methods should be mentioned, such as **nussl**[4] and **untwist** [364]. In the companion webpage of this paper[5], we list many different online resources such as datasets, implementations, and tools that we hope will be useful to the practitioner and provide some useful pointers to the interested reader.

D. Discussion

Finally, we summarize the core advantages and disadvantages for each one of the five groups of methods we identified.

Methods based on the harmonicity assumption for the lead are focused on sinusoidal modeling. They enjoy a very strong interpretability and allow for the direct incorporation of any prior knowledge concerning pitch. Their fundamental weakness lies in the fact that many singing voice signals are not harmonic, e.g., when breathy or distorted.

Modeling the accompaniment as redundant allows to exploit long-term dependencies in music signals and may benefit from high-level information like tempo or score. Their most important drawback is to fall short in terms of voice models: the lead signal itself is often redundant to some extent and thus partly incorporated in the estimated accompaniment.

Systems jointly modeling the lead as harmonic and the accompaniment as redundant benefit from both assumptions. They were long state-of-the-art and enjoy a good interpretability, which makes them good candidates for interactive separation methods. However, their core shortcoming is to be highly sensitive to violations of their assumptions, which proves to often be the case in practice. Such situations usually require fine-tuning and hence prevents their use as black-box systems for a broad audience.

Data-driven methods involve machine learning to directly learn a mapping between the mixture and the constitutive sources. Such a strategy recently introduced a breakthrough compared to everything that was done before. Their most important disadvantages are the lack of interpretability, which makes it challenging to design good user interactions, as well as their strong dependency on the size of the training data.

Finally, multichannel methods leverage stereophonic information to strongly improve performance. Interestingly, this can usually be combined with better spectrogram models such as DNNs to further improve quality. The price to pay for this boost in performance is an additional computational cost, that may be prohibitive for recordings of more than two channels.

IX. Conclusion

In this paper, we thoroughly discussed the problem of separating lead and accompaniment signals in music recordings. We gave a comprehensive overview of the research undertaken in the last 50 years on this topic, classifying the different approaches according to their main features and assumptions. In doing so, we showed how one very large body of research can be described as being model-based. In this context, it was evident from the literature that the two most important assumptions behind these models are that the lead instrument is harmonic, while the accompaniment is redundant. As we demonstrated, a very large number of methods on model-based lead-accompaniment separation can be seen as using one or both of these assumptions.

However, music encompasses a variety of signals of an extraordinary diversity, and no rigid assumption holds well for all signals. For this reason, while there are often some music pieces where each method performs well, there will also be some where it fails. As a result, data-driven methods were proposed as an attempt to introduce more flexibility at the cost of requiring representative training data. In the context of this paper, we proposed the largest freely available dataset for music separation, comprising close to 10 hours of data, which is 240 times greater than the first public dataset released 10 years ago.

At present, we see a huge focus on research utilizing recent machine learning breakthroughs for the design of singing voice separation methods. This came with an associated boost in performance, as measured by objective metrics. However, we have also discussed the strengths and shortcomings of existing evaluations and metrics. In this respect, it is important to note that the songs used for evaluation are but a minuscule fraction of all recorded music, and that separating music signals remains the processing of an artistic means of expression. As such it is impossible to escape the need for human perceptual evaluations, or at least adequate models for it.

After reviewing the large existing body of literature, we may conclude here by saying that lead and accompaniment

[4]https://github.com/interactiveaudiolab/nussl
[5]https://sigsep.github.io

separation in music is a problem at the crossroads of many different paradigms and methods. Researchers from very different backgrounds such as physics, signal or computer engineering have tackled it, and it exists both as an area for strong theoretical research and as a real-world challenging engineering problem. Its strong connections with the arts and digital humanities have proved attractive to many researchers.

Finally, as we showed, there is still much room for improvement in lead and accompaniment separation, and we believe that new and exciting research will bring new breakthroughs in this field. While DNN methods represent the latest big step forward and significantly outperform previous research, we believe that future improvements can come from any direction, including those discussed in this paper. Still, we expect future improvements to initially come from improved machine learning methodologies that can cope with reduced training sets, as well as improved modeling of the specific properties of musical signals, and the development of better signal representations.

REFERENCES

[1] R. Kalakota and M. Robinson, *e-Business 2.0: Roadmap for Success*. Addison-Wesley Professional, 2000.

[2] C. K. Lam and B. C. Tan, "The Internet is changing the music industry," *Communications of the ACM*, vol. 44, no. 8, pp. 62–68, 2001.

[3] P. Common and C. Jutten, *Handbook of Blind Source Separation*. Academic Press, 2010.

[4] G. R. Naik and W. Wang, *Blind Source Separation*. Springer-Verlag Berlin Heidelberg, 2014.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[6] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, Jun. 2000.

[7] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer Netherlands, 2007.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[9] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Press, 1990.

[10] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *14th International Workshop on Image Analysis for Multimedia Interactive Services*, Paris, France, Jul. 2013.

[11] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, May 2014.

[12] U. Zölzer, *DAFX - Digital Audio Effects*. Wiley, 2011.

[13] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[14] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.

[15] O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[16] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[17] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, May 2002.

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[19] N. Wiener, "Extrapolation, interpolation, and smoothing of stationary time series," 1975.

[20] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[21] G. Fant, *Acoustic Theory of Speech Production*. Walter de Gruyter, 1970.

[22] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum pseudo-autocovariance, cross-cepstrum, and saphe cracking," *Proceedings of a symposium on time series analysis*, pp. 209–243, 1963.

[23] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 296–302, 1964.

[24] ——, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[25] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[26] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *Journal of the Acoustical Society of America*, vol. 45, no. 2, pp. 458–465, 1969.

[27] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.

[28] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[29] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[30] A. J. Viterbi, "A personal history of the Viterbi algorithm," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 120–142, 2006.

[31] C. Bishop, *Neural networks for pattern recognition*. Clarendon Press, 1996.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[33] J. Salamon, E. Gómez, D. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Processing Magazine*, vol. 31, 2014.

[34] N. J. Miller, "Removal of noise from a voice signal by synthesis," Utah University, Tech. Rep., 1973.

[35] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun. 1968.

[36] R. C. Maher, "An approach for the separation of voices in composite musical signals," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1989.

[37] A. L. Wang, "Instantaneous and frequency-warped techniques for auditory source separation," Ph.D. dissertation, Stanford University, 1994.

[38] ——, "Instantaneous and frequency-warped techniques for source separation and signal parametrization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 1995.

[39] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *5th International Conference on Spoken Language Processing*, Sydney, Australia, Nov. 1998.

[40] T. F. Quatieri, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, Mar. 1992.

[41] A. Ben-Shalom and S. Dubnov, "Optimal filtering of an instrument sound in a mixed recording given approximate pitch prior," in *International Computer Music Conference*, Miami, FL, USA, Nov. 2004.

[42] S. Shalev-Shwartz, S. Dubnov, N. Friedman, and Y. Singer, "Robust temporal and spectral modeling for query by melody," in *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, Aug. 2002.

[43] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*. Swets & Zeitlinger, 1997, pp. 91–122.

[44] B. V. Veen and K. M. Buckley, "Beamforming techniques for spatial filtering," in *The Digital Signal Processing Handbook*. CRC Press, 1997, pp. 1–22.

[45] Y.-G. Zhang and C.-S. Zhang, "Separation of voice and music by harmonic structure stability analysis," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, Jul. 2005.

[46] ——, "Separation of music signals by harmonic structure modeling," in *Advances in Neural Information Processing Systems 18*. MIT Press, 2006, pp. 1617–1624.

[47] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, no. 2, pp. 155–182, Mar. 1979.

[48] Y.-G. Zhang, C.-S. Zhang, and S. Wang, "Clustering in knowledge embedded space," in *Machine Learning: ECML 2003*. Springer Berlin Heidelberg, 2003, pp. 480–491.

[49] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *6th International Conference on Music Information Retrieval*, London, UK, Sep. 2005.

[50] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, Mar. 2010.

[51] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, Sep. 2004.

[52] J. A. Moorer, "Signal processing aspects of computer music: A survey," *Proceedings of the IEEE*, vol. 65, no. 8, pp. 1108–1137, Aug. 2005.

[53] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Oct. 2007.

[54] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Oct. 2006.

[55] Z. Duan, Y.-F. Zhang, C.-S. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 766–778, May 2008.

[56] X. Rodet, "Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models," in *IEEE Time-Frequency and Time-Scale Workshop*, Coventry, UK, Aug. 1997.

[57] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *International Computer Music Conference*, Urbana, IL, USA, Aug. 1987.

[58] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaide, SA, Australia, Apr. 1994.

[59] M. Lagrange and G. Tzanetakis, "Sound source tracking and formation using normalized cuts," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, Apr. 2007.

[60] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 278–290, Feb. 2008.

[61] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[62] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *IEEE International Conference on Multimedia and Expo*, Hannover, Germany, Aug. 2008.

[63] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, Sep. 2008.

[64] Y. Ding and X. Qian, "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model," *Journal of the Audio Engineering Society*, vol. 45, no. 7/8, pp. 571–584, Jul. 1997.

[65] Y. Li and D. Wang, "Singing voice separation from monaural recordings," in *7th International Conference on Music Information Retrieval*, 2006.

[66] ——, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475–1487, May 2007.

[67] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *6th International Conference on Digital Audio Effects*, London, UK, Sep. 2003.

[68] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, USA, Mar. 2005.

[69] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 3, pp. 229–241, May 2003.

[70] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sep. 2002.

[71] Y. Han and C. Raphael, "Desoloing monaural audio using mixture models," in *7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Oct. 2007.

[72] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 793–799.

[73] C.-L. Hsu, J.-S. R. Jang, and T.-L. Tsai, "Separation of singing voice from music accompaniment with unvoiced sounds reconstruction for monaural recordings," in *AES 125th Convention*, San Francisco, CA, USA, Oct. 2008.

[74] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[75] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *9th International Conference on Digital Audio Effects*, Montreal, QC, Canada, Sep. 2006.

[76] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, USA, May 1996.

[77] C. Raphael and Y. Han, "A classifier-based approach to score-guided music audio source separation," *Computer Music Journal*, vol. 32, no. 1, pp. 51–59, 2008.

[78] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[79] E. Cano and C. Cheng, "Melody line detection and source separation in classical saxophone recordings," in *12th International Conference on Digital Audio Effects*, Como, Italy, Sep. 2009.

[80] S. Grollmisch, E. Cano, and C. Dittmar, "Songs2See: Learn to play by playing," in *AES 41st Conference: Audio for Games*, Feb. 2011, pp. P2–3.

[81] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Multimodal Music Processing*. Dagstuhl Publishing, 2012, pp. 95–120.

[82] E. Cano, C. Dittmar, and G. Schuller, "Efficient implementation of a system for solo and accompaniment separation in polyphonic music," in *20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[83] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *42nd AES Conference on Semantic Audio*, Ilmenau, Germany, Jul. 2011.

[84] E. Cano, C. Dittmar, and G. Schuller, "Re-thinking sound separation: Prior information and additivity constraints in separation algorithms," in *16th International Conference on Digital Audio Effects*, Maynooth, Ireland, Sep. 2013.

[85] E. Cano, G. Schuller, and C. Dittmar, "Pitch-informed solo and accompaniment separation towards its use in music education applications," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 23, Sep. 2014.

[86] J. J. Bosch, K. Kondo, R. Marxer, and J. Janer, "Score-informed and timbre independent lead instrument separation in real-world scenarios," in *20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[87] R. Marxer, J. Janer, and J. Bonada, "Low-latency instrument separation in polyphonic audio using timbre models," in *10th International Conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, Mar. 2012.

[88] A. Vaneph, E. McNeil, and F. Rigaud, "An automated source separation technology and its practical applications," in *140th AES Convention*, Paris, France, May 2016.

[89] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[90] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[91] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 556–562.

[92] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2003.

[93] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[94] C. Févotte, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[95] P. Common, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[96] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *6th International Conference on Music Information Retrieval*, London, UK, sep 2005.

[97] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[98] T. L. Nwe and Y. Wang, "Automatic detection of vocal segments in popular songs," in *5th International Conference for Music Information Retrieval*, Barcelona, Spain, Oct. 2004.

[99] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference*, Berlin, Germany, Sep. 2000.

[100] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in *International Conference on Advanced Technologies for Communications*, Hanoi, Vietnam, Oct. 2008.

[101] ——, "Singing voice separation in mono-channel music," in *International Symposium on Communications and Information Technologies*, Lao, China, Oct. 2008.

[102] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics*, vol. 4, pp. 1035–1038, 1963.

[103] R. Marxer and J. Janer, "A Tikhonov regularization method for spectrum decomposition in low latency audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[104] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[105] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 185–195, Jan. 2015.

[106] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. 4, pp. 1–17, Jan. 2009.

[107] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *8th International Conference on Independent Component Analysis and Signal Separation*, Paraty, Brazil, Mar. 2009.

[108] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *12th International Conference on Digital Audio Effects*, Como, Italy, Sep. 2009.

[109] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2009.

[110] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *7th International Conference on Independent Component Analysis and Signal Separation*, London, UK, Sep. 2007.

[111] T. Nakamuray and H. Kameoka, "L_p-norm non-negative matrix factorization and its application to singing voice enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[112] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Academic Press, 1970.

[113] H. Kameoka, M. Goto, and S. Sagayama, "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes," Information Processing Society of Japan, Tech. Rep., 2006.

[114] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[115] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[116] P. Sprechmann, A. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *13th International Society for Music Information Retrieval Conference*, Porto, Portugal, Oct. 2012.

[117] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[118] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Mathematical Programming Computation*, vol. 5, no. 2, pp. 201–226, Jun. 2013.

[119] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *27th International Conference on Machine Learning*, Haifa, Israel, Jun. 2010.

[120] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers of Electrical Electronic Engineering China*, vol. 6, no. 2, pp. 192–200, Jun. 2011.

[121] I.-Y. Jeong and K. Lee, "Vocal separation using extended robust principal component analysis with Schatten P/L_p-norm and scale compression," in *International Workshop on Machine Learning for Signal Processing*, Reims, France, Nov. 2014.

[122] F. Nie, H. Wang, and H. Huang, "Joint Schatten p-norm and l_p-norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, Mar. 2015.

[123] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *14th International Society for Music Information Retrieval conference*, Curitiba, PR, Brazil, Nov. 2013.

[124] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *26th Annual International Conference on Machine Learning*, Montreal, QC, Canada, Jun. 2009.

[125] T.-S. T. Chan and Y.-H. Yang, "Complex and quaternionic principal component pursuit and its application to audio separation," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 287–291, Feb. 2016.

[126] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach," in *International Symposium on Computer Music Multidisciplinary Research*, Montpellier, France, May 2003.

[127] R. B. Dannenberg and M. Goto, "Music structure analysis from acoustic signals," in *Handbook of Signal Processing in Acoustics*. Springer New York, 2008, pp. 305–331.

[128] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, Aug. 2010.

[129] Z. Rafii and B. Pardo, "A simple music/voice separation system based on the extraction of the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[130] ——, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73–84, Jan. 2013.

[131] Z. Rafii, A. Liutkus, and B. Pardo, "REPET for background/foreground separation in audio," in *Blind Source Separation*. Springer Berlin Heidelberg, 2014, pp. 395–411.

[132] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, Aug. 2001.

[133] P. Seetharaman, F. Pishdadian, and B. Pardo, "Music/voice separation using the 2d Fourier transform," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2017.

[134] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[135] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *13th International Society for Music Information Retrieval Conference*, Porto, Portugal, Oct. 2012.

[136] J. Foote, "Visualizing music and audio using self-similarity," in *7th ACM International Conference on Multimedia*, Orlando, FL, USA, Oct. 1999.

[137] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013.

[138] Z. Rafii, A. Liutkus, and B. Pardo, "A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[139] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *23rd IET Irish Signals and Systems Conference*, Maynooth, Ireland, Jun. 2012.

[140] A. Liutkus, Z. Rafii, B. Pardo, D. FitzGerald, and L. Daudet, "Kernel spectrogram models for source separation," in *4th Joint Workshop on Hands-free Speech Communication Microphone Arrays*, Villers-les-Nancy, France, May 2014.

[141] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[142] A. Liutkus, D. FitzGerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[143] T. Prätzlich, R. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Aug. 2015.

[144] D. F. Yela, S. Ewert, D. FitzGerald, and M. Sandler, "Interference reduction in music recordings combining kernel additive modelling and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017.

[145] M. Moussallam, G. Richard, and L. Daudet, "Audio source separation informed by redundancy with greedy multiscale decompositions," in *20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[146] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[147] H. Deif, D. FitzGerald, W. Wang, and L. Gan, "Separation of vocals from monaural music recordings using diagonal median filters and practical time-frequency parameters," in *IEEE International Symposium on Signal Processing and Information Technology*, Abu Dhabi, United Arab Emirates, Dec. 2015.

[148] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62–73, Jan. 2010.

[149] J.-Y. Lee and H.-G. Kim, "Music and voice separation using log-spectral amplitude estimator based on kernel spectrogram models backfitting," *Journal of the Acoustical Society of Korea*, vol. 34, no. 3, pp. 227–233, 2015.

[150] J.-Y. Lee, H.-S. Cho, and H.-G. Kim, "Vocal separation from monaural music using adaptive auditory filtering based on kernel back-fitting," in *Interspeech*, Dresden, Germany, Sep. 2015.

[151] H.-S. Cho, J.-Y. Lee, and H.-G. Kim, "Singing voice separation from monaural music based on kernel back-fitting using beta-order spectral amplitude estimation," in *16th International Society for Music Information Retrieval Conference*, Málaga, Spain, Oct. 2015.

[152] H.-G. Kim and J. Y. Kim, "Music/voice separation based on kernel back-fitting using weighted β-order MMSE estimation," *ETRI Journal*, vol. 38, no. 3, pp. 510–517, Jun. 2016.

[153] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[154] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *International Symposium on Frontiers of Research on Speech and Music*, 2007.

[155] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," MERL, Tech. Rep., 2006.

[156] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2005.

[157] J. Han and C.-W. Chen, "Improving melody extraction using probabilistic latent component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[158] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993.

[159] E. Gómez, F. J. C. nadas Quesada, J. Salamon, J. Bonada, P. V. Candea, and P. C. nas Molero, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing," in *13th International Society for Music Information Retrieval Conference*, Porto, Portugal, Aug. 2012.

[160] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *16th European Signal Processing Conference*, Lausanne, Switzerland, Aug. 2008.

[161] H. Papadopoulos and D. P. Ellis, "Music-content-adaptive robust principal component analysis for a semantically consistent separation of foreground and background in music audio signals," in *17th International Conference on Digital Audio Effects*, Erlangen, Germany, Sep. 2014.

[162] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the iKala dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[163] I.-Y. Jeong and K. Lee, "Singing voice separation using RPCA with weighted l_1-norm," in *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, Feb. 2017.

[164] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, Sep. 2008.

[165] Y. Wang and Z. Ou, "Combining HMM-based melody extraction and NMF-based soft masking for separating voice and accompaniment from monaural audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[166] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Oct. 2006.

[167] C.-L. Hsu, L.-Y. Chen, J.-S. R. Jang, and H.-J. Li, "Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement," in *10th International Society for Music Information Retrieval Conference*, Kyoto, Japan, Oct. 2009.

[168] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1884–1893, Sep. 2014.

[169] Z. Duan and B. Pardo, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[170] S. Venkataramani, N. Nayak, P. Rao, and R. Velmurugan, "Vocal separation using singer-vowel priors obtained from polyphonic audio," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[171] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2145–2154, Nov. 2010.

[172] V. Rao, C. Gupta, and P. Rao, "Context-aware features for singing voice detection in polyphonic music," in *International Workshop on Adaptive Multimedia Retrieval*, Barcelona, Spain, Jul. 2011.

[173] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, Oct. 2011.

[174] L. Zhang, Z. Chen, M. Zheng, and X. He, "Nonnegative matrix and tensor factorizations: An algorithmic perspective," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 54–65, May 2014.

[175] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[176] Y. Ikemiya, K. Itoyama, and K. Yoshii, "Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2084–2095, Nov. 2016.

[177] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[178] A. Dobashi, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A music performance assistance system based on vocal, harmonic, and percussive

source separation and content visualization for music audio signals," in *12th Sound and Music Computing Conference*, Maynooth, Ireland, Jul. 2015.

[179] Y. Hu and G. Liu, "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 643–653, Apr. 2015.

[180] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[181] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–347, Dec. 2001.

[182] Y. Li, J. Woodruff, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–1371, Sep. 2009.

[183] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, Sep. 2004.

[184] Y. Hu and G. Liu, "Monaural singing voice separation by non-negative matrix partial co-factorization with temporal continuity and sparsity criteria," in *12th International Conference on Intelligent Computing*, Lanzhou, China, Aug. 2016.

[185] X. Zhang, W. Li, and B. Zhu, "Latent time-frequency component analysis: A novel pitch-based approach for singing voice separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[186] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[187] B. Zhu, W. Li, and L. Li, "Towards solving the bottleneck of pitch-based singing voice separation," in *23rd ACM International Conference on Multimedia*, Brisbane, QLD, Australia, Oct. 2015.

[188] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, Apr. 2008.

[189] ——, "An iterative approach to monaural musical mixture de-soloing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009.

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

[191] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[192] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

[193] L. Benaroya, L. Mcdonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, Apr. 2003.

[194] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in Neural Information Processing Systems 18*. MIT Press, 2005, pp. 283–290.

[195] 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, Jan. 2006.

[196] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected F0 track," in *10th International Conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, Mar. 2012.

[197] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 2654–2658.

[198] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[199] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.

[200] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox," in *7th Sound and Music Computing Conference*, Barcelona, Spain, Jul. 2010.

[201] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[202] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *13th International Society for Music Information Retrieval Conference*, Porto, Portugal, Oct. 2012.

[203] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2385–2397, Nov. 2011.

[204] R. Zhao, S. Lee, D.-Y. Huang, and M. Dong, "Soft constrained leading voice separation with music score guidance," in *9th International Symposium on Chinese Spoken Language*, Singapore, Singapore, Sep. 2014.

[205] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *17th European Signal Processing Conference*, Glasgow, UK, Aug. 2009.

[206] J. Janer and R. Marxer, "Separation of unvoiced fricatives in singing voice mixtures with semi-supervised NMF," in *16th International Conference on Digital Audio Effects*, Maynooth, Ireland, Sep. 2013.

[207] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[208] R. Marxer and J. Janer, "Modelling and separation of singing voice breathiness in polyphonic mixtures," in *16th International Conference on Digital Audio Effects*, Maynooth, Ireland, Sep. 2013.

[209] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[210] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation*, St. Malo, France, Sep. 2010.

[211] ——, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[212] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014.

[213] R. Hennequin and F. Rigaud, "Long-term reverberation modeling for under-determined audio source separation with application to vocal melody extraction," in *17th International Society for Music Information Retrieval Conference*, New York City, NY, USA, Aug. 2016.

[214] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, Mar. 2010.

[215] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *9th International Conference on Music Information Retrieval*, Philadelphia, PA, USA, Sep. 2008.

[216] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *13th International Conference on Digital Audio Effects*, Graz, Austria, Sep. 2010.

[217] Y.-H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *20th ACM International Conference on Multimedia*, Nara, Japan, Oct. 2012.

[218] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1197–1200, Jun. 2014.

[219] E. Ochiai, T. Fujisawa, and M. Ikehara, "Vocal separation by constrained non-negative matrix factorization," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Hong Kong, China, Dec. 2015.

[220] T. Watanabe, T. Fujisawa, and M. Ikehara, "Vocal separation using improved robust principal component analysis and post-processing," in *IEEE 59th International Midwest Symposium on Circuits and Systems*, Abu Dhabi, United Arab Emirates, Oct. 2016.

[221] H. Raguet, J. Fadili, , and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, Jul. 2013.

[222] A. Hayashi, H. Kameoka, T. Matsubayashi, and H. Sawada, "Non-negative periodic component analysis for music source separation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Jeju, South Korea, Dec. 2016.

[223] D. FitzGerald, M. Cranitch, and E. Coyle, "Using tensor factorisation models to separate drums from polyphonic music," in *12th International Conference on Digital Audio Effects*, Como, Italy, Sep. 2009.

[224] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 228–237, Jan. 2014.

[225] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, Mar. 2010.

[226] H. Tachibana, N. Ono, and S. Sagayama, "A real-time audio-to-audio karaoke generation system for monaural recordings based on singing voice suppression and key conversion techniques," *Journal of Information Processing*, vol. 24, no. 3, pp. 470–482, May 2016.

[227] N. Ono, K. Miyamoto, H. Kameoka, J. L. Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Advances in Music Information Retrieval*. Springer Berlin Heidelberg, 2010, pp. 213–236.

[228] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[229] H. Deif, W. Wang, L. Gan, and S. Alhashmi, "A local discontinuity based approach for monaural singing voice separation from accompanying music with multi-stage non-negative matrix factorization," in *IEEE Global Conference on Signal and Information Processing*, Orlando, FL, USA, Dec. 2015.

[230] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2096–2107, Oct. 2013.

[231] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.

[232] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[233] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1584–1599, Aug. 2011.

[234] C.-L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1482–1491, Jul. 2012.

[235] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[236] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*. MIT Press Cambridge, 1986, pp. 318–362.

[237] N. J. Bryan and G. J. Mysore, "Interactive user-feedback for sound source separation," in *International Conference on Intelligent User-Interfaces, Workshop on Interactive Machine Learning*, Santa Monica, CA, USA, Mar. 2013.

[238] ——, "An efficient posterior regularized latent variable model for interactive sound source separation," in *30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[239] ——, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, BC, Canada, May 2013.

[240] K. Ganchev, J. ao Graça, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *Journal of Machine Learning Research*, vol. 11, pp. 2001–2049, Mar. 2010.

[241] A. Ozerov, N. Duong, and L. Chevallier, "Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation," Technicolor, Tech. Rep., 2013.

[242] X. Jaureguiberry, G. Richard, P. Leveau, R. Hennequin, and E. Vincent, "Introducing a simple fusion framework for audio source separation," in *IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, Sep. 2013.

[243] X. Jaureguiberry, E. Vincent, and G. Richard, "Variational Bayesian model averaging for audio source separation," in *IEEE Workshop on Statistical Signal Processing Workshop*, Gold Coast, VIC, Australia, Jun. 2014.

[244] ——, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1266–1279, Jul. 2016.

[245] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, Nov. 1999.

[246] M. McVicar, R. Santos-Rodriguez, and T. D. Bie, "Learning to separate vocals from polyphonic mixtures via ensemble methods and structured output prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016.

[247] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," in *IEEE International Conference on Systems, Man and Cybernetics*, Los Angeles, CA, USA, Nov. 1990.

[248] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[249] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher, "Block-coordinate Frank-Wolfe optimization for structural SVMs," in *30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[250] E. Manilow, P. Seetharaman, F. Pishdadian, and B. Pardo, "Predicting algorithm efficacy for adaptive, multi-cue source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2017.

[251] G. Wolf, S. Mallat, and S. Shamma, "Audio source separation with time-frequency velocities," in *IEEE International Workshop on Machine Learning for Signal Processing*, Reims, France, Sep. 2014.

[252] ——, "Rigid motion model for audio source separation," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1822–1831, Apr. 2016.

[253] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[254] C. P. Bernard, "Discrete wavelet analysis for fast optic flow computation," *Applied and Computational Harmonic Analysis*, vol. 11, no. 1, pp. 32–63, Jul. 2001.

[255] F. Yen, Y.-J. Luo, and T.-S. Chi, "Singing voice separation using spectro-temporal modulation features," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[256] F. Yen, M.-C. Huang, and T.-S. Chi, "A two-stage singing voice separation algorithm using spectro-temporal modulation features," in *Interspeech*, Dresden, Germany, Sep. 2015.

[257] T. Chi, P. Rub, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, Aug. 2005.

[258] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, Nov. 1999.

[259] T.-S. T. Chan and Y.-H. Yang, "Informed group-sparse representation for singing voice separation," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 156–160, Feb. 2017.

[260] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, Dec. 2006.

[261] S. Ma, "Alternating proximal gradient method for convex minimization," *Journal of Scientific Computing*, vol. 68, no. 2, p. 546572, Aug. 2016.

[262] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, Jan. 2007.

[263] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[264] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," 1993.
[265] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *132nd AES Convention*, Budapest, Hungary, Apr. 2012.
[266] M. Vinyes, "MTG MASS database," 2008, http://www.mtg.upf.edu/static/mass/resources.
[267] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *8th International Conference on Independent Component Analysis and Signal Separation*, Paraty, Brazil, Mar. 2009.
[268] S. Araki, A. Ozerov, B. V. Gowreesunker, H. Sawada, F. J. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): - audio source separation -," in *9th International Conference on Latent Variable Analysis and Signal Separation*, St. Malo, France, Sep. 2010.
[269] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation -," in *10th International Conference on Latent Variable Analysis and Signal Separation*, 2012.
[270] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, Aug. 2012.
[271] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic,, Aug. 2015.
[272] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, Feb. 2017.
[273] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 59, no. 7, pp. 3155–3167, Feb. 2011.
[274] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, , and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.
[275] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18, a dataset for audio source separation," Dec. 2017. [Online]. Available: https://sigsep.github.io/musdb
[276] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2005.
[277] W.-H. Tsai, D. Rogers, and H.-M. Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music Journal*, vol. 28, no. 3, pp. 68–78, 2004.
[278] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[279] E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, pp. 162–185.
[280] Z. Rafii, D. L. Sun, F. G. Germain, and G. J. Mysore, "Combining modeling of singing voice and background music for automatic separation of musical mixtures," in *14th International Society for Music Information Retrieval Conference*, Curitiba, PR, Brazil, Nov. 2013.
[281] N. Boulanger-Lewandowski, G. J. Mysore, and M. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014.
[282] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation*, St. Malo, France, Sep. 2010.
[283] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," *Proc. Interspeech 2017*, pp. 2013–2017, 2017.
[284] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3-4, pp. 197–387, Jun. 2014.
[285] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[286] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[287] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951.
[288] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.
[289] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *26th International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, Dec. 2013.
[290] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *International Conference on Learning Representations*, Banff, AB, Canada, Apr. 2014.
[291] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, 2015.
[292] ——, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014.
[293] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015.
[294] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017.
[295] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.
[296] J. Schlüter, "Learning to pinpoint singing voice from weakly labeled examples," in *17th International Society for Music Information Retrieval Conference*, New York City, NY, USA, Aug. 2016.
[297] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, Feb. 2017.
[298] S. I. Mimilakis, E. Cano, J. Abeßer, and G. Schuller, "New sonorities for jazz recordings: Separation and mixing using deep neural networks," in *2nd AES Workshop on Intelligent Music Production*, London, UK, Sep. 2016.
[299] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *IEEE International Workshop on Machine Learning for Signal Processing*, Tokyo, Japan, Sep. 2017.
[300] S. I. Mimilakis, K. Drossos, J. ao F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2018.
[301] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *18th International Society for Music Information Retrieval Conferenceng*, Suzhou, China, Oct. 2017.
[302] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2017.
[303] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016.
[304] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multispeaker separation using deep clustering," in *Interspeech*, 2016.
[305] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New-Orleans, LA, USA, Mar. 2017.

[306] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in *12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.

[307] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Dec. 2010.

[308] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Single channel audio source separation using deep neural network ensembles," in *140th AES Convention*, Paris, France, May 2016.

[309] ——, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Interspeech*, San Francisco, CA, USA, Sep. 2016.

[310] ——, "Discriminative enhancement for single channel audio source separation using deep neural networks," in *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, Feb. 2017.

[311] ——, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1773–1783, Sep. 2017.

[312] S. Nie, W. Xue, S. Liang, X. Zhang, W. Liu, L. Qiao, and J. Li, "Joint optimization of recurrent networks exploiting source auto-regression for source separation," in *Interspeech*, Dresden, Germany, Sep. 2015.

[313] J. Sebastian and H. A. Murthy, "Group delay based music source separation using deep recurrent neural networks," in *International Conference on Signal Processing and Communications*, Bangalore, India, Jun. 2016.

[314] B. Yegnanarayana, H. A. Murthy, and V. R. Ramachandran, "Processing of noisy speech using modified group delay functions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, Apr. 1991.

[315] Z.-C. Fan, J.-S. R. Jang, and C.-L. Lu, "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking," in *IEEE International Conference on Multimedia Big Data*, Taipei, Taiwan, Apr. 2016.

[316] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2003.

[317] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *AES 22nd International Conference*, Espoo, Finland, Jun. 2002.

[318] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *7th International Conference on Digital Audio Effects*, Naples, Italy, Oct. 2004.

[319] M. Vinyes, J. Bonada, and A. Loscos, "Demixing commercial music productions via human-assisted time-frequency masking," in *120th AES Convention*, Paris, France, May 2006.

[320] M. Cobos and J. J. López, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," *Digital Signal Processing*, vol. 18, no. 6, pp. 960–976, Nov. 2008.

[321] Özgür Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[322] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[323] S. Sofianos, A. Ariyaeeinia, and R. Polfreman, "Towards effective singing voice extraction from stereophonic recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, Mar. 2010.

[324] ——, "Singing voice separation based on non-vocal independent component subtraction," in *13th International Conference on Digital Audio Effects*, Graz, Austria, Sep. 2010.

[325] S. Sofianos, A. Ariyaeeinia, R. Polfreman, and R. Sotudeh, "H-semantics: A hybrid approach to singing voice separation," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 831–841, Oct. 2012.

[326] M. Kim, S. Beack, K. Choi, and K. Kang, "Gaussian mixture model for singing voice separation from stereophonic music," in *AES 43rd Conference*, Pohang, South Korea, Sep. 2011.

[327] M. Cobos and J. J. López, "Singing voice separation combining panning information and pitch tracking," in *AES 124th Convention*, Amsterdam, Netherlands, May 2008.

[328] D. FitzGerald, "Stereo vocal extraction using ADRess and nearest neighbours median filtering," in *16th International Conference on Digital Audio Effects*, Maynooth, Ireland, Jan. 2013.

[329] D. FitzGerald and R. Jaiswal, "Improved stereo instrumental track recovery using median nearest-neighbour inpainting," in *24th IET Irish Signals and Systems Conference*, Letterkenny, Ireland, Jun. 2013.

[330] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[331] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures with application to blind audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009.

[332] ——, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[333] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[334] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *9th International Conference on Latent Variable Analysis and Signal Separation*, St. Malo, France, Sep. 2010.

[335] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," in *7th International Symposium on Computer Music Modeling and Retrieval*, Málaga, Spain, Jun. 2010.

[336] A. Ozerov, N. Duong, and L. Chevallier, "On monotonicity of multiplicative update rules for weighted nonnegative tensor factorization," in *International Symposium on Nonlinear Theory and its Applications*, Luzern, Switzerland, Sep. 2014.

[337] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel NMF," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2011.

[338] ——, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[339] ——, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, May 2013.

[340] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. M. Cordovilla, S. Dalmia, I. Illina, and A. Liutkus, "Robust ASR using neural network based speech enhancement and feature simulation," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Scottsdale, AZ, USA, Dec. 2015.

[341] 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, Sep. 2016.

[342] ——, "Multichannel audio source separation with deep neural networks," Inria, Tech. Rep., 2015.

[343] ——, "Multichannel music separation with deep neural networks," in *24th European Signal Processing Conference*, Budapest, Hungary, Aug. 2016.

[344] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[345] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2011.

[346] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer-Verlag Berlin Heidelberg, 2013.

[347] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, USA, May 2001.

[348] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[349] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Scottsdale, AZ, USA, Dec. 2015.
[350] I. Recommendation, "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.
[351] E. Vincent, M. Jafari, and M. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *ICA Research Network International Workshop*, Southampton, UK, Sep. 2006.
[352] E. Cano, C. Dittmar, and G. Schuller, "Influence of phase, magnitude and location of harmonic components in the perceived quality of extracted solo signals," in *AES 42nd Conference on Semantic Audio*, Ilmenau, Germany, Jul. 2011.
[353] C. Févotte, R. Gribonval, and E. Vinvent, "BSS_EVAL toolbox user guide - revision 2.0," IRISA, Tech. Rep., 2005.
[354] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[355] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *7th International Conference on Latent Variable Analysis and Signal Separation*, London, UK, Sep. 2007.
[356] B. Fox and B. Pardo, "Towards a model of perceived quality of blind audio source separation," in *IEEE International Conference on Multimedia and Expo*, Beijing, China, Jul. 2007.
[357] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," *Journal of the Acoustical Society of America*, vol. 4, no. 1, 2008.
[358] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Multi-criteria subjective and objective evaluation of audio source separation," in *38th International AES Conference*, Pitea, Sweden, Jun. 2010.
[359] ——, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
[360] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *10th International Conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, Mar. 2012.
[361] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016.
[362] U. Gupta, E. Moore, and A. Lerch, "On the perceptual relevance of objective source separation measures for singing voice separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2005.
[363] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016.
[364] G. Roma, E. M. Grais, A. J. Simpson, I. Sobieraj, and M. D. Plumbley, "Untwist: A new toolbox for audio source separation," in *17th International Society on Music Information Retrieval Conference*, New York City, NY, USA, Aug. 2016.

Zafar Rafii received a PhD in Electrical Engineering and Computer Science from Northwestern University in 2014, and an MS in Electrical Engineering from both Ecole Nationale Supérieure de l'Electronique et de ses Applications in France and Illinois Institute of Technology in the US in 2006. He is currently a senior research engineer at Gracenote in the US. He also worked as a research engineer at Audionamix in France. His research interests are centered on audio analysis, somewhere between signal processing, machine learning, and cognitive science, with a predilection for source separation and audio identification in music.

Antoine Liutkus received the State Engineering degree from Télécom ParisTech, France, in 2005, and the M.Sc. degree in acoustics, computer science and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005. He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and obtained his PhD in electrical engineering at Télécom ParisTech in 2012. He is currently researcher at Inria, France. His research interests include audio source separation and machine learning.

Fabian-Robert Stöter received the diploma degree in electrical engineering in 2012 from the Leibniz Universität Hannover and worked towards his Ph.D. degree in audio signal processing in the research group of B. Edler at the International Audio Laboratories Erlangen, Germany. He is currently researcher at Inria, France. His research interests include supervised and unsupervised methods for audio source separation and signal analysis of highly overlapped sources.

Stylianos Ioannis Mimilakis received a Master of Science degree in Sound & Music Computing from Pompeu Fabra University and a Bachelor of Engineering in Sound & Music Instruments Technologies from Higher Technological Education Institute of Ionian Islands. Currently he is pursuing his Ph.D. in signal processing for music source separation, under the MacSeNet project at Fraunhofer IDMT. His research interests include, inverse problems in audio signal processing and synthesis, singing voice separation and deep learning.

Derry FitzGerald (PhD, M.A. B.Eng.) is a Research Felow in the Cork School of Music at Cork Institute of Technology. He was a Stokes Lecturer in Sound Source Separation algorithms at the Audio Research Group in DIT from 2008-2013. Previous to this he worked as a post-doctoral researcher in the Dept. of Electronic Engineering at Cork Institute of Technology, having previously completed a Ph.D. and an M.A. at Dublin Institute of Technology. He has also worked as a Chemical Engineer in the pharmaceutical industry for some years. In the field of music and audio, he has also worked as a sound engineer and has written scores for theatre. He has utilised his sound source separation technologies to create the first ever officially released stereo mixes of several songs for the Beach Boys, including Good Vibrations and I get around. His research interests are in the areas of sound source separation and, tensor factorizations.

Bryan Pardo, head of the Northwestern University Interactive Audio Lab, is an associate professor in the Northwestern University Department of Electrical Engineering and Computer Science. Prof. Pardo received a M. Mus. in Jazz Studies in 2001 and a Ph.D. in Computer Science in 2005, both from the University of Michigan. He has authored over 80 peer-reviewed publications. He has developed speech analysis software for the Speech and Hearing department of the Ohio State University, statistical software for SPSS and worked as a machine learning researcher for General Dynamics. While finishing his doctorate, he taught in the Music Department of Madonna University. When he's not programming, writing or teaching, he performs throughout the United States on saxophone and clarinet at venues such as Albion College, the Chicago Cultural Center, the Detroit Concert of Colors, Bloomington Indiana's Lotus Festival and Tucson's Rialto Theatre.

Gaussian Processes for Underdetermined Source Separation

Antoine Liutkus, Roland Badeau, *Senior Member, IEEE,* Gaël Richard, *Senior Member, IEEE*

***Abstract*—Gaussian process (GP) models are very popular for machine learning and regression and they are widely used to account for spatial or temporal relationships between multivariate random variables. In this paper, we propose a general formulation of underdetermined source separation as a problem involving GP regression. The advantage of the proposed unified view is firstly to describe the different underdetermined source separation problems as particular cases of a more general framework. Secondly, it provides a flexible means to include a variety of prior information concerning the sources such as smoothness, local stationarity or periodicity through the use of adequate covariance functions. Thirdly, given the model, it provides an optimal solution in the minimum mean squared error (MMSE) sense to the source separation problem. In order to make the GP models tractable for very large signals, we introduce *framing* as a GP approximation and we show that computations for *regularly sampled* and *locally stationary* GPs can be done very efficiently in the frequency domain. These findings establish a deep connection between GP and Nonnegative Tensor Factorizations with the Itakura-Saito distance and lead to effective methods to learn GP hyperparameters for very large and regularly sampled signals.**

***Index Terms*—Gaussian Processes, NMF, NTF, Source Separation, Probability Theory, Regression, Kriging, Cokriging**

I. Introduction

Gaussian processes [28], [35], [36], [44] are commonly used to model functions whose mean and covariances are known. Given some learning points, they enable us to estimate the values taken by the function at any other points of interest. Their main advantages are to provide a simple and effective probabilistic framework for regression and classification as well as an effective means to optimize a model's parameters through maximization of the *marginal likelihood* of the observations. For these reasons, they are widely used in many areas to model dependencies between multivariate random variables and their use can be traced back at least to works by Wiener in 1941 [43]. They have also been known in geostatistics under the name of *kriging* for almost 40 years [29]. A great surge of interest for Gaussian Process (GP) models occurred when they were expressed as a general purpose framework for regression as well as for classification (see [35] for a review). Their relation to other methods commonly used in machine learning such as multi-layer perceptrons, spline interpolation or support vector machines are now well understood.

Authors are with Institut Telecom, Telecom ParisTech, CNRS LTCI, France.
This work is partly funded by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

Source separation is another very intense field of research (see [10] for a review) where the objective is to recover several unknown signals called *sources* that were mixed together in observable *mixtures*. Source separation problems arise in many fields such as sound processing, telecommunications and image processing. They differ mainly in the relative number of mixtures per source signal and in the nature of the mixing process. The latter is generally modeled as *convolutive*, i.e. as a linear filtering of the sources into the mixtures. When the mixing filters reduce to a single amplification gain, the mixing is called *instantaneous*. When there are more mixtures than sources, the problem is called *overdetermined* and algorithms may rely on beamforming techniques to perform source separation. When there are fewer mixtures than sources, the problem is said to be *underdetermined* and is notably known to be very difficult. Indeed, in this case there are less observable signals than necessary to solve the underlying mixing equations. Many models were hence studied to address this problem and they all either restrict the set of possible source signals or assign prior probabilities to them in a Bayesian setting. Among the most popular approaches, we can mention Independent Component Analysis [6] that focuses both on probabilistic independence between the source signals and on high order statistics. We can also cite Non-negative Matrix Factorization (NMF) source separation that models the sources as locally stationary with constant normalized power spectra and time-varying energy [16], [27].

In this study, we revisit underdetermined source separation (USS) as a problem involving GP regression. To our knowledge, no unified treatment of the different underdetermined linear source separation problems in terms of classical GP is available to date and we thus propose here an attempt at providing such a formulation whose advantages are numerous. Firstly, it provides a unified framework for handling the different USS problems as particular cases, including convolutive or instantaneous mixing as well as single or multiple mixtures. Secondly, when prior information such as smoothness, local stationarity or periodicity is available, it can be taken into account through appropriate *covariance functions*, thus providing a significant expressive power to the model. Thirdly, it yields an optimal way in the minimum mean squared error (MMSE) sense to proceed to the separation of the sources given the model.

In spite of all their interesting features, GP models come at a high $\mathcal{O}\left(n^3\right)$ computational cost where n is the number of training points. For many applications such as audio signal processing where $n \approx 10^7$ is common, this cost is prohibitive. Hence, the GP framework has to come along with effective methods to simplify the computations in order to be of

practical use. Over the years, many approximation methods have been proposed [31], [34], [37], [38], [40] to address this issue and we show that the common practice of *framing* in audio signal processing can precisely be understood in terms of GP modeling as a particular choice for GP approximation. In particular, we give its connections with recently published Partially Independent Conditional (PIC) approximation [37] and Compact Support (CS) covariance functions [31], [40]. For the special case of locally stationary and regularly sampled signals, we furthermore show that computations can be performed extremely efficiently in the frequency domain and we establish a novel connection between GP models and the emerging techniques of Nonnegative Tensor Factorizations (NTF) [9] using the Itakura-Saito divergence.

The article is organized as follows. First, we present GP and particularly Gaussian Process Regression (GPR) in section II. Then, we set out the various linear underdetermined source separation problems in terms of GPR in section III. In order to make the GP models tractable for very large signals, we introduce *framing* as a GP approximation and we show that computations for *regularly sampled* and *locally stationary* GPs can be done very efficiently in the frequency domain in section IV. Finally, we illustrate the performance of the methods on synthetic and real data in section V and draw some conclusions in section VI.

II. Gaussian processes

A. Introduction

A Gaussian process [28], [35], [36], [44] is a possibly infinite set of scalar random variables $\{f(x)\}_{x\in\mathcal{X}}$ indexed by an *input space* $\mathcal{X}$, typically $\mathcal{X}=\mathbb{R}^D$, and taking values on $\mathbb{R}$, such that for any *finite* set of inputs $X=\{x_1\cdots x_n\}\in\mathcal{X}^n$, $\mathbf{f}\triangleq[f(x_1)\cdots f(x_n)]^\top$ is distributed with respect to a multivariate Gaussian distribution[1]. A GP is thus completely determined by a mean function $m(x)=\mathbb{E}[f(x)]$ and a covariance function $k(x,x')=\mathbb{E}[(f(x)-m(x))(f(x')-m(x'))]$.

More fundamentally, a GP may be understood as a process whose mean $m(x)$ and covariance $k(x,x')$ between any two inputs are known. Given only this prior information, assigning a multivariate Gaussian distribution to $\mathbf{f}$ given any finite set X of inputs from $\mathcal{X}$ is a sensible choice, since it is the probability distribution that maximizes entropy when only the first two moments are known [23].

It has been shown that the class of valid covariance functions coincides with the class of positive definite functions [1]. Let X be a finite set of elements from $\mathcal{X}$ that is possibly randomly drawn as in [19], the covariance matrix $K_{f,XX}$ is defined as $[K_{f,XX}]_{i,j}=k(x_i,x_j)$ and the probability of $\mathbf{f}$ given X is then given by[2]:

$$p(\mathbf{f}\mid X)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|K_{f,XX}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^\top K_{f,XX}^{-1}(\mathbf{f}-\mathbf{m})\right) \tag{1}$$

where $\mathbf{m}\triangleq[m(x_1)\cdots m(x_n)]^\top$. This is usually written:

$$f\sim\mathcal{GP}(m(x),k(x,x'))$$

Most studies in underdetermined source separation focus on the *single sensor* scenario $\mathcal{X}=\mathbb{R}$. Still, there is no difficulty involved in considering the general case $\mathcal{X}=\mathbb{R}^D$ and we will see examples of GPs defined on a multidimensional input space in sections III-C and IV. This framework thus easily allows modeling multivariate functions defined on arbitrary input spaces and many studies have used Gaussian processes for regression ($f(x)\in\mathbb{R}$) as well as for classification ($f(x)\in\mathbb{N}$). Their main advantages are to provide a probabilistic interpretation and a way to compute the variances of the estimates. From now on, we will focus on GPR, since our objective is to highlight the connections between GP and source separation, which is usually stated in terms of processes taking values in $\mathbb{R}$. For the sake of notational simplicity, we will assume *a priori* centered signals, i.e. $\forall x\in\mathcal{X}, m(x)=0$, as it is very common for audio signals. Still, there is no particular issue raised when considering arbitrary mean functions.

B. Gaussian processes regression

Suppose we observe $y(x)=f(x)+\epsilon(x)$, with $f(x)$ being the signal of interest and $\epsilon(x)$ being some additive signal — usually called *noise* — that is independent from $f(x)$, for a finite set X of input points from $\mathcal{X}$: $X=\{x_1\cdots x_n\}\in\mathcal{X}^n$. We want to estimate the values taken by f on a finite and possibly different set $X^*=\{x_1^*\cdots x_{n^*}^*\}\in\mathcal{X}^{n^*}$ of input points from $\mathcal{X}$. Let us furthermore assume that $f\sim\mathcal{GP}(0,k_f(x,x'))$ and $\epsilon\sim\mathcal{GP}(0,k_\epsilon(x,x'))$ where the covariance functions k_f and k_ϵ are known. As f and ϵ are supposed independent, we have:

$$f+\epsilon\sim\mathcal{GP}(0,k_f(x,x')+k_\epsilon(x,x')) \tag{2}$$

Let K_{f,XX^*} be the covariance matrix defined by $[K_{f,XX^*}]_{ij}=k_f(x_i,x_j^*)$. We define K_{f,X^*X}, K_{f,X^*X^*}, $K_{\epsilon,XX}$ in the same way. Let $\mathbf{f}\triangleq[f(x_1)\cdots f(x_n)]^\top$, $\mathbf{f}^*\triangleq[f(x_1^*)\cdots f(x_n^*)]^\top$ and similarly for $\mathbf{y}$. We have:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix}\sim\mathcal{N}\left(\mathbf{0},\begin{bmatrix}K_{f,XX}+K_{\epsilon,XX} & K_{f,XX^*}\\ K_{f,X^*X} & K_{f,X^*X^*}\end{bmatrix}\right)$$

Classical probability results then assert that the conditional distribution of $\mathbf{f}^*$ given $\mathbf{y}$ is (see [35]):

$$\mathbf{f}^*\mid\mathbf{y}\sim\mathcal{N}\left(\overline{\mathbf{f}^*},\mathrm{cov}\mathbf{f}^*\right) \tag{3}$$

with[3]:

$$\overline{\mathbf{f}^*}=K_{f,X^*X}\left[K_{f,XX}+K_{\epsilon,XX}\right]^{-1}\mathbf{y} \tag{4}$$

and

$$\mathrm{cov}\mathbf{f}^*=K_{f,X^*X^*}-K_{f,X^*X}\left[K_{f,XX}+K_{\epsilon,XX}\right]^{-1}K_{f,XX^*} \tag{5}$$

These expressions show that the maximum likelihood estimate $\hat{\mathbf{f}}^*$ of $\mathbf{f}^*\mid\mathbf{y}$ is found by setting $\hat{\mathbf{f}}^*=\overline{\mathbf{f}^*}$, which is also the Minimum Mean Squared Error (MMSE) estimate in the Gaussian case. This result will be fundamental when performing source separation using Gaussian processes. We can furthermore compute the covariance of the estimates.

[1] The symbol $\triangleq$ denotes a definition.

[2] Positive *semi*-definite covariance matrices are possible. In the case of singular $K_{f,XX}$, a characterization involving the characteristic function instead of (1) is required.

[3] In the case of singular covariance matrix $K_{f,XX}+K_{\epsilon,XX}$, numerical methods such as Moore-Penrose pseudo-inversion may be used.

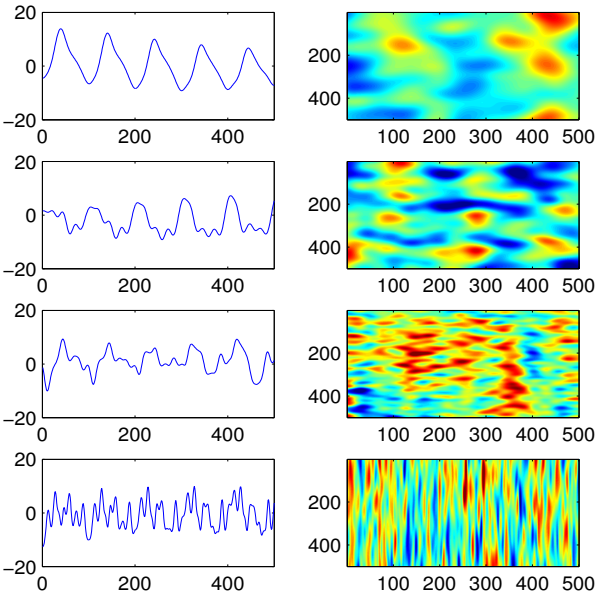


Figure 1. Typical realizations of GP with SE and periodic covariance functions with different values of the hyperparameters. On the left column, $D = 1$ and the processes are only parameterized by 3 scalars. On the right, $D = 2$ and the processes are parameterized by 6 scalars.

C. Covariance functions

Many studies (see [1] for a review) concern valid covariance functions. They belong to the general family of *kernels* and must be definite positive. Some properties of interest have been demonstrated:

- Sums and products of valid covariance functions are valid covariance functions.
- When it is *stationary*, i.e. when it can be expressed as a function of $\tau = x - x'$, then the covariance function can be parameterized by its Fourier transform.

Two examples of covariance functions for $\mathcal{X} = \mathbb{R}^D$ are:

- The Squared Exponential (SE) covariance function defined by: $k_{\text{SE}}(x, x'|\sigma, M) = \sigma^2 \exp\left(-\frac{(x-x')^\top M (x-x')}{2}\right)$ with $\sigma^2 > 0$ and M positive semidefinite. When $D = 1$, we have $k_{\text{SE}}(x, x'|\sigma, \lambda) = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$. λ is called a *characteristic length scale* in the sense that $|x - x'| \gg \lambda$ is required for two points x and x' of the process to be independent.
- The less common periodic covariance function of period T given by: $k_{\text{periodic}}(x, x'|T, \lambda) = \sigma^2 \exp\left(-\frac{2\sin^2\frac{\pi(x-x')}{T}}{\lambda^2}\right)$.

As can be seen, covariance functions are generally parameterized by a set of scalar values such as their Fourier transform when they are stationary, their characteristic length scales, their period, etc. These scalars are often called *hyperparameters* and are usually gathered in a *hyperparameter set* Θ. Typical realizations of GPs with SE and periodic covariance functions are given for $D = 1$ and $D = 2$ in Figure 1.

D. Optimization of the hyperparameters

In a Bayesian context, we may need to find the hyperparameters that maximize the marginal likelihood of the observations. In other words, we may need to find Θ^* such that $p(\mathbf{y} \mid X, \Theta^*)$ is maximum. Indeed, even if we may well guess the covariance functions that are adequate to the problem at hand, such as stationary covariance functions parameterized by their Fourier transform or SE covariance functions, it is likely that the hyperparameters that best explain the observations are not exactly known.

To this purpose, we can compute the closed-form expression of the marginal log-likelihood of the observations, $\log p(\mathbf{y} \mid X, \Theta)$ as (see [35]):

$$\log p(\mathbf{y} \mid X, \Theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\mathbf{y}^\top \left[K_{f,XX} + K_{\epsilon,XX}\right]^{-1}\mathbf{y} - \frac{1}{2}\log\left|K_{f,XX} + K_{\epsilon,XX}\right| \tag{6}$$

where each covariance matrix depends on Θ. Using the opposite of (6) as a cost function, we can proceed to the optimization of the hyperparameters using classical optimization algorithms, in a principled probabilistic framework. Note that depending on the covariance function and the hyperparameter considered, the corresponding optimization problem may or may not be convex.

III. Gaussian processes for source separation

A. Single mixture with instantaneous mixing

The presentation of GPR given in section II-B is actually slightly more general than what is usual in the literature. Indeed, it is often assumed that the covariance function k_ϵ of the additive signal ϵ is given by $k_\epsilon(x, x') = \sigma^2 \delta_{xx'}$ where $\delta_{xx'} = 1$ if and only if $x = x'$ and zero otherwise. This assumption corresponds to additive independent and identically distributed (i.i.d.) white Gaussian noise of variance σ^2.

In our presentation, the additive signal $\epsilon(x)$ is a GP itself and is potentially very complex. In any case, its covariance function is given by k_ϵ and the only assumption made is its independence with the signal of interest $f(x)$. A particular example of a model where k_ϵ non trivially depends on x and x' was for example studied in [20].

The results obtained can very well be generalized to the situation where y is the sum of M independent latent Gaussian processes:

$$\forall x \in \mathcal{X}, y(x) = \sum_{m=1}^{M} f_m(x)$$

with

$$f_m \sim \mathcal{GP}(0, k_m(x, x'))$$

In this case, if our objective is to extract the signal corresponding to the source m_0, we only need to replace k_f with k_{m_0} and k_ϵ with $\sum_{m \neq m_0} k_m$ in section II-B. Note that inversion of $K_{f,XX} + K_{\epsilon,XX}$ is needed only once for the extraction of all sources. Similarly, we can also jointly

optimize the hyperparameters of all covariance functions using exactly the same framework as in section II-D. We now consider the case of convolutive mixtures of independent GPs.

B. Single mixture with convolutive mixing

An important fact, which has already been noticed in the literature [2], [4], is that the convolution of a GP, as a linear combination of Gaussian random variables, remains a GP. Indeed, let us consider some GP $f_{0,m} \sim \mathcal{GP}\left(0, k_{0,m}\left(x, x'\right)\right)$ and let us define

$$f_m\left(x\right) = \int_{\mathcal{X}} a_m\left(x - z\right) f_{0,m}\left(z\right) dz \triangleq \left(a_m * f_{0,m}\right)\left(x\right)$$

where $a_m : \mathcal{X} \to \mathbb{R}$ is a stable *mixing filter* from $f_{0,m}$ to f_m. If the mean function of $f_{0,m}$ is identically 0, the mean function of f_m is easily seen to also be identically 0. The covariance function of f_m can be computed as $k_m\left(x, x'\right) = \mathbb{E}\left[f_m\left(x\right) f_m\left(x'\right)\right]$, that is:

$$k_m\left(x, x'\right) = \int_{\mathcal{X}} \int_{\mathcal{X}} a_m\left(x - z\right) a_m\left(x' - z'\right) k_{0,m}\left(z, z'\right) dz dz'$$

which is the convolution of $k_{0,m}$ by $a_m \times a_m \triangleq \left(x, x'\right) \in \mathcal{X}^2 \mapsto a_m\left(x\right) a_m\left(x'\right)$:

$$k_m\left(x, x'\right) = \left(\left(a_m \times a_m\right) * k_{0,m}\right)\left(x, x'\right) \tag{7}$$

Moreover, if several convolved GPs $\left\{f_m = \left(a_m * f_{0,m}\right)\right\}_{m=1\cdots M}$ are summed up in a mixture, it can readily be shown that the f_m are independent if the $f_{0,m}$ are independent. We thus get back to the instantaneous mixing model using modified covariance functions (7).

C. Multiple output GP

We have for now only considered GPs whose outputs lie in $\mathbb{R}$. A sizable body of literature focuses on possible extensions of this framework to cases where the processes of interest are multiple-valued, i.e. whose outputs lie in $\mathbb{R}^C$ for $C \in \mathbb{N}^*$. In geostatistics for example, important applications comprise the modeling of co-occurrences of minerals or pollutants in a spatial field. First attempts in this direction [24] include the so-called *linear model of coregionalization*, that considers each output as a linear combination of some latent processes. The name of *cokriging* has often been used for such systems in the field of geostatistics. If the latent processes are assumed to be GPs, the outputs are also GPs.

In the machine learning community, multiple-output GPs have been introduced [5] and popularized under the name of *dependent* GPs. Several extensions of such models have been proposed subsequently [2]–[4], [30] and we focus here on the model presented in [2] which is very close to the usual convolutive mixing model commonly used in multi-channel source separation, e.g. in [32].

Let $\left\{y_c\left(x\right)\right\}_{c=1\cdots C}$ be the C output signals called the *mixtures*. The *convolutive* GP model consists in assuming that each observable signal y_c is the sum of convolved versions of M latent GPs of interest $\left\{f_{0,m} \sim \mathcal{GP}\left(0, k_{0m}\left(x, x'\right)\right)\right\}_{m=1\cdots M}$ that we will call *sources*, plus one specific additional term $\epsilon_c \sim \mathcal{GP}\left(0, k_{\epsilon c}\left(x, x'\right)\right)$ that is often referred to as *additive noise*. We thus have:

$$y_c\left(x\right) = \sum_{m=1}^{M} \left(a_{cm} * f_{0,m}\right)\left(x\right) + \epsilon_c\left(x\right) \tag{8}$$

Instead of making a fundamental distinction between c and x, the GP framework allows us to consider that $\left\{y_c\left(x\right)\right\}_{(c,x)\in\{1\cdots C\}\times\mathcal{X}}$ is a single signal $\left\{y\left(x'\right)\right\}_{x'\in\{1\cdots C\}\times\mathcal{X}}$ indexed on an extended input space $\left\{1\cdots C\right\} \times \mathcal{X}$. If we assume that the different underlying sources $\left\{f_{0,m}\right\}_{m=1\cdots M}$ are independent, which is frequent in source separation and that the different $\left\{\epsilon_c\right\}_{c=1\cdots C}$ are also independent, we can express the covariance function $k\left(\left(c, x\right), \left(c', x'\right)\right)$ of y for two extended input points $\left(c, x\right)$ and $\left(c', x'\right)$ as:

$$k_{cc'}\left(x, x'\right) = \left(\sum_{m=1}^{M} k_{cc',m} + \delta_{cc'} k_{\epsilon c}\right)\left(x, x'\right) \tag{9}$$

$$\text{where } k_{cc',m}\left(x, x'\right) \triangleq \left(\left(a_{cm} \times a_{c'm}\right) * k_{0,m}\right)\left(x, x'\right) \tag{10}$$

For any given c, the different $\left\{f_{cm} \triangleq a_{cm} * f_{0,m}\right\}_{m=1\cdots M}$ are independent and are GPs with mean functions 0 and covariance functions $k_{cc,m}\left(x, x'\right)$. f_{cm} will be called the *contribution* of source m to mixture c. We can readily perform source separation on $\mathbf{y_c}$ to recover the different $\left\{\mathbf{f}_{cm}\right\}_{m=1\cdots M}$ using the standard formalism presented in section II-B. Let $\hat{\mathbf{f}}_{cm_0}$ be the estimate of $\mathbf{f}_{cm_0}$, we have:

$$\hat{\mathbf{f}}_{cm_0} = K_{cc,m_0}\left[\sum_{m=1}^{M} K_{cc,m} + K_{cc,\epsilon}\right]^{-1} \mathbf{y}_c \tag{11}$$

where $K_{cc,\epsilon}$ is the covariance matrix of the additive signal ϵ_c and where the covariance matrix $K_{cc,m}$ is defined as $\left[K_{cc,m}\right]_{x,x'} = k_{cc,m}\left(x, x'\right)$.

It is important to note here that even if the *sources* are the $\left\{f_{0,m}\right\}_{m=1\cdots M}$, many systems consider the signals of interest to actually be the different $\left\{f_{cm}\right\}_{c,m}$. For example, in the case of audio source separation, a stereophonic mixture can be composed of several monophonic sources such as voice, piano and drums. It is often considered sufficient to be able to separate the different instruments *within* the stereo mixtures and thus to obtain one stereo signal for each source, rather than trying to recover the original monophonic signals.

Still, for some m, given the estimates $\left\{\hat{\mathbf{f}}_{cm}\right\}_{c=1\cdots C}$ of all the different $\left\{\mathbf{f}_{cm}\right\}_{c=1\cdots C}$, we can for example estimate $\mathbf{f}_{0,m}$ using standard beamforming techniques.

D. Parameter optimization

Even in complex situations such as those presented in sections III-B or III-C, we can still use classical optimization methods to maximize the marginal log-likelihood (6) of the observations. Following [2], we will now give a simple way to include multiple output GPs in this framework.

Given a set X of n input points and the corresponding C column vectors $\left\{\mathbf{y}_c\right\}_{c=1\cdots C}$, we can build $\mathbf{y} \triangleq \left[\mathbf{y}_1^\top, \ldots, \mathbf{y}_C^\top\right]^\top$ as the Cn column vector containing all stacked outputs and use the expression (9) to build its covariance matrix K. We can ten proceed to parameters estimation through maximization of the

marginal log-likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations. Once more, depending on the covariance functions considered, this problem may or may not be convex.

E. Conclusion

In this section, we have derived a way to perform underdetermined source separation using GP models when the mixtures are the convolved sums of several independent GPs. Given some covariance functions and mixing filters, we saw that stating the problem in terms of GPs provides a principled way to estimate the source signals that minimize the mean squared error. In the GP framework, optimization of the hyperparameters is done through maximization of the marginal log-likelihood of the mixtures given the model.

To our knowledge, very few references are available to date on the topic. For example, [33] performs source separation using GPs in the determined case, but the covariance functions are therein applied on the outputs of the source signals rather than on the coordinates themselves (i.e. time or spatial position). A successful application of GPs to a subject close to source separation can also be found in [39] for echo cancellation.

IV. GP APPROXIMATIONS FOR LARGE SIGNALS

A. The need for approximations

The main issue with GP models is the need to invert the $n \times n$ covariance matrix of the learning points for inference (4) and for each evaluation of the observation likelihood in (6). In many areas of interest, we cannot afford to handle such a big matrix, since it is not computationally tractable. In audio signal processing for example, values such as $n \approx 10^7$ are common and GP models cannot be used without a significant reduction of the computational cost of the method.

In order to address this issue, many authors have proposed *sparse* approximation techniques [31], [34], [37], [38], [40] over the years that all aim at making GP inference possible for large datasets. As highlighted in [34], many methods rely on the choice of a small set of input points called *the inducing inputs* to approximate the posterior distribution at test points X^*. Among those methods, we can mention the Fully Independent Conditional (FIC) approximation [34], [38], that considers all the test points and the learning points independent given the inducing inputs. This leads to a very important reduction of the computational burden, but heavily relies on the density of the inducing points [37] to yield good estimates. Another approximation called Partially Independent Conditional (PIC) [37] no longer makes the assumption that both the training and test cases are independent given the inducing points, but rather that each of them not only depends on the possibly remote inducing points, but also on a limited number of other learning points nearby. This technique has the advantage of producing better estimates than FIC, while maintaining an easy inversion of the $n \times n$ covariance matrix that is now block-diagonal. Its main disadvantage is to lead to discontinuities of the estimates between the blocks, which may be problematic for some applications such as audio processing.

Another very attractive direction of research in the last few years has been the consideration of covariance functions with Compact Support (CS) [31], [40], i.e covariance functions $k(x, x')$ such that $\|x - x'\| > l \Rightarrow k(x, x') = 0$ for some given scale l. The idea underlying these techniques is to consider that if they are sufficiently far from each other, two points will be independent. If such covariance functions are used, the covariance matrix is sparse and inference through Cholesky decompositions is done much faster [40]. The main issue with this approach is to design covariance functions that correspond to some prior knowledge about the sources and that have CS at the same time.

In sections IV-B and IV-C, we introduce a general method for fast inference in GP models based on *framing* and that is a direct generalization of the common practice in audio signal processing.

Another important computational simplification is introduced in sections IV-D and IV-E when the signals are regularly sampled. In that case, we show that when the covariance functions are assumed stationary and separable, exact inference can be done extremely efficiently in the frequency domain.

When both approaches are combined into so called *locally-scaled* and *framewise-independent* stationary covariance functions, we show in section IV-F that inference and learning of hyperparameters become equivalent to recent and powerful Nonnegative Tensor Factorization (NTF) techniques [9].

B. Frames

In audio signal processing, it is common to split the signals into overlapping *frames* and to process the frames separately. Formally, the frames $\{y_i(x')\}_{i \in \mathbb{N}}$ are defined as small portions of the original signal. The advantage of the technique is that the frames are small and can be easily processed. The original signals can then be recovered through a deterministic *overlap-add* procedure: each frame is multiplied by a *weighting function* $g : \mathcal{X}' \to \mathbb{R}^+$ to ensure smooth transitions between the frames and is added to the reconstructed signal. g is often a HANN or a triangular window.

This idea can very well be generalized in any dimension D. Instead of considering the original signal y, we can split it into overlapping frames of smaller dimension. To this end, we consider a *frame input set* $\mathcal{X}' \subset \mathcal{X}$, a summable *weighting function* $g : \mathcal{X}' \to \mathbb{R}^+$ and a set of *frame positions* $\{t_i \in \mathcal{X}\}_{i \in \mathbb{N}}$ such that:

$$\forall x \in \mathcal{X}, I_x \triangleq \{i \in \mathbb{N} : x - t_i \in \mathcal{X}'\} \neq \varnothing \tag{12}$$

I_x is thus the set of frame numbers to which the input point x is mapped. Condition (12) ensures that each point of the signal is represented in at least one frame. Finally, given some signal $\{y(x)\}_{x \in \mathcal{X}}$, we can make the assumption that there is a set of *frames* $\{y_i(x)\}_{i \in \mathbb{N}, x \in \mathcal{X}'}$, also noted $\mathcal{G}\{y\}$ in the following, such that:

$$\forall x \in \mathcal{X}, y(x) = \frac{1}{\sum_{i \in I_x} g(x - t_i)} \sum_{i \in I_x} g(x - t_i)\, y_i(x - t_i) \tag{13}$$

When considering a finite set X of n input points, we only need to consider the frames $I \triangleq \bigcup \{I_x\}_{x \in X}$

that contain at least one input point from X. Let $X'_i = \{x' \in \mathcal{X}' \mid \exists x \in X : x' = x - t_i\}$ be the finite set of points from $\mathcal{X}'$ to which the elements of X in the scope of frame i are mapped and let[4] $L_i = \#(X'_i)$. Given some signal $\{\mathbf{y}(x)\}_{x \in X}$, it is always possible to build a set of frames obeying (13). This can be achieved by choosing X'_i and $\mathbf{y}_i(x)$ such that:

$$\forall (i, x') \in I \times \mathcal{X}', (t_i + x' \in X) \Rightarrow \begin{cases} x' \in X'_i \\ \mathbf{y}_i(x') = \mathbf{y}(t_i + x') \end{cases} \tag{14}$$

When they make use of framing, usual methods focus on the frames $\mathcal{G}\{y\}$ as the signals of interest rather than on y. Indeed, a good model for $\mathcal{G}\{y\}$ is *de facto* a good model for y since it can be computed deterministically from $\mathcal{G}\{y\}$. In such methods based on framing, the set (14) of frames is usually taken as being the observation. From our point of view, the frames are simply another process which is indexed on $\mathbb{N} \times \mathcal{X}'$ and from which we can deterministically recover y which is indexed on $\mathcal{X}$.

C. Frame-wise independence assumption

Given a signal $\{y(x)\}_{x \in \mathcal{X}}$ and a corresponding set of frames $\{y_i\}_{i \in \mathbb{N}}$, a classical assumption consists in writing that the different y_i are independent. As y can be deterministically computed from $\{y_i\}_{i \in \mathbb{N}}$, this is written:

$$\log p(\mathbf{y} \mid X, \Theta) = \sum_{i \in I} \log p(\mathbf{y}_i \mid X'_i, \Theta) \tag{15}$$

If $\mathcal{G}\{y\}$ is modeled as a GP, the frame-wise independence assumption is equivalent to modeling the covariance function $k((i, x), (i', x'))$ of $\mathcal{G}\{y\}$ as:

$$k((i, x), (i', x')) = \delta_{ii'} k_i(x, x') \tag{16}$$

with k_i being the covariance function of the GP $\{y_i(x)\}_{x \in \mathcal{X}'}$. Let X be a finite set of input points from $\mathcal{X}$, $\mathbf{y}$ a process indexed on X and $I \triangleq \bigcup \{I_x\}_{x \in X}$ be the corresponding frame indexes for a framing $\mathcal{G}$. Let n_I be the number of frames. If we model $\mathcal{G}\{\mathbf{y}\}$ as a GP, we readily see that it is equivalent to a multiple output GP as seen in section III-C with n_I outputs whose input set is $\mathcal{X}'$. We can thus stack its outputs and observe that the corresponding covariance matrix is block-diagonal due to the frame-wise independence assumption. Its inverse is thus easily computed.

The main computational trick involved by framing is hence to split the signal into overlapping frames, with a synthesis scheme that allows perfect reconstruction. Then, the frames are supposed to be independent and the corresponding covariance matrix becomes block diagonal. The advantage of this method is that when the frames are overlapping, each point estimate is a smoothing of several estimates computed in the different frames that contain this point, thus avoiding systematic discontinuities. In Figure 2, we illustrate this advantage of framing over PIC to produce smooth estimates in a very simple regression problem.

[4]For a countable set X, $\#(X)$ denotes the number of elements in X.

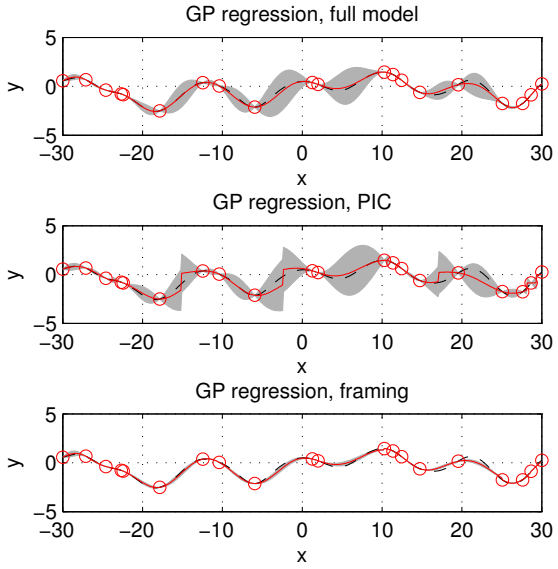


Figure 2. A simple regression example with a full GP model (up), PIC (middle) and framing (bottom) . The dotted and red lines are respectively the true and estimated signal. The grey area represents three standard deviations of the estimates around the mean and the circles stand for observed points.

The connections between *frame-wise independent* correlation functions and other existing models such as PIC or CS covariance functions [31], [40] are numerous. Firstly, we see that framing without overlap between the frames is very similar to PIC except for the fact that it does not take inducing points outside the frames into account. In framing, such remote inducing points are handled through the use of overlapping frames of different scales. Secondly, when considering (13), which gives the expression of the signal given its constitutive frames, we can straightforwardly compute the covariance function of the signal y itself given the covariance functions of the different frames. This computation is actually very similar to that led in [31] where the *basis function* used in the computation becomes the *weighting function* g we considered here. The basis function proposed in [31] is precisely the HANN window which is a very popular choice for g in audio processing.

Of course, as in practice the frames are built using expression (14), there is a duplication of the samples that belong to overlapping frames and the independance assumption between the frames may seem unjustified. Nevertheless, the idea underlying framing is that even if the occurrence of one point from X in some frame gets duplicated in another, and even if the corresponding observed values are connected or equal, they are supposed to be produced by two different underlying processes that do not share the same covariance function.

Still, there are interesting conceptual issues raised by framing that should be more thoroughly studied in the future. In particular, contrary to PIC, framing as it was exposed here suffers from overconfidence. This can be seen by computing the variance of the estimates for y given by (13) and then noticing that the more a point will get duplicated in different frames because of overlap, the smaller the a posteriori variance

of this point will become. This is due to the independence assumption between the frames. If this assumption was strictly legitimate, this diminution of the variance would be justified. However, as the overlap between the frames gets large, independence cannot be a valid assumption anymore and the variances get underestimated. This is illustrated in Figure 2 where the overlap was large.

Even if overlapping the different frames is common practice in audio signal processing, its consequences on statistical models has been largely neglected. In [26], LE ROUX et al. devise practical ways of consistently handling the dependencies between the frames as a postprocessing step. Still, to our knowledge, practical statistical models that fully take the overlap between the frames into account while remaining computationally tractable are yet to be proposed.

D. Stationarity assumption for regularly sampled signals

In this section, we assume that the signals are defined on $\mathcal{X} = \mathbb{R}^D$ for $D \geqslant 1$ and that $x \in \mathcal{X}$ can be written $x = (x_1, \cdots, x_D)$. We will moreover assume that all the covariance functions k that we consider are *separable*, i.e. there are D covariance functions $k^{(d)}$ such that:

$$\forall (x, x') \in \mathcal{X}^2, k(x, x') = \prod_{d=1}^{D} k^{(d)}(x_d, x'_d). \tag{17}$$

It is readily shown that this assumption implies that all the covariance matrices K considered can be expressed as a Kronecker product[5] of D covariance matrices $K^{(d)}$ of lower dimensions:

$$K = K^{(1)} \otimes K^{(2)} \cdots \otimes K^{(D)} \triangleq \bigotimes_{d=1}^{D} K^{(d)}. \tag{18}$$

From now on, we suppose that the points are regularly sampled. This is equivalent to assuming that any signal $\mathbf{y}$, $\mathbf{f}_m$ or $\mathbf{k}$ considered is the vectorization[6] of a corresponding underlying D-dimensional tensor $\underline{y}$, $\underline{f_m}$ or $\underline{k}$. Indeed, we will show in section IV-D1 that computations can be very concisely written using these tensors, which are actually natural to consider. For example, when $D = 2$, it makes sense to directly think of regularly sampled signals as matrices instead of their vectorized counterpart.

1) GPs for the separation of stationary mixtures: As seen in section II-C, a *stationary* covariance function $k(x, x')$ between two input points can be expressed as a function of their difference $\tau = x - x'$. It is noted $k(x - x')$. If all covariance functions considered are stationary, the computations become particularly simple.

Indeed, let us assume that a mixture $\{y(x)\}_{x \in \mathcal{X}}$ is the sum of several GPs $\{f_m(x)\}_{m=1\ldots M, x \in \mathcal{X}}$ whose covariance functions $k_m(x - x')$ are all stationary, and let us furthermore suppose that we are interested in separating the different sources for all points in X, thus having $X^* = X$. The covariance matrix K_y of y is given by : $K_y = \sum_{m=1}^{M} K_m$ where K_m is the covariance matrix of source m.

[5]See [9] for a concise introduction to tensor algebra.

[6]Vectorization is done recursively. For example, with $D = 2$ where tensors are matrices, it is done one row after the other.

Considering (18), K_m is given by: $K_m = \bigotimes_{d=1}^{D} K_m^{(d)}$ where $\left[K_m^{(d)}\right]_{i,j} = k_m^{(d)}(x_{i,d} - x_{j,d})$. $K_m^{(d)}$ can approximately be considered as circulant[7]. It is readily shown that any circulant matrix M can be expressed as $M = W_F^* \Lambda W_F$ where W_F is the discrete Fourier transform matrix[8] and where Λ is diagonal. Thus, for all m and d, there is a diagonal positive semidefinite matrix $\mathrm{diag} S_m^{(d)}$ such that $K_m^{(d)} \approx W_F^* \mathrm{diag} S_m^{(d)} W_F$ where the vector $S_m^{(d)}$ is the discrete Fourier transform of $\tau \mapsto k_m^{(d)}(\tau)$. We can thus write K_y as:

$$K_y = \sum_{m=1}^{M} \bigotimes_{d=1}^{D} W_F^* \mathrm{diag} S_m^{(d)} W_F. \tag{19}$$

Using classical results from tensor algebra, (19) can be written:

$$K_y = \left(\bigotimes_{d=1}^{D} W_F^*\right) \left(\sum_{m=1}^{M} \bigotimes_{d=1}^{D} \mathrm{diag} S_m^{(d)}\right) \left(\bigotimes_{d=1}^{D} W_F\right). \tag{20}$$

We can use this property to extract a given source m_0, and write[9] (4) as:

$$\overline{\mathbf{f}_{m_0}^*} = \left(\bigotimes_{d=1}^{D} W_F^*\right) \left(\frac{\bigotimes_{d=1}^{D} \mathrm{diag} S_{m_0}^{(d)}}{\sum_{m=1}^{M} \bigotimes_{d=1}^{D} \mathrm{diag} S_m^{(d)}}\right) \left(\bigotimes_{d=1}^{D} W_F\right) \mathbf{y}. \tag{21}$$

Introducing the D-dimensional tensor[10]

$$\underline{S}_m = S_m^{(1)} \circ S_m^{(2)} \cdots S_m^{(D)} \triangleq \bigcirc_{d=1}^{D} S_m^{(d)} \tag{22}$$

as the *model for source* m and $\mathcal{F}_D\left\{\underline{y}\right\}$ as the D-dimensional Fourier transform of $\underline{y}$, we can simply write (21) in tensor form as:

$$\mathcal{F}_D\left\{\overline{\underline{f}_{m_0}^*}\right\} = \left(\frac{\underline{S}_{m_0}}{\sum_{m=1}^{M} \underline{S}_m}\right) \cdot \mathcal{F}_D\left\{\underline{y}\right\} \tag{23}$$

which is similar to the classical Wiener filter for stationary processes. The differences between this expression and the classical one is firstly that it is valid for any dimension D of the input space and secondly that it is not restricted to the case of only two stationary sources. The sources themselves can be recovered through an inverse D-dimensional Fourier transform. The nonnegative tensor $\underline{S}_m$ can be understood as the D-dimensional Fourier transform of the stationary covariance function k_m. Note that the complexity of this *exact* GP inference method relying on stationarity of the covariance functions and on regular sampling is $\mathcal{O}(n \log n)$, and it is dominated by the computation of Fourier transforms, for which there exist very efficient and specialized algorithms. If $\mathcal{F}_D\left\{\underline{y}\right\}$ is known beforehand, the complexity of (23) decreases to $\mathcal{O}(n)$ which is remarkable for an exact GP inference technique.

[7]If the signal is regularly sampled, this approximation holds when the number n_d of points along dimension d tends to infinity or when $k^{(d)}(\tau)$ is periodic of period $\frac{n_d}{p}$ with $p \in \mathbb{N}^*$.

[8]W_F^* denotes the complex conjugate of W_F.

[9]$\frac{A}{B}$ and $A.B$ are respectively the element-wise division and multiplication of A and B.

[10]$\circ$ denotes the outer product.

2) Marginal likelihood for stationary sources: When all the covariance functions considered are stationary and parameterized by some hyperparameter set Θ that consists of their respective D-dimensional Fourier transforms, i.e. $\Theta = \{\underline{S}_1 \cdots \underline{S}_M\}$, it can readily be shown that the marginal log-likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations given regularly spaced input points and the hyperparameters simplifies from (6) to:

$$\log p(\mathbf{y} \mid X, \{\underline{S}_1 \cdots \underline{S}_M\}) = -\frac{1}{2} \sum_{i_1, \cdots i_D} \left[\frac{\left| \left[\mathcal{F}_D \left\{\underline{y}\right\}\right]_{i_1, \cdots i_D} \right|^2}{\sum_{m=1}^M \left[\underline{S}_m\right]_{i_1, \cdots i_D}} + \log \sum_{m=1}^M \left[\underline{S}_m\right]_{i_1, \cdots i_D} \right] + \text{Cte} \quad (24)$$

Considering (24), we see that it is equivalent up to an additive constant independent of Θ to half the opposite IS divergence[11] between[12] $\left|\mathcal{F}_D \left\{\underline{y}\right\}\right|^{.2}$ and $\sum_{m=1}^M \underline{S}_m$:

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2} D_{\text{IS}} \left(\left|\mathcal{F}_D \left\{\underline{y}\right\}\right|^{.2} \mid \sum_{m=1}^M \underline{S}_m \right) + \text{Cte} \quad (25)$$

The evaluation of the likelihood can be done in $\mathcal{O}(n \log n)$ operations when the signals are regularly sampled and the covariance functions are stationary. If the squared D-dimensional Fourier transform $\left|\mathcal{F}_D \left\{\underline{y}\right\}\right|^{.2}$ of the signal is known beforehand — it is typically computed only once — the computational complexity is reduced to $\mathcal{O}(n)$.

E. Locally stationary covariance functions

Let $\{y(x)\}_{x \in \mathcal{X}}$ be a particular signal, observed on a finite input set $X \in \mathcal{X}^n$ and let $\left\{y_i \in \mathbb{R}^{X_i'}\right\}_{i \in I}$ be a set of n_I corresponding frames. As in section IV-C, we can assume that the frames are independent and we can further suppose that the covariance function k_{im} of source m within frame i is stationary. This means that we model each source as being composed of several locally stationary frames, each of which has its own covariance function. The resulting signal is *not* supposed stationary with this assumption, only its restrictions to small regions of the input space $\mathcal{X}$ are assumed stationary.

Let us denote $\underline{\mathbf{Y}}$ the $(D+1)$-dimensional tensor whose last dimension goes over the frames and whose first D dimensions for a fixed frame contain the D-dimensional Fourier transform of the signal tensor for this frame as in section IV-D. As this tensor is called the Short Term Fourier Transform (STFT) of the signal when $D = 1$, it will be called the STFT tensor of the mixture. We define the STFT tensor $\underline{\mathbf{F}_m}$ of the sources and the STFT tensor $\underline{\mathbf{S}}_m$ of the covariance function of source m in the same way. We can use the results from the previous section for each frame and for source m_0: the MMSE estimate $\underline{\overline{\mathbf{F}}^*_{m_0}}$ of $\underline{\mathbf{F}_{m_0}}$ is given by:

$$\underline{\overline{\mathbf{F}}^*_{m_0}} = \frac{\underline{\mathbf{S}}_{m_0}}{\sum_{m=1}^M \underline{\mathbf{S}}_m} \cdot \underline{\mathbf{Y}}. \quad (26)$$

[11] $D_{\text{IS}}\left(\underline{x} | \underline{y}\right) \triangleq \sum_{i_1 \cdots i_D} \left[\frac{[\underline{x}]_{i_1 \cdots i_D}}{[\underline{y}]_{i_1 \cdots i_D}} - \log \frac{[\underline{x}]_{i_1 \cdots i_D}}{[\underline{y}]_{i_1 \cdots i_D}} - 1 \right]$.

[12] For a matrix M, $\left[M^{\cdot 2}\right]_{ij} \triangleq M_{ij}^2$

The sources can then be recovered by first applying an inverse D-dimensional Fourier transform to the estimate (26) for each frame, and then using the reconstruction scheme (13) to obtain the estimated sources in the original input space $\mathcal{X}$.

Let $\Theta = \{\underline{\mathbf{S}}_1, \cdots \underline{\mathbf{S}}_M\}$ be the models for the sources. The marginal likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations can similarly be shown to be:

$$\log p(\mathbf{y} \mid X, \Theta) = -\frac{1}{2} D_{\text{IS}} \left(|\underline{\mathbf{Y}}|^{.2} \mid \sum_{m=1}^M \underline{\mathbf{S}}_m \right) + Cte \quad (27)$$

where the constant is independent of Θ. This very simple expression can be computed in $\mathcal{O}(n)$ when $|\underline{\mathbf{Y}}|^{.2}$ is known and permits to efficiently proceed to hyperparameters learning as demonstrated in section IV-F.

F. Putting structures over the covariances

Given some regularly sampled signal tensor $\underline{y}$ and its corresponding STFT tensor $\underline{\mathbf{Y}}$ as defined in section IV-E, we have seen that source separation can be very efficiently performed provided some $(D+1)$-dimensional model $\underline{\mathbf{S}}_m$ is known for every source. As highlighted by CEMGIL *et al.* in [7] or [8] for the case of audio processing ($D = 1$), the important issue raised by this probabilistic framework becomes devising realistic but effective models for the nonnegative sources parameters $\underline{\mathbf{S}}_m$.

In audio signal processing (D=1), the result (26) is known as adaptive or generalized Wiener filtering and many methods for source separation such as [8], [32] use this technique in a principled way to recover the sources in the frequency domain. Those studies state their probabilistic model in the frequency domain where the time-frequency bins are supposed to be distributed with respect to independent Gaussian distributions. In our approach, the model is expressed directly in the original input space. The two points of view are actually equivalent: a stationary GP has an independently distributed Gaussian representation in the frequency domain. $\underline{\mathbf{S}}_m$ can hence be seen either as the STFT tensor of a covariance function or as a tensor containing the variances of the independent components of $\underline{\mathbf{Y}}$.

Focusing on the second interpretation of $\underline{\mathbf{S}}_m$, recent studies [8], [12], [13] proposed to model these tensors as Gamma Markov Random Fields (GMRF). This is a sensible choice indeed, because such models guarantee the nonnegativity of all the elements of $\underline{\mathbf{S}}_m$ while implementing the knowledge that for a given source, the spectrum is much likely to exhibit some continuity over time, or over the frequencies, or over both. As GMRF do not provide a closed-form expression for the marginal log-likelihood of the observations, the learning of hyperparameters has to be done using approximate methods. To this end, DIKMEN *et al.* [12] propose to use contrastive divergence [22] and report good results. To our knowledge, no generalization of GMRF has yet been published for input spaces of dimension greater than 2, but GP modeling may greatly benefit from such an extension.

Another point of view is to introduce some deterministic structure into the covariance functions of the GPs. A simple assumption to this end is to consider that for a given source

m, the covariance functions of the different *independent* frames are *stationary* and *locally scaled,* i.e identical up to an amplification gain depending on the frame. The model for source m and frame i can then be written:

$$\underline{\mathbf{S}}_{im} = H_{im}\, \underline{S}_{0,m} \tag{28}$$

where $\underline{S}_{0,m} = \bigcirc_{d=1}^{D} S_{0,m}^{(d)}$ is the D-dimensional Fourier transform of some *template* covariance function $k_{0,m}$ for source m that is independent of the frame index i. We get:

$$\underline{\mathbf{S}}_{m} = \left(\bigcirc_{d=1}^{D} S_{0,m}^{(d)}\right) \circ H_m \tag{29}$$

where $H_m = (H_{1m} \cdots H_{n_I m})$ denotes the amplification gains of the covariance function for source m on the different frames. Considering (29) we readily see that it is equivalent to a classical Nonnegative Tensor Factorization (NTF) model called Canonical Polyadic (CP) decomposition[13]. The different parameters become $\Theta = \left\{\left\{H_m, S_{0,m}^{(1)} \cdots S_{0,m}^{(D)}\right\}_{m=1\cdots M}\right\}$ and can be estimated by standard CP algorithms using the IS-divergence function. See [9] for a review of these models and algorithms.

G. Optimization

We have shown how GP learning can be connected to recent NTF techniques by factorizing the covariance structure of the GP model into a CP decomposition. More generally and depending on the application, the covariances can be factorized in many other ways to account for some prior knowledge we may have concerning the structure of the sources. For example, if the covariances are considered to be the outer product of some shared dictionaries, the tensor decompositions to be used become particular cases of Block Components Decompositions as introduced in [25]. Many very informative models can be designed this way, that decompose the covariance structure of the sources onto sophisticated dictionaries. In music processing ($D = 1$) for example, [41] decomposes the covariances into templates of harmonic bases. Other models of this type have also been used to model and extract singing voice signals from polyphonic mixtures with very promising results [14].

In any case, when an appropriate model has been chosen for $\{\underline{\mathbf{S}}_m\}_{m=1\cdots M}$, we have seen in section IV-E that hyperparameters learning can be done by minimizing the IS-divergence between $\sum_m \underline{\mathbf{S}}_m$ and $|\underline{\mathbf{Y}}|^{.2}$ through tensor factorizations. Efficient algorithms for IS-NTF can be found in the litterature, for example in [9].

V. Evaluation

In this section, we demonstrate the performance of the proposed approach based on GP models for the separation of real-valued mixtures. In section V-A, we first show that GP can easily be used for the separation of synthetic 2D random fields, or textures ($D = 2$). Then, we show in section V-B how GP can be used for the separation of drums signals in real polyphonic stereo recordings.

[13]CP is also called PARAFAC or CANDECOMP [9].

A. Synthetic additive textures

In this section, we set $D = 2$, which means that we aim at separating additive functions $f_m(x_1, x_2)$ defined on the plane and summed in an observable mixture signal $y(x_1, x_2)$. For this toy example, we will consider the case of one mixture ($K = 1$) that is the sum of $M = 2$ stationary sources[14]. Following the notations that were introduced in section IV-D, we will thus suppose that the mixture tensor $\underline{y}$ is the sum of two sources tensors $\underline{f}_1$ and $\underline{f}_2$. The corresponding vectors y, f_1 and f_2 will denote the vectorization of these tensors one row after the other. X denotes the corresponding coordinates.

In this experimental setup, the dimensions of the sources and mixtures tensors are 500×500 each, leading to $n = 250000$. In the following, we will assume that the covariance function of each source along each dimension is stationary. For the experiment, the covariance functions were arbitrarily set to:

$$k_m^{(d)}(x_d, x_d') = \exp\left(-\frac{2\sin^2\frac{\pi(x_d - x_d')}{T_{m,d}}}{l_{m,d}^2} - \frac{(x_d - x_d')^2}{2\lambda_{m,d}^2}\right) \tag{30}$$

where $\{T_{m,d}, l_{m,d}, \lambda_{m,d}\}_{m,d}$ are scalar parameters.

This model implements a particular prior knowledge where the sources are known to exhibit some kind of complex structure. More specifically, source m is known to be pseudo-periodic of period $(T_{m,1}, T_{m,2})$ and $(l_{m,1}, l_{m,2})$ controls the smoothness within one period. A further lengthscale $(\lambda_{m,1}, \lambda_{m,2})$ controls global covariance between two input points. In the particular example shown in Figure 3, the parameters were:

m	$\lambda_{m,1}$	$\lambda_{m,2}$	$T_{m,1}$	$T_{m,2}$	$l_{m,1}$	$l_{m,2}$
1	100	100	50	20	0.5	0.7
2	40	4	25	$+\infty$	0.7	N/A

1) Data synthesis: Generating a realization of a GP with some known covariance matrix K is generally addressed through Cholesky or Singular Value Decompositions (SVD) of the covariance matrix [35]. As we have $n = 250000$, we cannot naively implement this idea here. A simple way to circumvent this problem is to write K as in (18) and then to perform a Cholesky decomposition of each $K^{(d)}$ to get $K^{(d)} = L^{(d)} L^{(d)\top}$. The Cholesky decomposition of $K = \bigotimes_{d=1}^{D} L^{(d)} L^{(d)\top}$ is finally obtained by $K = \left(\bigotimes_{d=1}^{D} L^{(d)}\right)\left(\bigotimes_{d=1}^{D} L^{(d)}\right)^{\top}$ and a realization of this GP can be very easily generated. Indeed, let R be a vector of length n whose entries are i.i.d. Gaussian random variables of unit variance. K is the covariance matrix of $\left(\bigotimes_{d=1}^{D} L^{(d)}\right) R$.[15]

2) Source separation: In this very simple experimental setup, we consider that the 12 hyperparameters for the sources covariance functions (30) are known beforehand.

[14]This usecase is common in geostatistics: the observed signal is often modeled as the sum of the signal of interest whith a contaminating white Gaussian noise [11]. Estimating the value of the target signal through Kriging is hence a special case of source separation with GP priors.

[15]We can further speed up this computation by using the fact that for $\mathbf{c} = \mathrm{vec}(\underline{C})$ and matrices A and B of appropriate size, $(A \otimes B)\,\mathbf{c} = A\underline{C}B^{\top}$. This avoids considering such a big matrix as $\left(\bigotimes_{d=1}^{D} L^{(d)}\right)$.

We can perform separation through the exact method presented in section IV-D1. To this purpose, we can build the spectral covariance tensor $\underline{\mathbf{S}}_m$ of each source as the outer product of the Fourier transforms of (30) along each dimension and then perform separation in the frequency domain as in (23). The sources are recovered through an inverse 2-dimensional Fourier transform. It is worth noticing here that the computations are performed extremely rapidly since they only involve element-wise multiplications of 500×500 images, instead of the inversion of the 250000×250000 covariance matrix required by the basic GP setup. Overall computations for this example — synthesis and source separation — are achieved in less than 3 seconds on a standard laptop computer.

Results for one example are shown in Figure 3. The average Signal to Error ratio obtained on 50 experiments was of 8dB, which is very encouraging.

B. Separation of drums signal in polyphonic music

In this section, we apply the general framework we have presented in sections III and IV to the separation of drums signals in polyphonic music. The regularly sampled signals we consider are thus defined on the input space $\mathcal{X} = \mathbb{Z}$ of dimension $D = 1$.

Separation of the drums track from polyphonic music is a challenging task that has already been addressed in several studies such as [18], [21]. Whereas HÉLEN and VIRTANEN [21] perform a Nonnegative Matrix Factorization (NMF) of the mixture and then group the different components obtained through a classification procedure, GILLET and RICHARD [18] decompose the mixture signal with spectral templates learned from a drums database.

In section V-B1, we introduce a GP model for this task and in section V-B2, we compare its performance with the state of the art [18].

1) GP model: The observed mixtures $y(x)$ are supposed to be the sum of two independent GPs $s_d(x)$ and $s_r(x)$ corresponding respectively to the *drums* and the *musical residual* tracks. We assume that some framing $\mathcal{G}\{y\}$ with n_I frames of same length as defined in section IV-B is available for the mixtures and we aim at estimating the framings $\mathcal{G}\{s_d\}$ and $\mathcal{G}\{s_r\}$ of the different sources such that $\mathcal{G}\{y\} = \mathcal{G}\{s_d\} + \mathcal{G}\{s_r\}$.

We suppose that each of the signals s_d and s_r are themselves the sum of several independent processes called *components*. In our example, the R_d different components of s_d are the five most common sources we find in a drums signal, e.g. kick drum, snare drum, hihat, bells and clap sounds. The R_r different components of the musical residual are all the other elements composing the polyphonic mixture. This assumption can be written $s_d = \sum_{m=1}^{R_d} f_m$ and $s_r = \sum_{m=R_d+1}^{R_d+R_r} f_m$.

In the model we are considering, we will assume that all the components f_m are GP whose covariance functions k_m are *locally scaled, frame-wise independent* and *stationary* as defined in section IV-F. For some frame i, they can thus be expressed as:

$$k_{im} = H_{im}\, k_{0,m} \tag{31}$$

where $k_{0,m}$ denotes the template stationary covariance function for component m and $H_m = (H_{1m} \cdots H_{n_I m})$ are the nonnegative activation gains of this component within the frames. Introducing the Fourier transform $S_{0,m}$ of $k_{0,m}$ and using the method presented in section IV-F, the MMSE estimate $\hat{\underline{F}}_d$ of the STFT $\underline{F}_d$ of the drums signal is given by:

$$\hat{\underline{F}}_d = \frac{\sum_{m=1}^{R_d} S_{0,m} \circ H_m}{\sum_{m=1}^{R_d+R_r} S_{0,m} \circ H_m} \cdot \underline{\mathbf{Y}} \tag{32}$$

where $\underline{\mathbf{Y}}$ is the STFT of the mixtures. The model $\underline{\mathbf{S}} \triangleq \sum_m \underline{\mathbf{S}}_m$ becomes $\underline{\mathbf{S}} = \sum_{m=1}^{R_d+R_r} S_{0,m} \circ H_m$. Since $D = 1$, this can be written in matrix form as $\underline{\mathbf{S}} = WH$ where $S_{0,m}$ is the m^{th} column of W and H_m is the m^{th} row of H. As we have shown in section IV-F, the optimization of the hyperparameters $\Theta = \{W, H\}$ through likelihood maximization is thus equivalent to the minimization of the IS distance between the power STFT $|\underline{\mathbf{Y}}|^{\cdot 2}$ and the product WH, yielding a NMF model as in [9], [17], [27], [32].

Some other kind of knowledge has now to be put into the model so that it can be useful in practice, since we have not yet made any distinction between the covariance functions of the drums components and those of the musical residual signal. A very simple and computationally cheap solution to this problem is to appropriately initialize some of the hyperparameters. In this experiment, we will focus on the meaning of the activation gains H_m of the components as introduced in (31). H_{im} can be understood as a magnitude parameter for component m into frame i. A good way to initialize all these parameters $\{H_m\}_{m=1\cdots R_d}$ for the drums signal is simply to use an *onset detector* such as [15]. Indeed, if an onset detector feature has a high magnitude in some frame i, then some drums component must be active in it. The onset detector of [15] was hence used in R_d different frequency bands of the mixture STFT, yielding R_d signals. These signals were used to initialize the activation gains $\{H_m\}_{m=1\cdots R_d}$ and all the other hyperparameters of the model were randomly initialized. A NMF was then applied using this initialization and separation was performed using (32).

2) Results: The proposed GP separation method was tested on ten 30-second excerpts sampled at 44.1kHz from the Quaero[16] source separation corpus. The excerpts featured many different kinds of music signals, including pop, electropop, rock, reggae and bossa. For each of these excerpts, the ground truth drums and musical residual signals are known for evaluation but the separation systems can only observe their mixtures. On average, the relative amplitude $20 \log_{10} \frac{\sum_x |s_r(x)|}{\sum_x |s_d(x)|}$ of the musical residual signal was set to $+6$dB compared to the drums signal.

We applied the method proposed by GILLET and RICHARD in [18] on the same mixtures and the quality of the results were quantified through the BSSEVAL toolbox [42]. The separation quality was evaluated both on the drums signals and on the musical residual signal.

The metrics obtained through BSSEVAL include the Source to Distortion Ratio (SDR), the Source to Artifact Ratio (SAR)

[16] http://www.quaero.org

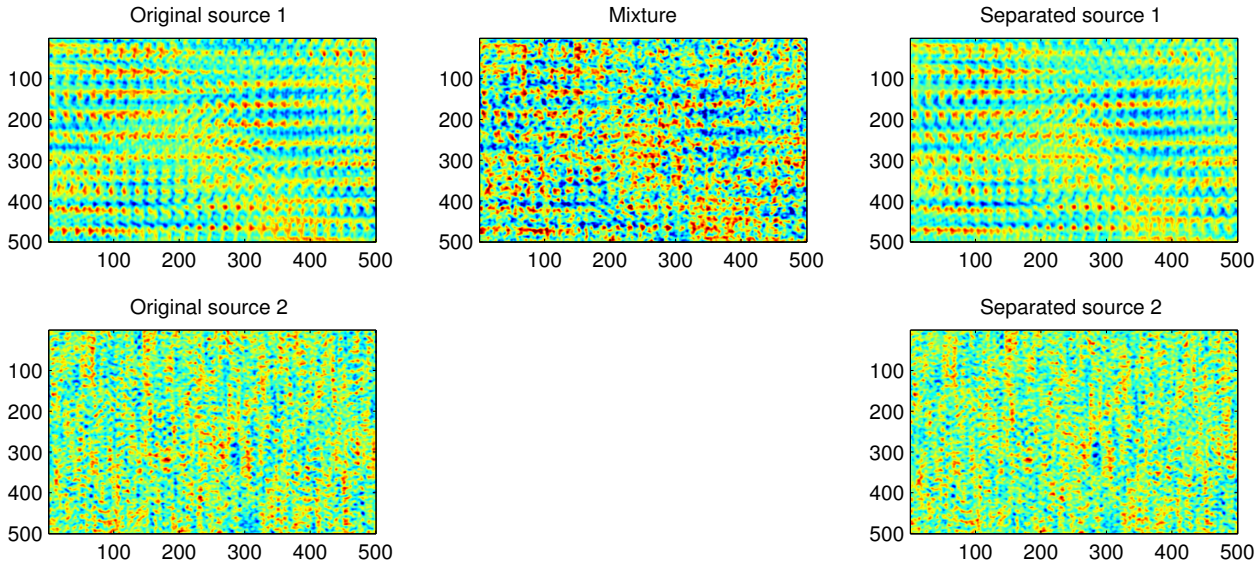


Figure 3. GP for the separation of two stationary random fields ($D = 2$) using a Gaussian Process model. On the left are the original sources. On the center is the mixture and on the right are the estimated sources.

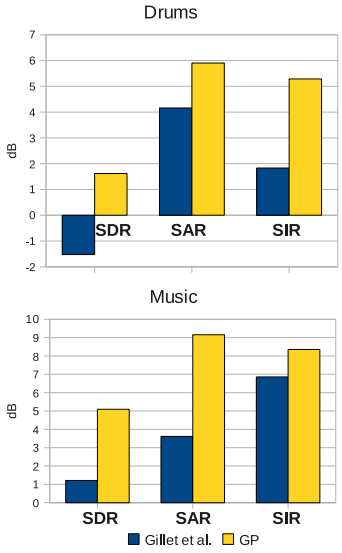


Figure 4. Evaluation of the separation quality for the extraction of the drums track (top) and the musical residual signal (bottom). Higher is better.

and the Source to Interference Ratio (SIR) that are all expressed in dB. Whereas the SDR is a global measure of separation quality, the SAR and SIR respectively measure the amount of separation/reconstruction artifacts and the amount of energy from the other sources. Results are given in Figure 4.

From Figure 4, we can see that the GP model presented in section V-B1 very well manages to separate the drums and musical residual signals on many different kinds of music and both signals are well recovered. A further feature of this technique is that it is extremely fast: on average, 30 seconds are needed to handle a 30-second long excerpt whereas 300 seconds are needed by [18]. Sound excerpts and a full implementation in Python of this separation technique are freely available on our website[17].

VI. Conclusion

In this study, we have stated the linear underdetermined, instantaneous, convolutive and multiple-output source separation problems in terms of Gaussian processes regression and have shown that it leads to simple formulas to optimally proceed to signals separation w.r.t. the MMSE. The advantages of setting out the source separation problem in terms of GP are numerous.

First, there is neither notational burden nor any conceptual issue raised when using input spaces $\mathcal{X}$ different from $\mathbb{R}$ or $\mathbb{Z}$, thus enabling a vast range of source separation problems to be handled within the same framework. Multi-dimensional signal separation may include audio, image or video sensor arrays as well as geostatistics.

Secondly, GP source separation can perfectly be used for the separation of non locally-stationary signals. Of course, some important simplifications of the computations as presented in sections IV-D and IV-E are lost when using non-stationary covariance functions. Still, the frame-wise independence assumption presented in section IV-C may nonetheless be used in order to make the estimations computationally tractable.

Thirdly, it provides a coherent probabilistic way to take many sorts of relevant prior information into account. Indeed, prior information is encapsulated in the choice of the covariance functions and the framework proposed here thus clearly distinguishes between the optimal separation methods and the particular models considered.

[17] http://www.telecom-paristech.fr/~liutkus/GPSS/

Finally, we have seen that under appropriate assumptions, optimization of the hyperparameters of a GP model is equivalent to a classical NTF using the Itakura-Saito divergence on the spectrogram tensor of the mixtures, thus enabling efficient estimation of the hyperparameters.

Setting the source separation problem in such a unified framework allows it to be considered from a larger perspective where its objective is to separate additive independent functions on arbitrary input spaces that are mathematically characterized by their first and second moments only.

REFERENCES

[1] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 878, Norsk Regnesentral, Oslo, Norway, April 1997.

[2] M. Alvarez and N. D. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Neural Information Processing Systems (NIPS)*, pages 57–64. MIT Press, 2008.

[3] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In *Neural Information Processing Systems (NIPS)*, pages 153–160. MIT Press, 2007.

[4] P. Boyle and M. Frean. Multiple output Gaussian process regression. Technical report, Victoria University of Wellington, April 2005.

[5] P. Boyle and M. R. Frean. Dependent Gaussian processes. In *Neural Information Processing Systems (NIPS)*, pages 217–224. MIT Press, 2004.

[6] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90:2009–2026, October 1998.

[7] A. T. Cemgil, S. J. Godsill, P. H. Peeling, and N. Whiteley. *The Oxford Handbook of Applied Bayesian Analysis*, chapter Bayesian Statistical Methods for Audio and Music Processing. Number ISBN13: 978-0-19-954890-3. Oxford University Press, 2010.

[8] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for Time-Frequency energy distributions. In *Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'07)*, pages 151–154, NY, USA, October 2007.

[9] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, September 2009.

[10] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[11] P.J. Diggle and P. J. Ribeiro. *Model-based Geostatistics*. Springer series in statistics. Springer, 1 edition, March 2007.

[12] O. Dikmen and A. T. Cemgil. Gamma markov random fields for audio source modelling. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):589–601, March 2010.

[13] O. Dikmen and A.T. Cemgil. Unsupervised single-channel source separation using Bayesian NMF. In *Proc. of the 2009 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'09)*, pages 93–96, NY, USA, October 2009.

[14] J.-L. Durrieu, A. Ozerov., C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. In *Proc. 17th European Signal Proc. Conf. (EUSIPCO'09)*, pages 15–19, Glasgow, UK, August 2009.

[15] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. In *In Proc. Digital Audio Effects Workshop (DAFx*, London, UK, September 2003.

[16] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[17] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Proc. of Irish Sig. and Systems Conf. (ISSC'08)*, 2008.

[18] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Trans. on Audio, Speech, and Language Processing,*, 16(3):529 –540, 2008.

[19] A. Girard, C. E. Rasmussen, J. Quiñonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In *Neural Information Processing Systems (NIPS)*, pages 529–536. MIT Press, 2002.

[20] P. W. Goldberg, C. K. I. Williams, and C. M. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Neural Information Processing Systems (NIPS)*, pages 493–499. The MIT Press, 1997.

[21] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. 13th European Signal Processing Conference (EUSIPCO)*, Antalaya, Turkey, 2005.

[22] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002.

[23] E. T. Jaynes and G. L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[24] A.G. Journel and C.J. Huijbregts. *Mining geostatistics*. Academic Press, London ; New York, 1978.

[25] L. De Lathauwer. Decompositions of a higher-order tensor in block terms—part II: Definitions and uniqueness. *SIAM J. Matrix Anal. Appl.*, 30(3):1033–1066, September 2008.

[26] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama. Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency. In *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, pages 89–96, St. Malo, France, September 2010.

[27] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 556–562. The MIT Press, April 2001.

[28] D. MacKay. Gaussian processes - a replacement for supervised neural networks? In *Neural Information Processing Systems (NIPS)*. MIT Press, 1997.

[29] G. Matheron. The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468, 1973.

[30] A. Melkumyan and F. Ramos. Multi-kernel Gaussian processes. In *Neural Information Processing Systems (NIPS)*. MIT Press, 2009.

[31] A. Melkumyan and F. Ramos. A sparse covariance function for exact gaussian process inference in large datasets. In *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*, pages 1936–1942, San Francisco, CA, USA, July 2009. Morgan Kaufmann Publishers Inc.

[32] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 18(3):550–563, March 2010.

[33] S. Park and S. Choi. Gaussian processes for source separation. In *Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'08)*, volume 18, pages 1909–1912, Las Vegas, USA, March 2008.

[34] J. Quiñonero-Candela, C. E. Rasmussen, and R. Herbrich. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, December 2005.

[35] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[36] M. Seeger. Gaussian processes for machine learning. *Int. J. Neural Syst.*, 14(2):69–106, April 2004.

[37] E. Snelson. Local and global sparse gaussian process approximations. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, volume 2, pages 524–531, San Juan, Puerto Rico, March 2007.

[38] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Neural Information Processing Systems (NIPS)*, pages 1257–1264. MIT press, 2006.

[39] J. Ichiro Tomita and Y. Hirai. Acoustic echo cancellation using Gaussian processes. In *(ICONIP'08) 15th Int. Conf. on Neural Information Processing.*, volume 5507 of *Lecture Notes in Computer Science*, pages 353–360. Springer, 2008.

[40] J. Vanhatalo and A. Vehtari. Modelling local and global phenomena with sparse gaussian processes. In *Proc. of 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 571–578, Helsinky, Finland, July 2008. AUAI Press.

[41] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE trans. on Audio, Speech and Language Proc. (TASLP)*, 18(3):528–537, March 2010.

[42] E. Vincent, C. Févotte, and R. Gribonval. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[43] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications.* MIT Press, 1949.

[44] C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models*. Kluwer, 1999.

Antoine Liutkus was born in France on February 23rd, 1981. He received the State Engineering degree from Telecom ParisTech, France, in 2005, and the M.Sc. degree in acoustics, computer science and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005. He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and is currently pursuing the Ph.D. degree in the Department of Signal and Image Processing, Telecom ParisTech.
His research interests include statistical music processing, source separation and machine learning methods applied to signal processing.

Roland Badeau (M'02-SM'10) was born in Marseille, France, on August 28, 1976. He received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D degree from the ENST in 2005, in the field of signal processing. He received the ParisTech Ph.D. Award in 2006, and the Habilitation à Diriger des Recherches degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010.
In 2001, he joined the Department of Signal and Image Processing, Telecom ParisTech (ENST), as an Assistant Professor, where he became Associate Professor in 2005. From November 2006 to February 2010, he was the manager of the DESAM project, funded by the French National Research Agency (ANR), whose consortium was composed of four academic partners. His research interests include high resolution methods, adaptive subspace algorithms, nonnegative factorizations, audio signal processing, and musical applications. Roland Badeau is a Senior Member of the IEEE Signal Processing Society and he is a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state). He is the author of 17 journal papers and 40 international conference papers.

Gaël Richard (SM'06) received the State Engineering degree from Telecom ParisTech, France (formerly ENST) in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in September 2001.
After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 to 2001, he successively worked for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In particular, he was the Project Manager of several large scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, Telecom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves research group. He is a coauthor of over 80 papers and inventor in a number of patents. He is also one of the experts of the European commission in the field of speech and audio signal processing.
Prof. Richard is a member of the EURASIP and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

GENERALIZED WIENER FILTERING WITH FRACTIONAL POWER SPECTROGRAMS

Antoine Liutkus[1] *Roland Badeau*[2]

[1]Inria, Speech processing team, Villers-lès-Nancy, France
[2]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France

ABSTRACT

In the recent years, many studies have focused on the single-sensor separation of independent waveforms using so-called soft-masking strategies, where the short term Fourier transform of the mixture is multiplied element-wise by a ratio of spectrogram models. When the signals are wide-sense stationary, this strategy is theoretically justified as an optimal Wiener filtering: the power spectrograms of the sources are supposed to add up to yield the power spectrogram of the mixture. However, experience shows that using fractional spectrograms instead, such as the amplitude, yields good performance in practice, because they experimentally better fit the additivity assumption. To the best of our knowledge, no probabilistic interpretation of this filtering procedure was available to date. In this paper, we show that assuming the additivity of fractional spectrograms for the purpose of building soft-masks can be understood as separating locally stationary α-stable harmonizable processes, α-harmonizable in short, thus justifying the procedure theoretically.

***Index Terms*—audio source separation, probability theory, harmonizable processes, α-stable random variables, soft-masks**

I. INTRODUCTION

In the past ten years, much research has focused on the *demixing* of musical signals. The objective of such research is to process a musical track so as to recover the original individual sounds that were used for its making. For instance, such a process would permit to automatically recover the voice signal from a song and thus automatically generate a karaoke version as well as solo vocals that could be used for resampling. In the scientific community, each constitutive component —or *stem*— from the mixture is called a *source*, and the problem of demixing is commonly called *audio source separation* [6], [32], [25], [19]. In the literature, both single-channel and multichannel audio source separation have been considered, depending on the number of channels of the mixture signal. For the sake of simplicity, we will only consider the single channel case in this study and leave the multichannel case for future developments.

For achieving single channel audio source separation, an efficient approach is focused on a *filtering* paradigm: each source estimate is obtained by applying a time-varying filter to the mixture. In practice, a time-frequency (TF) representation of the mixture is computed, such as its short-term Fourier transform (STFT), and each source is recovered by multiplying each element in this representation by a gain between 1 and 0, according to whether this point is identified as rather belonging to this source or not, respectively [4], [34], [8], [3]. For one given source, those gains form a *time-frequency mask*, and several ways of designing such masks have been considered in the past.

In the audio source separation literature, an important path of research is to consider the devising of TF masks as a *classification* problem. In that setting, the entries of the mask are either 0 or 1: it is typically assumed that only one source is active for any TF bin, so that the problem becomes to determine the source to which each entry of the mixture STFT is associated to. The separation algorithm hence inputs the mixture and performs a multi-class classification task, where each class corresponds to one source. Among those techniques, we can mention the celebrated DUET [34] and ADRESS [2] algorithms, that classify TF bins according to panning positions in the stereo plane. In the single-sensor case, other works attempt to separate sources with binary masks by using harmonicity assumptions: a melody line is first extracted, and then a binary comb-filter is generated to extract the corresponding source [27]. Other recent research considers deep neural network structures to generate the binary mask used to separate target sources [33].

Even if reducing the separation problem to a classification task is convenient, it comes with the drawback of bringing a characteristic and annoying *musical noise*, due to abrupt phase and amplitude transitions in the estimates. To address this issue, many researchers have focused on a *soft masking* strategy, where the TF mask is no longer binary, but rather lies in the continuous $[0\,1]$ interval. It has long been acknowledged that such strategies have the noticeable advantage of strongly reducing musical noise. Many different approaches were undertaken in the past for the purpose of building a soft TF mask. Among them, we can mention some studies where this mask is based on a divergence measure between the mixture and some model for the source: the further the observation is from the model, the smaller the weight, as in [21], [10], [26]. This approach has the advantage of requiring a model only for the target source to separate, but has the inconvenient to be unpractical if more than one source is to be extracted from the mixture.

The most popular approach to soft-masking for source separation today is based on estimating a nonnegative time-frequency energy distribution for each source, which is most commonly called a "spectrogram" in a loose acceptation. Then, the soft mask is computed for each source as the ratio of its estimated spectrogram over the sum of them all. This strategy guarantees that the sum of all soft masks equals 1 for each TF bin, so that the sum of all estimated sources is identical to the mixture, which is a desirable property. For the purpose of estimating those spectrograms, it is typically assumed that they simply add up to yield the observable spectrogram of the mixture, notwithstanding destructive interferences. Given some assumptions on how those spectrograms should look like, such as a specific parametric form [25] or local regularities [20], [22], estimation is performed as a latent variable decomposition of the spectrogram of the mixture.

It has long been acknowledged [4], [3], [5], [18] that when the spectrogram is understood as an estimate of the time-varying Power Spectral Density (PSD) of the source, this weighting strategy is theoretically justified as an optimal Wiener filtering performed independently in each frame. This filter provides the Minimun Mean Squared Error (MMSE) linear estimator of the sources given the mixture. Furthermore, theory does suggest that the PSDs of uncorrelated wide-sense stationary (WSS) processes do add up to yield the PSD of their sum [18]. For all this framework to hold,

This work was partly supported under the research programme EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

the spectrograms to be used must hence be estimates of PSDs, i.e. *squared* modulus of STFTs. We should emphasize here that this acceptation is actually the original and only rigorous one.

However, much research undertaken in the recent years has commonly understood the term "spectrogram" with a different meaning. Instead of seeing it as an estimate of the PSD, many researchers have used the word "spectrogram" to denote the modulus of the STFT raised to some arbitrary exponent $\alpha \in]0\,2]$ (see [29], [11], [14], [30]). Choosing $\alpha = 1$ is common. In the sequel, the term α-spectrogram will be used for clarity to denote this wider acceptation of the word. Just like in the WSS case with $\alpha = 2$, it is then typically assumed that the α-spectrograms of the sources add up to form the α-spectrogram of the mixture, and soft masks are derived in the same way as for the Wiener filter. Experience shows that such a procedure *does* often lead to improved performance. However, no theoretical foundation was available to explain and support this approach: to the best of our knowledge, both additivity of the α-spectrograms and soft-masking filtering are only justified theoretically for $\alpha = 2$.

In this paper, we show that using general α-spectrograms for sources modeling and separation is the optimal procedure if the sources are not understood as WSS processes, but rather as *locally stationary stable harmonizable* processes [28], α-harmonizable processes in short. Note that for $\alpha = 2$, such processes coincide with Gaussian processes [18]. They fall under the umbrella of α-stable distributions [24], [28]. Several studies demonstrated that those distributions are often better models for audio signals than the Gaussian distribution, due to their ability to handle very large deviations from the mean, which is important for such impulsive phenomena as music or sound signals in general that exhibit a large dynamic range [16], [12]. Whereas some papers focused on the separation of independent and identically distributed (i.i.d.) α-stable random variables [16], no study so far considered the separation of locally stationary and harmonizable stable processes. As we show, they provide the exact probabilistic framework needed to assume additivity of α-spectrograms as well as a justification for the design of the corresponding soft-masks.

This paper is structured as follows. In section II, we study the empirical validity of the additivity assumption for α-spectrograms. In section III, we quickly introduce α-harmonizable processes and show how they can be separated using soft masking strategies. In section IV, we compare the music separation performance of this stable harmonizable model as a function of the exponent α. Finally, we draw some tracks for future research as a conclusion.

II. ADDITIVITY OF α-SPECTROGRAMS

II-A. Notations and background

Let $\tilde{x}(t)$ be the audio signal to be separated, which is assumed regularly sampled. In typical audio applications, it is the waveform of the single channel song to be unmixed and for this reason, $\tilde{x}$ is called the *mixture* in the following. The mixture is assumed to be the simple sum of J underlying signals $\tilde{s}_j(t)$ called sources, that correspond to the individual waveforms of the different instruments playing in the mixture, such as voice, bass, guitar, percussions, etc.

In typical source separation procedures, the mixture is processed so as to compute its STFT denoted $x(f,n)$, where f is a frequency index and n is a frame index. x is thus a $N_f \times N_n$ matrix, where N_f is the total number of frequency bands[1] and N_n the total number of time frames. (f,n) is called a TF bin. For music source separation, experience shows that having frames approximately 80ms long with 80% overlap yields good results. Since the STFT is a linear transform, the simple mixing model we choose leads to:

$$\forall (f,n), x(f,n) = \sum_{j=1}^{J} s_j(f,n),$$

[1]Since $\tilde{x}$ is a real signal in audio, its spectrum is Hermitian. We assume that the redundant information in the Fourier transform of each frame has been discarded.

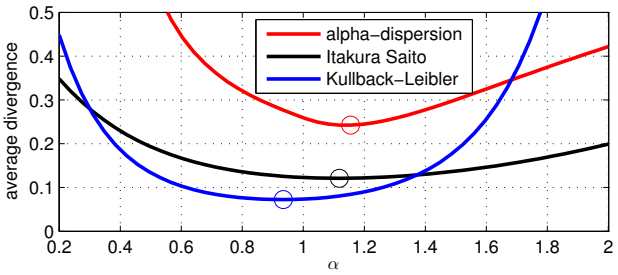


Fig. 1. Average L_α, Itakura-Saito and Kullback-Leibler divergences between the sum of the α-spectrograms of the sources and the α-spectrogram of the mixture, as a function of α. Minimal values are marked with a circle.

where s_j is the STFT of source j. For convenience, the modulus of the STFT is denoted p in the following[2]:

$$p(f,n) \triangleq |x(f,n)|.$$

Throughout this paper, the α-spectrogram p^α is defined as $p^\alpha(f,n) \triangleq p(f,n)^\alpha$. Similarly, p_j^α corresponds to the α-spectrogram of source j. As we see, the 2-spectrogram is the *power* spectrogram, which is the estimate of the PSD.

Most audio source separation methods can be understood as assuming that we basically have:

$$\forall (f,n), p^\alpha(f,n) \approx \sum_{j=1}^{J} p_j^\alpha(f,n). \quad (1)$$

As seen above, this assumption is justified theoretically when $\alpha = 2$ if we assume that the sources are locally stationary Gaussian processes [18]. For any other $\alpha \in]0\,2]$, no such probabilistic framework is available even though (1) is often assumed [29], [11], [14], [30].

II-B. Experimental study

The objective of this section is to study the validity of the additivity assumption (1) for α-spectrograms, as a function of α. To this purpose, we consider the 8 complete songs of different musical genres found in the QUASI database[3], for which the constitutive sources are available. For a set of 50 α values ranging from 0.2 to 2, we computed the *α-dispersion* between the mixture α-spectrogram and the sum of the α-spectrograms of the sources:

$$L_\alpha(f,n) = \left| p^\alpha(f,n) - \sum_{j=1}^{J} p_j^\alpha(f,n) \right|^{1/\alpha}, \quad (2)$$

as well as the popular Itakura-Saito (IS) and Kullback-Leibler (KL) divergences, commonly used in audio source separation [9], [7]. Then, the average of each divergence over all songs and all TF bins was computed, as a function of α. The results are displayed in Fig. 1.

II-C. Discussion

As can be noticed in Fig. 1, the additivity assumption (1) is not equally valid for all α. On the contrary, we clearly see that a value $\alpha \approx 1$ is much more empirically appropriate than the value $\alpha = 2$, for all divergences considered.

[2]$\triangleq$ denotes a definition.

[3]www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

This result has already been noticed, e.g. in [13], [17], and demonstrates that assuming additivity of the power spectrograms, even if justified theoretically under Gaussian assumptions, is not mostly appropriate. On the contrary, assuming additivity of the moduli p_j of the STFT of the sources for audio processing, as in most Probabilistic Latent Component Analysis studies (PLCA, see [30] and references therein), is indeed a good idea[4].

However, this empirical fact does raise an important question. When estimates of the α-spectrograms p_j^α of the sources have been obtained by any appropriate method, the estimation of the STFT of source j is then typically achieved through:

$$\hat{s}_j(f,n) = \frac{p_j^\alpha(f,n)}{\sum_{j'} p_{j'}^\alpha(f,n)} x(f,n), \tag{3}$$

which we call an α-Wiener filter in the following. Is this procedure any good and does it come with any flavor of optimality? If so, in which sense? The current lack of a probabilistic model justifying (1) for $\alpha \neq 2$ also prevented answering these questions so far. As we now show, assuming that each source is a locally stationary and α-stable harmonizable process naturally leads to (1) and, for $0 < \alpha \leq 2$, establishes (3) as the conditional expectation of $s_j(f,n)$ given $x(f,n)$, thus providing a theoretical understanding for the validity of the procedure.

III. α-HARMONIZABLE PROCESSES

We define an α-harmonizable process as a process that can locally be approximated as a stationary harmonizable α-stable process. In practice, the audio is split into overlapping frames, which are then assumed independent and each one of them is assumed stationary α-stable harmonizable.

In this section, we briefly present stationary α-stable harmonizable processes, which have been the topic of much research since the 70s and are particular cases of α-stable processes [23], [28], [24], [31], [12]. Due to space constraints, only some important facts which are of interest in our study are recalled here and the interested reader is referred to the very thorough overview of α-stable processes given in [28] and references therein for a more comprehensive treatment.

III-A. Symmetric α-stable distributions and processes

Let v be a random vector of dimension $T \times 1$. We say that v is strictly stable if for any positive numbers A and B, there is a positive number C such that

$$Av^{(1)} + Bv^{(2)} \overset{d}{=} Cv, \tag{4}$$

where $v^{(1)}$ and $v^{(2)}$ are independent copies of v and $\overset{d}{=}$ denotes equality in distribution. It can be shown [28, p. 58] that for any random vector v satisfying (4), there is one constant $\alpha \in]0\,2]$ called the *characteristic exponent* such that C in (4) is given by:

$$C = (A^\alpha + B^\alpha)^{1/\alpha}.$$

We then say that v is α-stable. If v and $-v$ furthermore have the same distribution, v is called symmetric α-stable, abbreviated as SαS. An important result is that the simple property (4) of an α-stable random vector permits to derive its characteristic function. No expression for the α-stable probability density functions is available in general, but only for $\alpha = 2$ and $\alpha = 1$, that respectively coincide with the Gaussian and Cauchy distributions.

α-stable distributions have an important number of desirable properties. One of the most famous is their ability to model data with very large deviations, making them a practical model for impulsive data in the field of robust signal processing [24]. In practice, the closest α is to 0, the heavier are the tails of an α-stable distribution. In a source separation context, the *stability* property (4) is fundamental. It basically means that provided the sources are modeled as α-stable, so will be their mixture.

We say that a collection $\{\tilde{z}(t)\}_t$ of random variables is an α-stable random *process* if the vector $\tilde{z}_T \triangleq [\tilde{z}(t_1), \ldots, \tilde{z}(t_T)]^\top$ (where $^\top$ denotes transposition) is α-stable for any choice and any number of sample positions $t_1, \ldots, t_T$.

[4]Remarkably, Fig. 1 also suggests to use KL rather than IS for $\alpha = 1$, and IS rather than KL for $\alpha = 2$, as done in the literature.

III-B. Isotropic complex SαS random variables

Because it will be useful in the sequel, we mention here that a complex random variable (r.v.) $z = v_1 + iv_2$ is called SαS if the random vector $\left[v_1^\top\ v_2^\top\right]^\top$ is SαS. A particular case of interest in our context is the special case where a complex SαS r.v. z is isotropic, or circular, abbreviated $S\alpha S_c$, meaning that:

$$\forall \theta \in [0\,2\pi[\,, \exp(i\theta)\, z \overset{d}{=} z.$$

It can be shown that in the Gaussian case $\alpha = 2$ this is equivalent to v_1 and v_2 being independent and identically distributed (i.i.d.) Gaussian r.v., whereas for the case $\alpha < 2$, isotropy leads to the particular characteristic function [28, p. 85]:

$$z = v_1 + iv_2 \sim S\alpha S_c \Leftrightarrow \mathbb{E}\left[\exp\left(i\left(\theta_1 v_1 + \theta_2 v_2\right)\right)\right] = \exp\left(-\sigma^\alpha |\boldsymbol{\theta}|^\alpha\right), \tag{5}$$

where $|\boldsymbol{\theta}|$ is the Euclidean norm of the vector $[\theta_1\,\theta_2]$, and $\sigma > 0$ is a scale parameter[5]. The real and imaginary parts of an isotropic complex SαS r.v. are *not* independent in general. As can be seen, the isotropic complex SαS distribution is only parameterized by the scale parameter σ. For convenience, we denote it $S\alpha S_c(\sigma^\alpha)$. We trivially have:

$$z_1 \sim S\alpha S_c(\sigma_1^\alpha) \text{ and } z_2 \sim S\alpha S_c(\sigma_2^\alpha),\ z_1 \text{ and } z_2 \text{ independent} \Rightarrow z_1 + z_2 \sim S\alpha S_c(\sigma_1^\alpha + \sigma_2^\alpha). \tag{6}$$

III-C. Stationary harmonizable α-stable processes

An harmonizable process $\tilde{z}(t)$ is defined as the inverse Fourier transform of a complex random measure $z(\omega)$ with independent increments:

$$\tilde{z}(t) = \int_{-\infty}^{\infty} \exp(i\omega t)\, z(\omega)\, d\omega. \tag{7}$$

In expression (7), the r.v. $z(\omega)$ may be understood as the spectrum of $\tilde{z}$, taken at angular frequency ω. Stating that z has independent increments basically means that all frequencies of the spectrum of $\tilde{z}$ are asymptotically independent, if the frame is long enough. It is a classical result that when $z(\omega)$ is an isotropic complex Gaussian random measure, $\tilde{z}(t)$ is furthermore stationary. Since audio signals can be considered stationary for the whole duration of each frame, assuming $z(\omega)$ to be an isotropic complex Gaussian is a popular assumption in the audio processing literature (see e.g. [18]).

However, assuming an isotropic complex Gaussian spectral measure is not the only way of guaranteeing that an harmonizable process $\tilde{z}$ is stationary. In particular, a very important result in our context [28, p. 292] is that taking z as an isotropic complex SαS random measure is equivalent to having $\tilde{z}$ being both a stationary and an SαS random process, which is the natural extension of the Gaussian case to $\alpha < 2$. We then model $z(\omega) \sim S\alpha S_c(\sigma_z^\alpha(\omega))$, where σ_z^α is called the fractional power spectral density of $\tilde{z}$ [31], abbreviated α-PSD in the following.

[5]Since we only consider isotropic complex SαS r.v., we do not linger here on the topic of the so-called "spectral measure" of $\left[v_1^\top\ v_2^\top\right]^\top$, which is important for general SαS multivariate distributions [28, p. 65].

The main interest of the α-harmonizable model is to account for signals that both include large deviations and are stationary. It is thus interesting for audio signals, because they are stationary on short time-frames and often feature large dynamic ranges.

III-D. Separation

Let the J source waveforms $\tilde{s}_1, \ldots, \tilde{s}_J$ defined in section II be modeled as independent α-harmonizable processes. Due to the stability property (4), their mixture $\tilde{x}$ is also α-harmonizable and using (6), we have:

$$x\left(f,n\right) \sim S\alpha S_c\left(\sum_{j=1}^{J} \sigma_j^\alpha\left(f,n\right)\right),$$

where σ_j^α is the α-PSD of source j. Since the α-spectrogram p_j^α defined in section II-A is an estimate of the α-PSD [6], we see that the α-harmonizable model indeed leads to the additivity assumption (1) over the α-spectrograms of the sources.

Now, given $x\left(f,n\right)$ and assuming the α-PSD σ_j^α of the sources are known, is there a way to estimate $s_j\left(f,n\right)$ in order to proceed to source separation? Interestingly, the answer is yes. If $0 < \alpha \leq 2$, and considering that (i) $x\left(f,n\right)$ is the sum of J independent $S\alpha S_c$ r.v. $s_j\left(f,n\right)$ and that (ii) $x\left(f,n\right)$ and $s_j\left(f,n\right)$ are jointly SαS, we have[7]:

$$\mathbb{E}\left[s_j\left(f,n\right) \mid x\left(f,n\right), \left\{\sigma_j^\alpha\right\}_j\right] = \frac{\sigma_j^\alpha\left(f,n\right)}{\sum_{j'} \sigma_{j'}^\alpha\left(f,n\right)} x\left(f,n\right). \quad (8)$$

Equation (3) can thus be interpreted as a practical estimate $\hat{s}_j(f,n)$ of $s_j(f,n)$ given $x(f,n)$, where the α-PSD σ_j^α in equation (8) has been replaced by its estimate p_j^α. We can conclude that for $0 < \alpha \leq 2$, the α-Wiener filter (3) corresponds to estimating the separated sources as their conditional expectation given the mixture x under an α-harmonizable model.

IV. EVALUATION

IV-A. Data and metrics

For evaluating the performance of the proposed α-Wiener filter for source separation, we processed the 8 songs of the QUASI database in the following way:

First, the α-spectrograms p_j^α of the true sources were computed. Then, separation was performed through (3) to obtain the best possible estimates $\hat{s}_j$ under an α-harmonizable model. After this, the resulting waveforms were obtained through an inverse STFT.

For evaluation, all separated sources were split into 30s excerpts, yielding a total of 182 separated source excerpts. The Perceptual Similarity Measure (PSM, from PEMO-Q [15]) was finally used to compare the estimated sources with the true ones, on all the excerpts and for 19 values of α between 0.2 and 2. The PSM lies between 0 (mediocre) to 1 (identical) and is frequently used in assessing audio quality. Results are displayed in Fig. 2.

IV-B. Discussion

As can be noticed in Fig. 2, the α-Wiener filter yields approximately the same performance for $\alpha \in [1, 2]$. This justifies both common practice in the source separation community and the α-harmonizable model that establishes it on solid theoretical grounds for $0 < \alpha \leq 2$. That said, two further remarks may be done here.

[6] Actually, p_j^α as defined in section II should be multiplied by a constant depending only on α, in order to get an asymptotically unbiased estimate of σ_j^α. Even so, it is important to note that this constant would vanish in equation (3). In [31], an asymptotically unbiased and consistent estimator of σ_j^α is proposed, which additionally involves a stage of spectral smoothing.

[7] The proof of this result is available in [1]. It is the natural extension of [28, th. 4.1.2 p. 175] to the isotropic complex $S\alpha S_c$ case, and to the whole range $\alpha \in]0, 2]$.

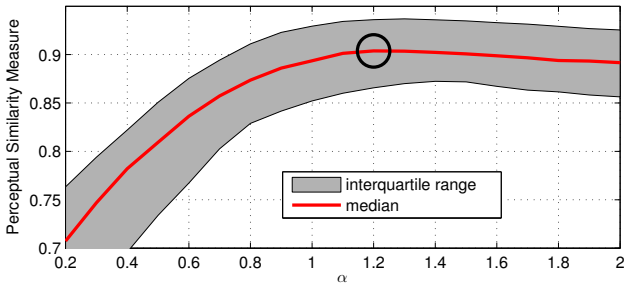


Fig. 2. Distribution of the Perceptual Similarity Measure between the true sources and those obtained by the α-Wiener filter (3), as a function of α. $\alpha = 2$ corresponds to classical Wiener filtering. The best performance is marked with a circle.

First, we see that choosing an α-harmonizable model with $\alpha < 2$ *does* improve the separation performance. In particular, the classical 2-Wiener filter is outperformed in our experiments by an α-Wiener filter with $\alpha \approx 1.2$, even if the improvement is only of a few percents.

Second, these scores correspond to the oracle performance of the method, i.e. when the true α-spectrograms of the sources are known. In real applications, they need to be estimated from the mixture and the additivity assumption (1) is critical for this purpose. Since we saw in section II that (1) is much better verified when $\alpha \approx 1$ than in the Gaussian case, we see that the α-harmonizable model may be advantageous in practice, because it is the only one we know of that justifies both this popular assumption and the resulting filtering procedure (3).

V. CONCLUSION

In a single channel audio source separation context, it is often convenient to assume some linear relationship between the spectrogram of the mixture and the spectrograms of the sources. Identifying the spectrograms of the sources is indeed important to devise soft TF masks used for separation.

When we model the sources as independent and locally wide-sense stationary processes, we have recalled that this assumption is valid for power spectrograms. In that case, a natural TF mask is the classical Wiener filter.

However and as we empirically showed here, assuming the power spectrograms of the sources to add up to form the power spectrogram of the mixture is generally a rough assumption for real audio signals. After introducing the α-spectrogram as the magnitude of the STFT raised to the power $\alpha \in]0\,2]$, we demonstrated that the additivity assumption rather holds for α-spectrograms for some $\alpha < 2$. This fact has already been pointed out by some studies in the dedicated literature.

In this paper, we have modeled the sources as locally stationary α-stable harmonizable processes, abbreviated α-harmonizable, and showed that this naturally leads to the additivity of their α-spectrograms. Furthermore, that probabilistic framework does yield a natural way of separating such signals through a soft TF mask which is analogous to the Wiener filter.

This study could be extended in two main and important directions. First, the case of multichannel mixtures is important for audio processing, because audio signals often come in several channels, as in stereophonic music. Second, this paper was only concerned with the oracle performance of the separation of stationary α-harmonizable processes, i.e. assuming that the true α-spectrograms were known. An interesting question concerns the implications of this model with respect to the blind estimation of the α-spectrograms of the sources when only the mixture is available.

VI. REFERENCES

[1] R. Badeau and A. Liutkus. Proof of Wiener-like linear regression of isotropic complex symmetric alpha-stable random variables. Technical report, September 2014.

[2] D. Barry, B. Lawlor, and E. Coyle. Real-time sound source separation using azimuth discrimination and resynthesis. In *117th Audio Engineering Society (AES) Convention*, San Francisco, CA, USA, October 2004.

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

[4] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. In *IEEE International Conference Acoustics Speech Signal Processing (ICASSP)*, pages 613–616, Hong-Kong, April 2003.

[5] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for time-frequency energy distributions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 151–154, New Paltz, NY, USA, October 2007.

[6] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[8] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 78–81, New Paltz, NY, USA, Oct. 2005.

[9] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Irish Signals and Systems Conference (ISSC)*, Galway, Ireland, June 2008.

[10] D. Fitzgerald and R. Jaiswal. On the use of masking filters in sound source separation. In *International Conference on Digital Audio Effects, (DAFx-12)*, York, UK, September 2012.

[11] J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders. Source separation by score synthesis. In *International Computer Music Conference (ICMC)*, New York, NY, USA, June 2010.

[12] P. Georgiou, P. Tsakalides, and C. Kyriakakis. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE Transactions on Multimedia*, 1(3):291–301, September 1999.

[13] R. Hennequin. *Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale*. PhD thesis, Telecom ParisTech, Paris, France, December 2011.

[14] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, Kyoto, Japan, March 2012.

[15] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902 –1911, November 2006.

[16] P. Kidmose. *Blind separation of heavy tail signals*. PhD thesis, Technical University of Denmark, Lyngby, Denmark, 2001.

[17] B. King, C. Févotte, and P. Smaragdis. Optimal cost function and magnitude power for nmf-based speech separation and music interpolation. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[18] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[19] A. Liutkus, J-L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, Paris, France, July 2013.

[20] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16):4298–4310, Aug 2014.

[21] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56, Kyoto, Japan, March 2012.

[22] A. Liutkus, Z. Rafii, B. Pardo, D. Fitzgerald, and L. Daudet. Kernel Spectrogram models for source separation. In *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[23] G. Miller. Properties of certain symmetric stable distributions. *Journal of Multivariate Analysis*, 8(3):346 – 360, 1978.

[24] C. Nikias and M. Shao. *Signal processing with alpha-stable distributions and applications*. Wiley-Interscience, 1995.

[25] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, May 2012.

[26] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech & Language Processing*, 21(1):71–82, January 2013.

[27] C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *Computer Music Journal*, 32(1):51–59, March 2008.

[28] G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[29] P. Smaragdis. Separation by humming : User-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2009.

[30] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[31] G.A. Tsihrintzis, P. Tsakalides, and C.L. Nikias. Spectral methods for stationary harmonizable alpha-stable processes. In *European signal processing conference (EUSIPCO)*, pages 1833–1836, Rhodes, Greece, September 1998.

[32] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. Gowreesunker, D. Lutter, and N. Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, August 2012.

[33] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1381–1390, July 2013.

[34] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.

Kernel additive models for source separation

Antoine Liutkus*[1,2,3], Derry Fitzgerald[4], Zafar Rafii[5], Bryan Pardo[5], Laurent Daudet[6]

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[4]NIMBUS Centre, Cork Institute of Technology, Ireland
[5]Northwestern University, Evanston, IL, USA
[6]Institut Langevin, Paris Diderot Univ., France

***Abstract*—Source separation consists of separating a signal into additive components. It is a topic of considerable interest with many applications that has gathered much attention recently. Here, we introduce a new framework for source separation called Kernel Additive Modelling, which is based on local regression and permits efficient separation of multidimensional and/or nonnegative and/or non-regularly sampled signals. The main idea of the method is to assume that a source at some location can be estimated using its values at other locations *nearby,* where nearness is defined through a source-specific *proximity kernel*. Such a kernel provides an efficient way to account for features like periodicity, continuity, smoothness, stability over time or frequency, self-similarity, etc. In many cases, such *local dynamics* are indeed much more natural to assess than any global model such as a tensor factorization. This framework permits one to use different proximity kernels for different sources and to separate them using the iterative *kernel backfitting* algorithm we describe. As we show, kernel additive modelling generalizes many recent and efficient techniques for source separation and opens the path to creating and combining source models in a principled way. Experimental results on the separation of synthetic and audio signals demonstrate the effectiveness of the approach.**

***Index Terms*—MLR-SSEP, SSP-SSEP, MLR-GRKN, MLR-MUSI**

I. Introduction

Source separation (see [1] for a review) is a research field that has gathered much attention during the last 30 years. Its objective is to recover several unknown signals called *sources* that were mixed into observable *mixtures*. It has applications in telecommunication, audio processing, latent components analysis, biological signals processing, etc.

The objective of *blind* source separation is to estimate the sources given the mixtures only. There are many ways to formulate this problem and many different approaches have been undertaken to address this challenging task. Among them, we can mention three different paradigms that have attracted much of the attention of researchers in the field.

LD is partly supported by the DReaM project of the French Agence Nationale de la Recherche (ANR-09-CORD-006, under CONTINT program). This work is partly supported by LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under references ANR-10-LABX-24 and ANR-10-IDEX-0001-02 PSL.

One of the first efficient approaches to source separation is denoted Independent Components Analysis (ICA, see [2], [1]). It performs signal separation through the assumptions that the sources are all probabilistically independent and distributed with respect to a non-Gaussian distribution. Given these assumptions, contrast features representative of both non-Gaussianity and independence are maximized, leading to the recovery of separated signals. The main issue with these approaches is that they are hard to extend to *underdetermined* source separation, i.e. when less mixtures than sources are available. Furthermore, many signals of interest are usually poorly modelled as independent and identically distributed (i.i.d.), a common assumption in ICA.

A second set of techniques for source separation is grounded in state-space modelling. Indeed, it can be expressed in terms of an adaptive filtering problem where the hidden state of the system is composed of the sources and where the observation process leads to the resulting mixtures [3], [4], [5]. Source separation in this context can be performed through state-inference techniques. The main issue with this approach is that the sources can rarely be modelled as obeying linear dynamic models. Meanwhile, tractable nonlinear adaptive filtering is often restricted to very local dependencies. Furthermore, the computational cost of the methods has hindered their widespread use in practice. Still, some studies [6], [7], [8] have demonstrated that a state-space model is appropriate to account for the dynamics of audio spectra in many cases. Following from this, nonnegative dynamical systems were recently introduced [9] to perform efficient separation of nonnegative sources defined through a state-space model.

Finally, a third approach, which is currently the dominating paradigm for the underdetermined separation of waveforms, is the use of generalized Wiener filtering [10], [11], [12] under Gaussian assumptions [13]. In practice, it can be shown [14] that this approach reduces to decomposing the spectrograms of the mixtures into the spectrograms of the sources. The corresponding waveforms are then easily recovered. Most related methods rely on the idea that the sources are likely to exhibit some kind of spectral redundancy. This can be efficiently captured through dimension reduction methods like Nonnegative Matrix/Tensor Factorizations (NMF/NTF, [15], [16], [17], [18]).

In spite of their appealing tractability and their performance on many signals, the aforementioned ideas often have limitations. First, some sources like the human voice are hard to model with a few fixed spectral templates. Studies such as [19], [18] address this issue and introduce more sophisticated models, but they may require a careful tuning

in practice [20]. Second, these techniques typically assume that different sources are characterized by different sets of spectra, which may not be realistic in many cases, like in a string quartet for instance.

In this study, we do not attempt to decompose the sources as combinations of fixed patterns. Instead, we focus on their *regularities* to identify them from the mixtures. To motivate this, we can consider the case of musical signals. Audio sources can exhibit extremely complex spectrograms that sometimes cannot be modelled using block structures such as NTF. However, their *local dynamics* may be understood as obeying more simple rules. Auditory Scene Analysis [21] demonstrates on perceptual grounds that our ability to discriminate sources within a mixture largely depends on local features such as repetitivity, continuity or common fate. These dynamic features do not depend on any particular spectral template, but rather on local *regularities* concerning their evolution over time and frequency. If several studies have already addressed the problem of introducing regularities within the parameters of block models [22], [16], [23], [9], only a few focused on modelling the correlations within nonnegative sources [24]. In any case, these techniques can for now only account for a small number of correlations, thus strongly limiting their expressive power.

We focus on the modelling and separation of signals which are defined on arbitrary input spaces, meaning that the approach is applicable to both 1D signals (e.g. audio time-series) and multi-dimensional data. In order to model local dependencies within a source, we assume that it can locally be approximated by a parametric model such as a polynomial. If its values at some locations are not directly observable from the data, they can be estimated using its values at other locations through *local regression* [25], [26], [27], [28]. Usually, this local fitting is handled using a sliding window of adjacent locations, yielding smooth estimates. Instead, we introduce the general concept of a source-dependent *proximity kernel,* that gives the proximity of any two locations to use for local fitting. This direct generalization permits to account for signals that are not necessarily smooth, which is often the case in practice.

If we observe a mixture and want to estimate the value of one of the sources at some location, the method we describe assumes that the contribution of all other sources will average out during local regression and that separation can hence be performed in an iterative manner. In practice, we introduce a variant of the *backfitting* algorithm [29], which can use a different proximity kernel for each source. This approach is flexible enough to take prior knowledge about the dynamics of many kinds of signals into account. In the context of audio processing, we show that it encompasses a large number of recently proposed methods for source separation [30], [31], [32], [33], [34], [18] and provides an efficient way to devise new specific separation algorithms for sources that are characterized by local features, rather than by a global additive model such as NTF.

This text is organized as follows. First, we introduce kernel local parametric models for the sources in section II. Then, we consider the case of mixtures of such sources in section III and present an algorithm for their separation that we call *kernel backfitting* (KBF). In section IV, we illustrate the effectiveness of the approach for the separation of 1D mixtures contaminated by strong impulsive noise. Finally, we discuss the application of the framework to audio sources in section V and show that KAM is efficient for the separation of the vocal part in musical signals.

II. Kernel local parametric models

Throughout this paper, all signals are understood as functions $f(l)$ giving the value of the signal at any location l. For example, in the case of a time series, l will be a time position or a sample index, while $f(l) \in \mathbb{R}$ is the corresponding value for the waveform. In another setting, for image or field measurements for instance, l may be a spatial location and $f(l)$ the signal value at that position. Such a formulation permits one to handle both regularly and non-regularly sampled data in a principled way.

In this section, we present an approach to model the local dynamics of signals. Its principle is to locally approximate a signal through a parametric model. For this purpose, it is necessary to choose a parametric family of approximations but also to define the particular weighting function used for the local fitting. This weighting function, called a *proximity kernel*, may not be based on the standard Euclidean distance, but rather on some knowledge concerning the signal. This kernel local regression is an important building block of the separation procedure we present in section III.

A. Kernel local regression

Local regression [26], [28], [35] is a method that was initially introduced for smoothing scatter plots. Formally, let $\mathbb{L}$ denote an arbitrary space called input space and let us assume that our objective is to estimate a signal $f : \mathbb{L} \to \mathbb{R}$ based on N noisy observations (l_n, z_n), where $l_n \in \mathbb{L}$ is a location and $z_n \in \mathbb{R}$ is the observed value at that location. We write

$$\mathcal{D} = \{(l_n, z_n)\}_{n=1\cdots N}$$

as the set gathering the N available measurements.

The first step in local parametric modelling is to assess a relation between the signal we seek to estimate and the observations. Usually, each observation z_n is assumed to be the sum of $f(l_n)$ with some white additive noise ϵ_n:

$$z_n = f(l_n) + \epsilon_n. \tag{1}$$

More generally, we will consider that the negative log-likelihood $\mathcal{L}(z_n \mid f(l_n))$ of the observations z_n given $f(l_n)$, also called the *model cost function* in the following, is known:

$$\mathcal{L}(z_n \mid f(l_n)) = -\log p(z_n \mid f(l_n)). \tag{2}$$

This probabilistic formulation permits us to handle noise in a more flexible manner than (1). By selecting the appropriate

probability density function, noise may be modeled as additive, multiplicative, or some other relation. Throughout this paper, we suppose that all observations z_n are independent[1].

The main idea of local parametric modelling, reminiscent of Taylor series, is to consider that for a particular position $l^\star \in \mathbb{L}$, the signal f can be *locally* approximated using a member $\mathcal{F}_{\theta(l^\star)}$ of some given parametric set of functions $\mathcal{F}$. In that case, $\theta(l^\star)$ denotes the set of parameters defining the function $\mathcal{F}_{\theta(l^\star)}$. For example, $\mathcal{F}$ can be the set of all polynomials of a given order and $\theta(l^\star)$ is then a particular set of coefficients. Finally, $f(l^\star)$ is estimated as:

$$\hat{f}(l^\star) = \mathcal{F}_{\theta(l^\star)}(l^\star), \tag{3}$$

with local parameters $\theta(l^\star)$ chosen as:

$$\theta(l^\star) = \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_{\mathcal{D}}(l_n, l^\star)\, \mathcal{L}(z_n \mid \mathcal{F}_\theta(l_n)), \tag{4}$$

where $w_{\mathcal{D}}(l_n, l^\star)$ is a *known* nonnegative weight giving the importance of having a good fit $\mathcal{L}(z_n \mid \mathcal{F}_\theta(l_n))$ at l_n for estimating $f(l^\star)$. The rationale behind the weight function in (4) is that a good choice for $\theta(l^\star)$ is only required to be good locally. In the literature, having $\mathbb{L} = \mathbb{R}^d$ with some $d \in \mathbb{N}$ is common, $\mathcal{F}$ is mostly chosen as the set of linear functions $\mathcal{F}_{\alpha,\beta} = \left\{l \mapsto \alpha + \beta^\top l\right\}_{\alpha,\beta}$ and $\mathcal{L}$ as the squared error, leading to the following cost function:

$$(\alpha^\star, \beta^\star) = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{n=1}^{N} w_{\mathcal{D}}(l_n, l^\star)\left(z_n - \alpha - \beta^\top l_n\right)^2. \tag{5}$$

This is easily solved and leads to the estimate $\hat{f}(l^\star) = \alpha^\star + \beta^{\star\top} l^\star$. When $\mathcal{F}$ is chosen as the class of constant functions (having $\beta = 0$), and $\mathcal{L}$ as the squared error, (5) is solved by the Nadaraya-Watson [36], [25] estimate:

$$\hat{f}(l^\star) = \frac{\sum_{n=1}^{N} w_{\mathcal{D}}(l_n, l^\star)\, z_n}{\sum_{n=1}^{N} w_{\mathcal{D}}(l_n, l^\star)}, \tag{6}$$

which is essentially a weighted average of the observations around $l^\star$. The parametric space $\mathcal{F}$ and the penalty function $\mathcal{L}$ to use can strongly depend on the application.

The presentation above is slightly more general than what is common in the literature. First, $\mathbb{L}$ is often taken as Euclidean equipped with a norm $l \in \mathbb{L} \mapsto \|l\| \in \mathbb{R}_+$. Second, the weight function $w_{\mathcal{D}}(l_n, l^\star)$ that gives the proximity of l_n to $l^\star$ in (4) is typically given in terms of their *distance* $\|l_n - l^\star\|$, justifying the name *local regression* for the approach. Here, we purposefully did not make these assumptions, because we allow the value $f(l^\star)$ at some location $l^\star$ to depend not necessarily on its neighbours under the initial input metric, but rather on neighbours under some arbitrary metric defined by a proximity kernel $w_{\mathcal{D}}(l, l^\star)$. This further flexibility adds improved expressive power.

[1]Assuming the observations to be independent does not mean that no relationship is to be expected between them. In the additive formulation (1) for instance, it only means that the additive noises ϵ_n are independent.

B. Proximity kernels

We refer to the weight function $w_{\mathcal{D}}(l, l^\star)$ as a proximity kernel. Its output must be non-negative and should increase as the importance of using $f(l)$ to estimate $f(l^\star)$ increases. It may be implemented using a distance metric based on the location of l and $l^\star$. This leads to the kind of kernel typical in the local regression literature. We begin with an example of such kernels and then show how they can be generalized to significantly increase the power of the approach.

1) An example proximity kernel for local regression: Most existing proximity kernels $w_{\mathcal{D}}(l, l')$ found in the local regression literature are *stationary*, i.e. functions of the *distance* $\|l - l'\|$ between their operands. This can for instance be written as:

$$w_{\mathcal{D}}(l, l') = \delta\left(\frac{\|l - l'\|}{h}\right), \tag{7}$$

where h is called the *bandwidth* in this context and is chosen using the measurements $\mathcal{D}$. δ is usually taken as a smoothly decreasing function[2] of its argument like the tricube function [28]:

$$\delta(\tau) = \begin{cases} \left(1 - |\tau|^3\right)^3 & \text{for } |\tau| \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

This kind of choice for $w_{\mathcal{D}}(l, l')$ is motivated by its intuitive connection with the notion of proximity of l from l' when $\mathbb{L}$ has an Euclidean topology.

2) Generalized proximity kernels: In the present study, we show how more general proximity kernels can be used to significantly reduce the size of $\mathcal{F}$, and thus the computational complexity of the optimization problem (4). The main idea is to choose a kernel $w_{\mathcal{D}}(l, l')$ that is not necessarily related to the Euclidean distance $\|l - l'\|$ between l and l' in $\mathbb{L}$, but rather to how much we expect the parametric approximations of f at l and l' to share the same parameters. This way, estimation of the parametric model $\mathcal{F}_{\theta(l^\star)}$ to estimate $f(l^\star)$ in (4) may depend on points l that are "far" from $l^\star$ in the Euclidean distance, but for which $w_{\mathcal{D}}(l, l^\star)$ is high.

Furthermore, even if a proximity kernel $w_{\mathcal{D}}(l, l^\star)$ is a function of the locations l and $l^\star$, it must be emphasized that it may also depend on the data $\mathcal{D}$ as is notably the case in robust local regression [26]. In short, $w_{\mathcal{D}}(l, l^\star)$ is a positive quantity, which is high whenever we expect f to share the same parameters at l and $l^\star$, *in light of the data*.

3) Example: pseudo-periodic signals: In order to illustrate these ideas, consider the example given in figure 1. We assume $\mathbb{L} = \mathbb{R}$ and we suppose that f is a function from $\mathbb{R}$ to $\mathbb{R}_+$, which is known to be periodic with period T, but not necessarily smooth. To model f, one can either pick $\mathcal{F}$ as the set of all positive trigonometric polynomials of period T and choose a global fitting strategy, or simply choose $\mathcal{F}$ as the set of constant functions and, for example, define $w_{\mathcal{D}}(l, l') = 1$ if and only if $l - l' = pT$ ($p \in \mathbb{Z}$) and 0

[2]A function is usually called smooth if it is derivable. The more derivable, the smoother it is.

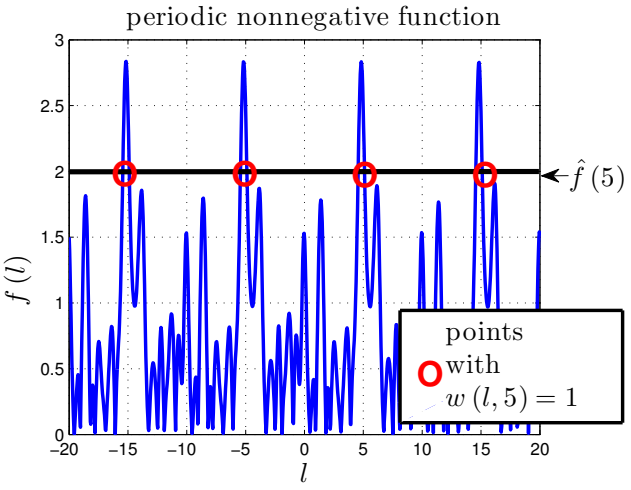


Figure 1. Nonnegative periodic function of period 10. Since f is periodic, $f(l)$ is identical to $f(l+10p)$ with $p \in \mathbb{Z}$ for any l, justifying the use of a locally constant model with a periodic proximity kernel.

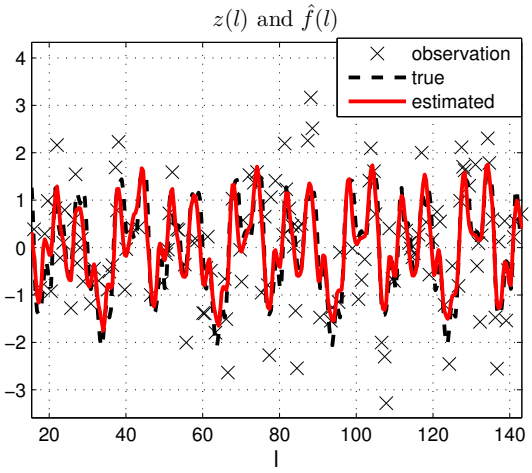


Figure 2. Kernel local regression of a non-regularly sampled and pseudo-periodic time-series mixed with white Gaussian noise of variance $\sigma^2 = 1$, using a constant model with a pseudo-periodic proximity kernel (9). Estimation is done using the Nadaraya-Watson estimate (6).

otherwise. This accounts for the fact that for any l, $f(l)$ is identical to $f(l+pT)$ with $p \in \mathbb{Z}$, because f is periodic. Further if f is assumed to be smooth, this can easily be expressed with a proximity kernel that additionally includes some proximity within each period [13]:

$$w_{\mathcal{D}}(l,l') = \exp\left(-\frac{2}{\lambda^2}\sin^2\left(\pi\frac{l-l'}{T}\right) - \frac{|l-l'|^2}{2P^2T^2}\right), \quad (9)$$

where P is a parameter indicating for how many periods the signal is known to be self-similar, while λ denotes the phase distance $\sin^2\left(\pi\frac{l-l'}{T}\right)$ required for two samples $f(l)$ and $f(l')$ to become independent.

On figure 2, we show an example of kernel local regression where the observations z_n are the sum of a non-regularly sampled locally-periodic signal f with additive white Gaussian noise of variance $\sigma^2 = 1$. f is modelled as locally constant with proximity kernel (9). Estimation is hence performed using the classical Nadaraya-Watson estimate (6) using this non-conventional proximity kernel.

The above example is representative of the expressive power of the proposed method. Instead of focusing on complex parametric spaces to globally model the signals, kernel local parametric modelling fits simpler models, but adapts the notion of proximity for the estimation of f based on some prior knowledge about its dynamics. Put otherwise, when the proximity kernel $w_{\mathcal{D}}(l,l')$ is high whenever the values of $f(l)$ and $f(l')$ are similar and negligible otherwise, simple spaces $\mathcal{F}$ of *smooth* functions such as low order polynomials may be used in (4), even if f is not smooth with respect to the canonical topology of $\mathbb{L}$. Above, $\mathcal{F}$ has been simplified from the heavily parameterized set of nonnegative periodic functions to the trivial set of constant functions.

4) Some examples of proximity kernels: Many studies in the literature [28], [35], [37], [38], [39] focus on the case of an Euclidean input space $\mathbb{L} = \mathbb{R}^d$ with $d \in \mathbb{N}$ and the proximity kernel is chosen [35], [39] as:

$$w_{\mathcal{D}}(l,l') = |H_{\mathcal{D}}|^{-1}\,\delta\left(H_{\mathcal{D}}^{-\frac{1}{2}}(l-l')\right)$$

where $H_{\mathcal{D}}$ is a $d\times d$ symmetric positive definite matrix called the *bandwidth matrix*, because it is the direct multivariate extension of the *bandwidth parameter* h of (7). δ is a smoothly decreasing function of the norm of its argument like the tricube function (8). The choice of $H_{\mathcal{D}}$ is generally data dependent and a local choice of $H_{\mathcal{D}}$ has provided good edge-preserving estimates in image processing [38], [39].

Other proximity kernels of interest include k-nearest neighbours ($k-$NN) kernels [27]. For a location $l^\star \in \mathbb{L}$, such kernels are defined as assigning a non-zero proximity value $w_{\mathcal{D}}(l,l^\star) > 0$ to at most $k \in \mathbb{N}$ locations l, called the the *nearest neighbours* of $l^\star$ and denoted $\mathcal{I}(l^\star) \in \mathbb{L}^k$. Several cases of $k-$NN kernels can be found, such as the uniform $k-$NN, that assigns the same proximity to all k nearest neighbours of $l^\star$. Some examples of $k-$NN kernels will be given in sections IV and V.

Finally, we also mention kernels that are obtained through the *embedding* $\phi_{\mathcal{D}}$ of $\mathbb{L}$ into a *feature space* Φ of arbitrary dimension through:

$$\phi_{\mathcal{D}} : l \in \mathbb{L} \mapsto \phi(l) \in \Phi. \quad (10)$$

The feature space Φ is assumed to be equipped with a dot product $\langle \phi(l), \phi(l') \rangle_\Phi$ and the proximity kernel $w_{\mathcal{D}}(l,l')$ to use can be chosen as:

$$w_{\mathcal{D}}(l,l') = \langle \phi(l), \phi(l') \rangle_\Phi. \quad (11)$$

Alternatively, the proximity kernel $w_{\mathcal{D}}(l,l')$ can be a k-NN kernel based on the distance in the feature space. This is notably the case for the *nonlocal means* method [40] for image denoising that computes the similarity $w_{\mathcal{D}}(l,l')$ of two locations based on the similarity of the observations in

their respective neighbourhoods. When the embedding (10) does not depend on $\mathcal{D}$, proximity kernels (11) have the noticeable property of always being positive definite [41], [42], which is a key element in many kernel methods. Conversely, any positive definite kernel can be shown [42] to be the dot product of some feature space. Thus, the embedding (10) may either involve the effective computation of a set of features for each l_n, or one may simply choose any positive definite function as a proximity kernel, a method which is known as the *kernel trick* in the specialized literature.

C. Comparison with Bayesian nonparametric methods

The local regression framework can be seen as a particular instance of the kernel method [42]. In our context, it has several advantages compared to other regression frameworks such as Gaussian Processes (GP, [41]). First, it allows the proximity kernel $w_{\mathcal{D}}$ to be a function of the observations $\mathcal{D}$, which is not possible through a consistent Bayesian treatment based on GP [43]. Second, it can easily permit non-negativity of the estimates given nonnegativity of the observations. Indeed, provided $w_{\mathcal{D}}(l, l') \geq 0$ and $\forall n, z_n \geq 0$, the simple Nadaraya-Watson estimate (6) is for instance also nonnegative. This feature may be important in some applications like audio processing as we show in section V. Third, contrary to the GP case, noise distributions that are not Gaussian can easily be taken into account in the local regression case. Finally, kernel local regression has the important advantage of being computationally very efficient for some choices of $w_{\mathcal{D}}$ and $\mathcal{L}$. For example, computations involved in the example displayed in figure 2 involve $\mathcal{O}\left(N^2\right)$ operations whereas consistent regression using GP would have involved the inversion of a $N \times N$ covariance matrix, requiring $\mathcal{O}\left(N^3\right)$ operations in the general case. Whenever $w_{\mathcal{D}}$ has limited support, meaning that $w_{\mathcal{D}}(\cdot, l')$ is nonzero for at most $k \ll N$ locations, complexity of local regression can drop down to $\mathcal{O}(kN)$.

Of course, this approach is not always the most appropriate for signal modelling, because its performance strongly depends on the assumption that the true underlying function can locally be approximated as lying in some given —and known— parametric set, which may not be the case. On the contrary, fully nonparametric Bayesian methods such as GP do not require such an assumption. Still, there are many cases of practical interest where a parametric model may locally be a very good fit, which is for example demonstrated by the ubiquitous use of Taylor series in science.

III. Kernel additive models

In this section, we assume that the measured signal, called the *mixture* and written x, is a noisy observation which depends on $J > 1$ functions s_j called the *sources*. A common example is the case of a sum $x_n = \sum_j s_j(l_n)$ of the sources. Our objective becomes the estimation of all J sources s_j and thus to achieve *source separation*.

The particularity of the approach we propose is that each source s_j is modelled locally using a kernel local parametric model as defined above, with its own parametric family $\mathcal{F}^{(j)}$ and proximity kernel[3] w_j. As we will illustrate in sections IV and V, this feature is central and permits the combination of different source separation methods in a principled way.

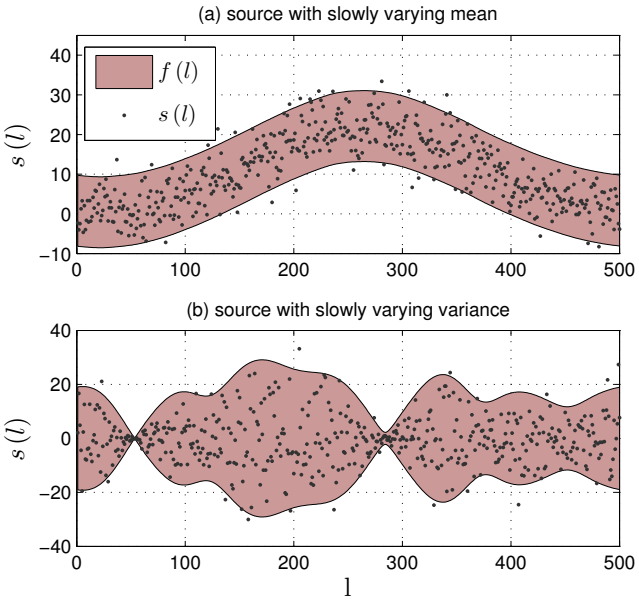


Figure 3. Two examples of sources modelled through local kernel models. (a) Independent Gaussian samples with a varying mean. (b) Independent Gaussian samples with a varying variance.

A. Formalization

Let x be a signal called the *mixture*, defined on an arbitrary input space $\mathbb{L}$ and taking values in $\mathbb{C}^I$, which means that for all $l \in \mathbb{L}$, $x(l)$ is a $I \times 1$ complex vector. The mixture depends on J underlying signals s_j called the *sources*. Each source s_j is also a function from $\mathbb{L}$ to $\mathbb{C}^I$. We assume that x is observed at N locations, yielding the observed data $\mathcal{D} = \{(l_n, x_n)\}_{n=1,\ldots,N}$.

The first step in Kernel Additive Modelling (KAM) is to model the source signals. Formally, for a source j, all samples $s_j(l)$ are assumed independent and their distribution is driven by some location-dependent hyperparameters. To illustrate this, consider the examples given in figure 3. In figure 3(a), the source samples have a slowly varying mean. In figure 3(b), they have a slowly varying variance. In both cases, their distribution depends on a slowly varying latent variable. Other parameters may also be provided, such as the variance σ^2 of all samples in figure 3(a).

In the general case, all samples $s_j(l)$ are assumed independent and their likelihood is known and given by the

[3]For conciseness, we drop the $\mathcal{D}$ subscript for the proximity kernels, but it must be emphasized that every proximity kernel considered may depend on the data.

source cost function[4]:

$$\mathcal{L}_{s_j|\Gamma_j}\left(s_j\left(l\right)\right) = -\log p\left(s_j\left(l\right) \mid \Gamma_j\left(l\right)\right), \tag{12}$$

where $\Gamma_j\left(l\right)$ is a set of source parameters that identifies which distribution to use. It is split in two parts:

$$\Gamma_j\left(l\right) = \left\{f_j\left(l\right), R_j\left(l\right)\right\}. \tag{13}$$

Among all parameters, $f_j\left(l\right)$ is called a *latent explanatory function* and is the one that undergoes local modelling, while $R_j\left(l\right)$ gathers all other parameters. For instance, the source cost function for figure 3(a) corresponds to a Gaussian distribution with mean $f\left(l\right)$ and variance $\sigma^2 \in R_j\left(l\right)$ and is thus $\mathcal{L}_{s|f}\left(s\left(l\right)\right) = \left|s\left(l\right) - f\left(l\right)\right|^2/\sigma^2$. As such, model (12) is classical and simply assesses the relation between source signals s_j and some parameters of their distributions.

The second and most noticeable step in KAM is to model the latent explanatory functions f_j. In the literature, it is common to assume that they are well described as a member $\mathcal{F}^{(j)}_{\theta_j}$ of some parametric family $\mathcal{F}^{(j)}$ of functions, such as in variance modelling with NTF [44], [18]. Here, we drop this global assumption and rather focus on a *local* model. Each latent explanatory function f_j is approximated in the vicinity of location $l^\star$ as:

$$f_j\left(l^\star\right) \approx \mathcal{F}^{(j)}_{\theta_j\left(l^\star\right)}\left(l^\star\right), \tag{14}$$

where $\mathcal{F}^{(j)}$ is a parametric family of functions and $\theta_j\left(l^\star\right)$ some location-dependent parameters. If we assume as in section II that N noisy observations $z_j\left(l_n\right)$ of $f_j\left(l_n\right)$ are available, the local parameters $\theta_j\left(l^\star\right)$ in (14) are chosen as:

$$\theta_j\left(l^\star\right) = \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} w_j\left(l_n, l^\star\right) \mathcal{L}_j\left(z_j\left(l_n\right) \mid \mathcal{F}^{(j)}_{\theta}\left(l_n\right)\right), \tag{15}$$

where w_j is the proximity kernel of source j as defined above in section II-B and $\mathcal{L}_j\left(z_j\left(l_n\right) \mid u\right)$ is a known *model cost function* for source j. It is the penalty of choosing $f_j\left(l_n\right) = u$ when its noisy observation is $z_j\left(l_n\right)$.

The final step is to assess the relation between the mixture x, the sources $s_1, \ldots, s_J$ and their parameters $\Gamma_1, \ldots, \Gamma_J$. This is done by specifying the *separation cost function*, $\mathcal{L}_{s|x,\Gamma}$, which describes our knowledge on how to perform a good separation given some set of parameters. If we adopt a probabilistic perspective, it may be understood as the negative log-likelihood of the sources given all their parameters and the mixture:

$$\begin{aligned}&\mathcal{L}_{s|x,\Gamma}\left(s_1\left(l_n\right), \ldots, s_J\left(l_n\right)\right)\\ &\quad = -\log p\left(s_1\left(l_n\right), \ldots, s_J\left(l_n\right) \mid x_n, \Gamma_1 \ldots \Gamma_J\right).\end{aligned} \tag{16}$$

In that case, it may be derived by combining the source cost function (12) with a known mixing model $p\left(x_n \mid s_1\left(l_n\right), \ldots, s_J\left(l_n\right)\right)$ through Bayes' theorem. However, some studies have demonstrated that sticking to that probabilistic interpretation may not always be advantageous and that user-defined separation cost functions may yield very good results [45]. For that reason, we retain an optimization perspective and simply assume $\mathcal{L}_{s|x,\Gamma}$ is given. Two different examples are given in sections IV and V. Furthermore, this broad definition (16) permits handling more complex mixing scenarios than simple sums, like a product of sources or non-linear mixing.

With the source, model and separation cost functions in hand, source separation amounts to computing the estimates $\hat{s}_j$, $\hat{f}_j$and $\hat{R}_j$ that jointly minimize all cost functions (12), (15) and (16). Whereas this may seem daunting to solve in full generality, we now adapt the ideas of backfitting in order to perform the estimation iteratively.

B. Kernel backfitting algorithm

The problem above has been extensively studied. In particular, if $\mathbb{L} = \mathbb{R}^J$ and in the additive case $x = \sum_j s_j$, if we assume that the sources coincide with the latent explanatory variables[5], having $s_j = f_j$ and that each source $s_j\left(l\right)$ only depends on the j^{th} coordinate of l, this problem has been extensively studied under the name of Generalized Additive Models (GAM, [29], [46]). The more general case where each function $s_j\left(l\right) = f_j\left(l\right)$ depends on a particular projection $a_j^\top l$ of the input has been considered by FRIEDMAN et al. as *projection pursuit regression* (PPR, [47]).

In this work, we propose to adapt the GAM and PPR models so that they can be used for source separation without their original assumptions that $s_j = f_j$ with each $f_j\left(l\right)$ depending on the projection of l into the real line. We instead only assume that the mixing, source and model cost functions as defined above are given, along with the parametric families $\mathcal{F}^{(j)}$ of functions and proximity kernel*s* w_j. Even if this is a generalization of both approaches, the separation algorithm we present is strongly inspired by the original *backfitting* procedure described in [47] and further studied in [29], [46] for the estimation of GAMs. Logically, we propose the term *kernel backfitting* for this algorithm.

Intuitively, the algorithm goes as follows. For a set of source estimates $\hat{s}_j$ and for each location l_n, we compute the parameters $\hat{\Gamma}_j\left(l_n\right)$ that minimize the sources cost functions $\mathcal{L}_{s_j|\Gamma_j}$without taking the model cost function into account. This leads to a set of parameters $\hat{\Gamma}_j = \left\{z_j, \hat{R}_j\right\}$, where z_j is a noisy observation of the true latent explanatory function f_j. Then, new estimates $\hat{f}_j$ are computed by kernel-smoothing z_j through kernel local regression as in section II-A, using the model cost function (15). Finally, with this new set of parameters $\left\{\hat{f}_j, \hat{R}_j\right\}$, new sources estimates are computed by minimizing the separation cost function (16). The process then repeats using those new source estimates, until a stopping criterion is reached, such as iteration number or the difference between new and old estimates. The different steps are outlined in algorithm 1.

[4]In the following, each notation $\mathcal{L}_\star\left(\cdot\right)$ denotes a cost function which depends on the location l considered. For ease of notation, this dependence on l is not made explicit.

[5]For instance, we have $s_j = f_j$ when $\sigma^2 = 0$ in figure 3(a).

Algorithm 1 Kernel backfitting (KBF). General formulation.

1) **Input**:
   - Mixture data $\mathcal{D} = \{(l_n, x_n)\}_{n=1,\dots,N}$
   - Number of source J
   - Sources (12), model (15) and separation (16) cost functions
   - Kernel models $\left\{w_j, \mathcal{F}^{(j)}\right\}_{j=1,\dots,J}$
   - Stopping criterion
2) **Initialization**
   a) $\forall (j,n), \hat{s}_j(l_n) \leftarrow x_n/J$
3) **Parameters update step**
   a) $\forall j, \left\{z_j, \hat{R}_j\right\} \leftarrow \underset{\Gamma}{\text{argmin}} \sum_n \mathcal{L}_{s_j|\Gamma_j}\left(\hat{s}_j(l_n)\right)$
   b) $\forall (j,n), \hat{f}_j(l_n) \leftarrow \mathcal{F}^{(j)}_{\theta_j(l_n)}(l_n)$, where $\theta_j(l_n)$ is estimated as in (15)
   c) $\forall (j,n), \hat{\Gamma}_j(l_n) \leftarrow \left\{\hat{f}_j(l_n), \hat{R}_j(l_n)\right\}$
4) **Separation step**
   a) $\{\hat{s}_1, \dots, \hat{s}_J\} \leftarrow \underset{s}{\text{argmin}} \mathcal{L}_{s|x,\hat{\Gamma}}$
5) If stopping criterion is not met, go to step 3
6) **Output**: sources and parameter estimates $\left\{\hat{s}_j, \hat{\Gamma}_j\right\}_{j=1\cdots J}$

Apart from its similarity with the backfitting procedure, this algorithm also coincides in some cases with the Expectation-Maximization approach (EM [48]) undertaken for underdetermined source separation, e.g. in [17], [12], [18]. This happens when the proximity kernels w_j for the sources are uniformly one $\forall (l,l'), w_j(l,l') = 1$, leading to a global fitting of θ_j in step 3b, and when the model cost function is chosen in a probabilistically coherent way with the source and separation models. For instance, it has been argued in [49], [44], [12], [14], [13], [18] that the Itakura-Saito (IS) divergence should be chosen as the model cost function (15) if the parameters θ_j are to be used for variance modelling of Gaussian random variables. The corresponding source and separation distributions are derived consequently. However, many studies have demonstrated that the use of other model cost functions such as the Kullback-Leibler (KL) divergence may provide better performance in the same context [50], [51]. This is our motivation in using user definable cost functions for the sources, models and separation steps. As we show later, this can be important in the case of strong impulsive noise in the measurements.

Finally, the KBF algorithm is close in spirit to the Denoising Source Separation framework (DSS [52]), which is limited to overdetermined source separation. Indeed, it includes a local smoothing of the latent functions z_j in step 3b to yield the updated estimates $\hat{f}_j$. Even if this smoothing actually includes arbitrary proximity kernels as defined in section II-A, the main idea remains the same: prior knowledge is used within the separation algorithm to improve estimates through some kind of *procedural* denoising operation, which permits to model sources. In a sense, the KBF algorithm 1 may be considered as a counterpart for DSS in the case of underdetermined mixtures.

IV. Toy example : robust source separation of locally constant sources

In this section, we study the separation of synthetic signals mixed with impulsive noise and show that KAM gives good performance in this context, unlike linear methods such as GP [13]. MATLAB code for these toy examples is available at the web page dedicated to this paper[6]

A. KAM formulation

To illustrate the use of KAM for source separation, assume that the observation is composed of N real measurements $x_1, \dots, x_N$ of the mixture at locations $l_1, \dots, l_N$. They are a simple sum of J sources $s_j(l_n)$:

$$x_n = \sum_{j=1}^{J} s_j(l_n). \tag{17}$$

The first step in KAM is to model each source. We will assume for now that all samples $s_j(l_n)$ from each source s_j are independent and are Gaussian distributed with mean $f_j(l_n)$ and variance σ^2 as in figure 3(a):

$$s_j(l_n) \sim \mathcal{N}\left(f_j(l_n), \sigma^2\right). \tag{18}$$

Based on the observation of a single sample $\hat{s}_j(l_n)$, the maximum likelihood estimate $z_j(l_n)$ of the mean $f_j(l_n)$, which minimizes the source cost function (12) is the trivial

$$z_j(l_n) = \hat{s}_j(l_n), \tag{19}$$

to be used in KBF at step 3a.

The second element required for KBF is to set a kernel local parametric model to each latent mean function f_j. This is achieved by specifying a parametric family $\mathcal{F}^{(j)}$ of functions, a proximity kernel w_j and a model cost function $\mathcal{L}_j$. First, f_j is simply assumed locally constant, so that (14) collapses to $f_j(l_n) = \theta_j(l_n)$. Second, we choose a nearest neighbours proximity kernel w_j as described in section II-B. Its particular shape depends on the source and is described below. Finally, the model cost function is arbitrarily chosen as the absolute deviation:

$$\mathcal{L}_j(z_j(l_n) \mid f_j(l_n)) = |z_j(l_n) - f_j(l_n)|. \tag{20}$$

This choice is motivated by the fact that $z_j(l_n)$ is likely to be a very poor estimate of $f_j(l_n)$, because it is based on a single observation. The absolute deviation is widely known to be more robust to the presence of outliers in the data. It is readily shown [53] that minimization of the binary-weighted model cost function (20) is achieved by the median value of $\{z_j(l) \mid l \in \mathcal{I}_j(l_n)\}$, denoted:

$$\hat{f}_j(l_n) = \text{median}\{z_j(l) \mid l \in \mathcal{I}_j(l_n)\}, \tag{21}$$

to be used in KBF at step 3b.

[6] www.loria.fr/~aliutkus/kam/

Finally, the last step in KAM is to specify the separation cost function. Provided all latent mean functions f_j are known, the posterior distribution $p\left(s_1, \ldots, s_J \mid x, \Gamma\right)$ of the sources given the mixture is Gaussian. Their a posteriori mean thus minimizes the separation cost function and may hence be used during KBF at step 4a:

$$\hat{s}_j\left(l_n\right) = f_j\left(l_n\right) + \frac{x_n - \sum_{j'=1}^{J} f_{j'}\left(l_n\right)}{J}. \qquad (22)$$

Using expressions (19), (21) and (22) in the corresponding steps of the KBF algorithm, separation of all sources and estimation of the latent mean functions f_j can be achieved. If all proximity kernels have limited support $k \ll N$, complexity of the KBF algorithm is $\mathcal{O}\left(kN\right)$.

B. GP formulation

The same problem can be handled using Gaussian Processes (GP) for source separation [13]. Combining (17) and (18), we get:

$$x_n = \sum_{j=1}^{J} f_j\left(l_n\right) + \epsilon_n, \qquad (23)$$

where all ϵ_n are independent and identically distributed (i.i.d.) with respect to a Gaussian distribution. For reasons that will become clear soon, their common variance $J\sigma^2$ is rewritten as $2\gamma^2$, where γ^2 is called the noise *power*. Provided each source is modelled as a GP with known mean and covariance functions (see [41], [13]), their separation is readily achieved as a particular case of GP regression:

$$\left[\hat{f}_j\left(l_1\right) \ldots \hat{f}_j\left(l_n\right)\right]^\top = K_j \left[\sum_{j'=1}^{J} K_{j'} + 2\gamma^2 \mathbf{I}_N\right]^{-1} \left[x_1, \ldots, x_N\right]^\top, \qquad (24)$$

where $\cdot^\top$ denotes conjugation, K_j is a known $N \times N$ covariance matrix of $\left[\hat{f}_j\left(l_1\right) \ldots \hat{f}_j\left(l_n\right)\right]^\top$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix. In the GP framework, prior information about each source comes as a particular choice for the covariance function, which encodes our knowledge about the regularities of f_j and permits the building of K_j. As demonstrated for instance in [13], in the case of regularly sampled signals and stationary covariance functions, separation (24) may be achieved in $\mathcal{O}\left(N^2 \log N\right)$ operations. Many techniques for underdetermined source separation can be understood as such GP regression [13].

C. Results and discussion

In order to compare the KAM and GP frameworks for source separation, we synthesize two latent explanatory functions f_1 and f_2 as realizations of two GP whose covariance functions are known. In other words, the covariance matrices K_1 and K_2 in (24) are assumed known, which is the ideal case for the GP approach. x is then built as in (23), and separation is performed with both KAM and GP. The metric considered for performance evaluation is the signal to error ratio SER_j:

$$SER_j = 10 \log_{10} \frac{\left\|f_j\right\|_2}{\left\|f_j - \hat{f}_j\right\|_2},$$

which is higher for better separation. 50 independent trials of this experiment are performed and results are reported as the median and interquartile range of all SER_j.

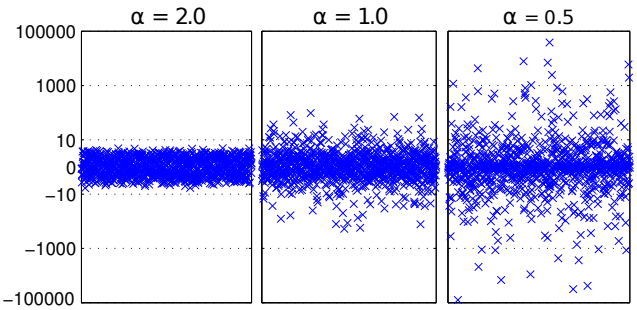


Figure 4. Independent and identically distributed symmetric and centered α-stable noise with power 1 and different α. Plots are semi-logarithmic. The case $\alpha = 2$ is Gaussian and $\alpha \to 0$ features strong outliers.

Our objective in this toy-study is to test the robustness of KAM and GP to violations of the Gaussian assumption for the additive noise ϵ_n. More precisely, we check for their performance when some outliers are present among the ϵ_n. In practice, instead of taking all ϵ_n as i.i.d. Gaussian, they are drawn from a symmetric α-stable distribution of power γ^2. The family of α-stable distributions includes Gaussian ($\alpha = 2$), Cauchy ($\alpha = 1$) and Levy ($\alpha = 1/2$) distributions as special cases. Their main characteristic is that a sum of α-stable random variables remains an α-stable random variable. Their *stability* parameter $\alpha \in [0\,2]$ controls the tail of the distribution, ($\alpha \to 0$ leads to heavy tails) and its *power* γ^2 controls its spread. In figure 4, we show independent and identically distributed samples from such symmetric α-stable distributions. They have been largely studied in the field of nonlinear signal processing because they are good models for impulsive data, yielding estimates that are robust to outliers (see [53] for a review).

For many different values of α between $\alpha = 0.5$ and $\alpha = 2$, we perform separation with both KAM and GP as described above. Of course, the assumptions underlying GP separation do not hold except for $\alpha = 2$. Still, estimation can nonetheless be performed as in (24), where $2\gamma^2$ is used instead of the variance of noise. As can be noticed, none of the KAM updates (19), (21) and (22) involve the noise variance and all can hence be used as is. Periodic kernels are chosen for the fitting of f_j, with the true periods assumed known as in figure 1. For GP, the true covariance matrices K_j are used for separation, which is a stronger prior information than the periods only.

Results are displayed in figure 5 and clearly show that while KAM provides good performance for all $\alpha \in [0.5\,;\,2]$,

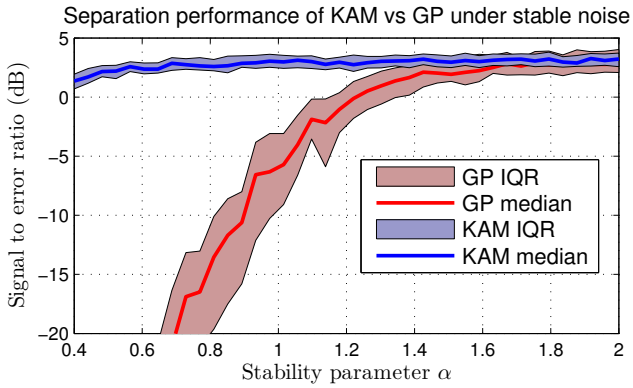


Figure 5. Performance of separation under impulsive α-stable noise of unit power as a function of α for both GP and KAM separation algorithms. 50 independent trials are considered for each α. Whereas KAM is robust to impulsive noise, the performance of GP separation is good for $\alpha \geq 1.6$ only, i.e. for non-impulsive noise. IQR stands for InterQuartile Range.

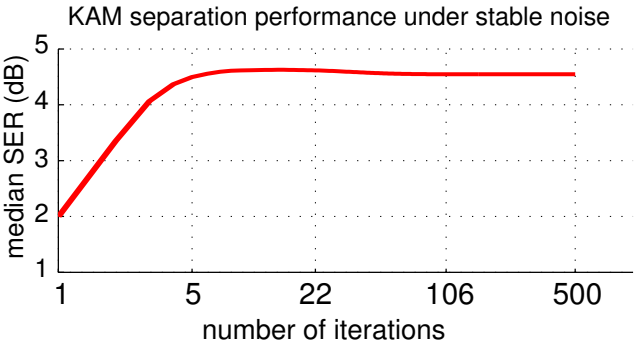


Figure 6. Median signal to error ratio for KAM separation of periodic signals from α-stable noise with $\alpha = 1$, as a function of the number of iterations, for 10 independent trials. Performance plateaus after 5 iterations.

the scores obtained by GP rapidly drop below $\alpha = 1.6$. Remarkably, even for the Gaussian case $\alpha = 2$, GP separation is not better than KAM. We can conclude that GP cannot handle outliers as well as KAM. This is an expected result, since (24) boils down to a linear combination of observations. On the contrary, separation using the KBF algorithm involves a robust estimation of f_j at step 3b, which permits excellent performance even in case of $\alpha-$stable noise. On figure 6, we show the performance of KAM as a function of the number of iterations for this example. As can be seen, performance plateaus in about 5 iterations.

Finally, we tested KAM for the separation of step-like signals from periodic oscillations under stable noise. Some illustrative results are given in figure 7. Remarkably, it is impossible to use a GP with a stationary covariance function to model such step-like signals. In KAM, the only difference from the scenario above is the use of a classical proximity kernel $\mathcal{I}_j(l) = [l-p, \dots l+p]$. This leads to a median filtering of z_j in the corresponding KBF step 3b. Separation with KAM is still of linear complexity and done in a few seconds using a standard laptop computer for $N = 5000$ observations and 10 iterations of KBF.

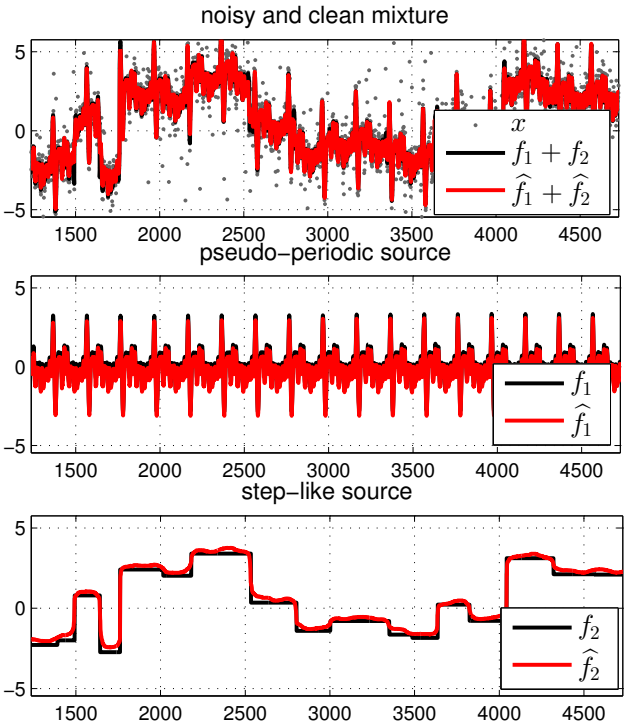


Figure 7. Example of kernel additive modelling. The noisy observed signal (top) is the sum of two sources (middle, bottom) and very adverse Cauchy noise ($\alpha = 1$). KAM permits the separation of step-like signals, for which no stationary GP model is available.

V. Application : audio source separation

In this section, we illustrate how to use KAM for a particular real-world scenario: the separation of music recordings. After some theoretical background on audio source separation, we show how to instantiate the KAM framework to devise efficient audio separation methods.

A. Separation of Gaussian processes : principles

The observed mixture consists of I audio waveforms denoted $\tilde{x}$. Each one of them is called a *channel* of the mixture. In music processing, the case $I = 2$ of stereophonic mixtures $\tilde{x}_n$ is common. The mixture $\tilde{x}$ is assumed to be the sum of J unknown signals $\{\tilde{s}_j\}_{j=1,\dots,J}$ called *sources*, that are also multichannel waveforms:

$$\tilde{x} = \sum_{j=1}^{J} \tilde{s}_j. \tag{25}$$

The Short Term Fourier Transforms (STFTs) of the J sources and of the mixture are written $\{s_j\}_{j=1\cdots J}$ and x, respectively. They all are $N_\omega \times N_t \times I$ tensors, where N_ω is the number of frequency bins and N_t the number of frames. $N = N_\omega N_t$ is the total number of Time-Frequency (TF) bins. $x(\omega,t)$ and $s_j(\omega,t)$ are $I \times 1$ vectors that gather the value of the STFT of all channels (e.g. left and right) of x and s_j at TF bin (ω,t). We denote $\mathbb{L}$ as a set of all TF bins (ω,t): $\mathbb{L} = [1 \cdots N_\omega] \times [1 \cdots N_t]$.

In the monophonic case $I = 1$, it can be shown that under local stationarity assumptions [13], all TF bins $s_j(\omega, t)$ of the STFT are independent and normally distributed. In the multichannel case, a popular related model is the Local Gaussian Model [12]. It assumes that all vectors $s_j(\omega, t)$ are independent, each one of them being distributed with respect to a multivariate centered complex Gaussian distribution:

$$\forall (\omega, t),\ s_j(\omega, t) \sim \mathcal{N}_c\left(0, f_j(\omega, t) R_j(\omega)\right), \tag{26}$$

where $f_j(\omega, t) \geq 0$ is the *power spectral density* (PSD) of source j at TF bin (ω, t). It is a nonnegative scalar that corresponds to the energy of source j at TF bin (ω, t). The *spatial covariance matrix* $R_j(f)$ is a $I \times I$ positive semidefinite matrix that encodes the covariance between the different channels of s_j at frequency ω. As shown in [12], this model generalizes the popular linear instantaneous and convolutive cases [1] and permits more flexibility in the modelling of the spatial dispersion of a source. As can be seen, the source model (26) is a multichannel extension of the heteroscedastic model depicted in figure 3(b), which includes both a latent explanatory function f_j and other parameters R_j, gathered in $\Gamma_j(\omega, t) = \{f_j(\omega, t), R_j(\omega)\}$. Being the sum of J independent random Gaussian vectors $s_j(\omega, t)$, the mixture $x(\omega, t)$ is itself distributed as:

$$\forall (\omega, t),\ x(\omega, t) \sim \mathcal{N}_c\left(0, \sum_{j=1}^{J} f_j(\omega, t) R_j(\omega)\right). \tag{27}$$

If the parameters $\Gamma_j = \{f_j, R_j\}$ are known or estimated as $\hat{f}_j$ and $\hat{R}_j$, the Minimum Mean-Squared Error (MMSE) estimates $\hat{s}_j$ of the STFTs of the sources are obtained via generalized spatial Wiener filtering [10], [11], [13], [12]:

$$\hat{s}_j(\omega, t) = \hat{f}_j(\omega, t) \hat{R}_j(\omega) \left[\sum_{j'=1}^{J} \hat{f}_{j'}(\omega, t) \hat{R}_{j'}(\omega)\right]^{-1} x(\omega, t). \tag{28}$$

The waveforms $\tilde{s}_j$ of the sources in the time domain are easily recovered by inverse STFTs.

B. Locally constant models for audio sources

Setting this in the KAM methodology, we see that (26) readily provides a source cost function while (28) permits minimization of the separation cost function. We now choose a kernel local parametric model for the PSD f_j of the sources, to be used in the KBF algorithm at step 3b. We model all PSDs f_j as locally constant and use k-NN proximity kernels as presented in section II-B. In other words, for each TF bin $(\omega, t) \in \mathbb{L}$, we specify a set of k neighbours $\mathcal{I}_j(\omega, t) \in \mathbb{L}^k$, for which the PSD has a value close to $f_j(\omega, t)$:

$$\forall l \in \mathcal{I}_j(\omega, t),\ f_j(l) \approx f_j(\omega, t).$$

Some examples of such binary proximity kernels are given below in section V-C.

With the proximity kernels in hand, the only missing part for the use of KAM is the definition of the model cost function $\mathcal{L}_j$. Just like in the toy example above in section IV, we choose the absolute deviation (20), because it is known to be less sensitive to outliers in the estimates z_j, which are numerous during convergence. Indeed, $z_j(l_n)$ is computed using I observations only and is likely to be contaminated with interferences from other sources. This leads to the following cost function to be minimized at KBF step 3b:

$$\hat{f}_j(\omega, t) = \operatorname*{argmin}_f \sum_{l \in \mathcal{I}_j(\omega, t)} |z_j(l) - f|, \tag{29}$$

which is achieved by:

$$\hat{f}_j(\omega, t) = \operatorname{median}\left(z_j(l) \mid l \in \mathcal{I}_j(\omega, t)\right). \tag{30}$$

The application of the general KBF algorithm 1 to this audio setup is summarized in algorithm 2, where $\cdot^\star$ denotes conjugate transpose and $\operatorname{tr}(\cdot)$ is the trace operator. Steps 4b and 4c of this algorithm correspond to maximum likelihood estimation of z_j and $\hat{R}_j$ given $\hat{s}_j$. The interested reader is referred to [12], [18] for further details . A noticeable feature of this algorithm is that all sources can be handled in parallel during both steps 3 and 4, permitting computationally efficient implementations. On a current desktop computer, typical total computing time is about 5 times slower than real time and the computational complexity of KBF scales linearly with track length and number of iterations.

Algorithm 2 Kernel backfitting for multichannel audio source separation with locally constant spectrogram models and k-NN proximity kernels.

1) **Input**:
   - Mixture STFT $x(\omega, t)$
   - Neighbourhoods $\mathcal{I}_j(\omega, t)$ as in figure 8.
   - Number of iterations
2) **Initialization**
   - $\forall j, \hat{f}_j(\omega, t) \leftarrow x(\omega, t)^\star x(\omega, t) / IJ$
   - $R_j(\omega) \leftarrow I \times I$ identity matrix
3) Compute estimates $\hat{s}_j$ of all sources using (28)
4) For each source j:
   a) $C_j(\omega, t) \leftarrow \hat{s}_j(\omega, t) \hat{s}_j(\omega, t)^\star$
   b) $\hat{R}_j(\omega) \leftarrow \frac{I}{N_t} \sum_t \frac{C_j(\omega, t)}{\operatorname{tr}(C_j(\omega, t))}$
   c) $z_j(\omega, t) \leftarrow \frac{1}{I} \operatorname{tr}\left(\hat{R}_j(\omega)^{-1} C_j(\omega, t)\right)$
   d) $\hat{f}_j(\omega, t) \leftarrow \operatorname{median}\left\{z_j(l) \mid l \in \mathcal{I}_j(\omega, t)\right\}$
5) For another iteration, go to step 3
6) **Output**:
   sources PSDs $\hat{f}_j$ and spatial covariance matrices $\hat{R}_j(\omega)$ to use for filtering (28).

C. Examples of kernels for audio sources

Many methods for audio source separation can be understood as instances of the framework presented above, including the many variants of REPET [31], [32], [33], [54], [34] or the median filtering approach presented in [30]. From

the point of view of KAM, those methods simply correspond to different choices for the proximity kernels of the sources.

As highlighted in [32], most of those studies rely on ad-hoc filtering approaches and are suboptimal in light of the developments above. In particular, when several local source models are provided as in [30], the estimation is performed independently for each source and no special care is taken in correctly modelling the observation as the *mixture* of the sources. As such, these techniques can be understood as performing only one iteration of the kernel backfitting procedure described in algorithm 2.

In this section, we illustrate the capacity of KAM to combine completely different approaches to source separation which use different assumptions. To this end, we present four families of proximity kernels to use with algorithm 2.

1) Models for percussive and harmonic sounds: In musical signals, percussive elements are known to be self-similar along the frequency axis, while harmonic steady sounds are self-similar along time [30]. This prior knowledge can be exploited in the KAM framework by choosing k-NN proximity kernels that are either vertical or horizontal, as depicted in figure 8(a) and 8(b) respectively. Using them in algorithm 2 leads to a generalization of the procedure presented in [30] that allows for multichannel mixtures.

2) Models for repetitive patterns: The musical accompaniment of a song may often be considered as locally repetitive. For instance, it may contain drum loops or guitar riffs. This has already been exploited for audio source separation in the REPET approach [31], [32] and is reminiscent of pioneering work by CLEVELAND et. al [55] on the separation of seasonal and trend components in time series. Here, we show that REPET fits well within the KAM framework and can be extended to account for superpositions of different repetitive patterns at different time scales.

Formally, the PSD f_j of a repeating source j is assumed to be locally periodic along time with period T_j, which means that $f_j(\omega, t)$ ought to be similar to $f_j(\omega, t + pT_j)$ with $p = -P, \ldots, P$. Following the discussion in section II-B3, this can be accounted for by choosing $\mathcal{I}_j(\omega, t) = \{(\omega, t + pT_j)\}$, as depicted in figure 8(c) for $P = 2$.

Then, the repeating part of a song can be modelled as the sum of J_a such spectrally pseudo-periodic signals. This formulation encompasses the REPET model discussed in [31], [32], [33] that is limited to 1 repeating pattern only.

For the purpose of estimating the periods T_j of all repeating sources, we use a peak detection of the average autocorrelation for all frequency bands of the spectrogram of the mixture. More sophisticated approaches may be considered to allow for non-integer periods.

3) Weak models for natural sounds: When devising models for the PSD of a voice signal, we are faced with the extraordinary diversity of sounds it may produce. In the past, many studies exploited the fact that sung melodies are often composed of harmonic parts obeying the classical source-filter model for phonation, including the renowned IMM model [19], [20]. Even if it often obtains good performance, this approach has issues with the separation of consonants and breathy voices, that do not fit well the harmonic model.

In this study, *natural sounds* such as the human voice are simply assumed to have smooth variations in their PSD, along time or along frequency, e.g. during voiced or voiceless parts, respectively. Since this assumption is rather loose and is valid for a large variety of signals, we call it a *weak* model for natural sounds. Formally, such a model considers that $f_j(l)$ and $f_j(l')$ are close whenever l and l' are close either along time or frequency. This is achieved by choosing the cross-like kernel depicted in figure 8(d).

4) NMF model within KAM: Even if the KAM approach encompasses a large number of recent methods for source separation [30], [31], [32], [33], [34], it can also be used to combine such approaches with a more classical NMF model. Some audio sources are indeed well explained as the activation over time of fixed and global patterns. To this purpose, the PSD f_j of source j may be modelled as:

$$f_j(\omega, t) = \sum_{k=1}^{K} W_j(\omega, k) H_j(k, t), \tag{31}$$

where W_j and H_j are parameters to be fitted globally. This is readily achieved in the KAM framework by setting $w_j = 1$. During step 4d of the KBF algorithm 2, median filtering is then simply replaced for such a source by a global fitting of z_j by the NMF model (31) through standard procedures. The model cost function $\mathcal{L}_j$ to use may be any divergence seen fit, such as IS or KL. Remarkably, if all sources are modelled this way and if the IS divergence is chosen, algorithm 2 coincides with the EM procedure described, e.g. in [18], [12].

D. Voice extraction performance

In our experiments, we processed 50 full-length stereo tracks from the ccMixter[7] database, featuring many different musical genres. For each track, the accompaniment was modelled as a sum of $J_a = 6$ repeating patterns along with a 2-seconds steady harmonic source. Vocals were modelled using a cross-like kernel of height 15Hz and width 20ms. Framelength is set to 90ms , with 80% overlap.

Kernel backfitting as described in algorithm 2 was applied for 6 iterations. A MATLAB implementation of KAM may be found in the companion webpage of this paper[8], along with the audio database and separation examples. We also performed vocal separation on these 50 full-length tracks with 3 techniques from the state of the art: IMM [19], RPCA [56] and REPETsim [54], [34]. Since RPCA and REPETsim do not handle stereo signals explicitly, they were applied on left and right channels independently. Once the tracks have been separated, they are split into 30s excerpts and performance is evaluated on the 350 resulting excerpts. The metric considered is the Source to Distortion Ratio (SDR) computed with the BSSeval toolkit [57], which is

[7] www.ccmixter.org

[8] www.loria.fr/~aliutkus/kam/

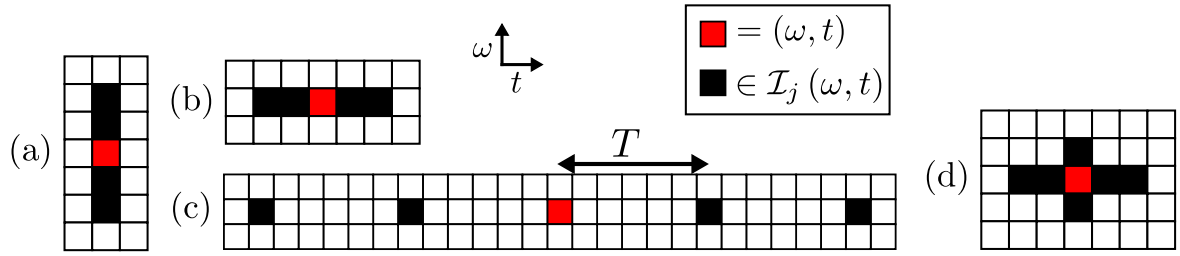


Figure 8. Some examples of k-Nearest Neigbours proximity kernels for modelling audio sources. (a) *vertical*, for percussive elements, (b) *horizontal*, for stable harmonic elements, (c) *periodic*, for repetitive elements, (d) *cross-like*, for smoothly varying power spectral densities such as vocals.

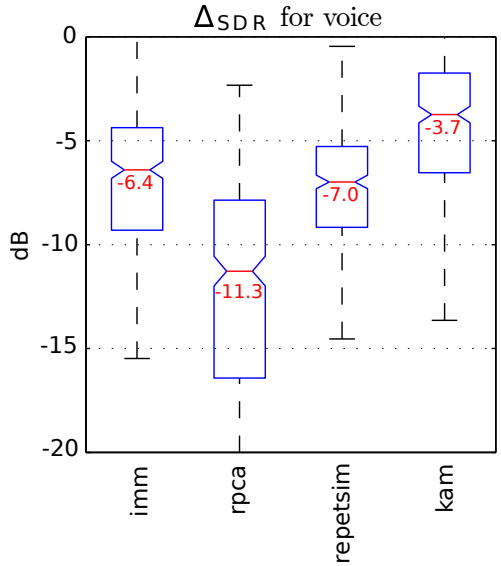


Figure 9. Δ_{SDR} scores for the separation of vocals over 50 full-length tracks, including those of the proposed KAM setup. Higher is better.

given in dB. In order to normalize separation results along the different tracks, Δ_{SDR} is given instead of SDR and indicates the loss in performance as compared to the soft-mask oracle [58]. In other words, $\Delta_{SDR} = 0$dB indicates that separation is as good as oracle Wiener filtering and the higher Δ_{SDR} is, the better the separation. Boxplots of the results are displayed on figure 9. As can be noticed, performance of the proposed KAM setup for vocal separation beats other competing methods by approximately 3dB. A multiple comparison test using a non-parametric Kruskal-Wallis analysis of variance, at the 5% confidence interval level, shows that KAM is significantly better in terms of Δ_{SDR} than all other methods. In any case, these scores only hold for the choice of proximity kernels we made in this voice/music separation task. Indeed, KAM may be used in many other settings or yield improved performance with more adequate proximity kernels and careful tuning.

VI. Conclusion

In this paper, we have proposed a new framework for source separation, where each source is modelled both *locally* and *parametrically*. Each source taken at some location is assumed to be correctly predicted using its values at other locations *nearby*. In this case, estimation can be performed using local regression.

However, not all sources are well understood by stating that neighbouring locations necessarily induce close values. This would only be true for smooth signals, which are not a good fit to the data in many cases. Instead, there may be a more sophisticated way to decide whether two locations give similar values. More generally, we introduced *proximity kernels*, which give the proximity of two points from the perspective of a source model. There are several ways of building such kernels and many methods from the literature come as special cases of this framework.

Separation of a mixture in this context can be performed using a variant of the *backfitting* algorithm, termed *kernel backfitting*, for which topological distance is replaced by source-specific proximity kernels. We showed how this Kernel Additive Modelling approach permits separation of sources that are defined through different proximity kernels.

A first feature of this method is that it is flexible enough to account for the dynamics of many kinds of signals and we indeed showed that it comes as a unifying framework for many state-of-the-art methods for source separation. Second, it yields an easy and principled way to create and combine kernel models in order to build sophisticated mixture models. Finally, the corresponding algorithms are very easy to implement in some cases and provide good performance, as demonstrated in our evaluations.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] P. Comon and C. Jutten, eds., *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, eds., *Independent Component Analysis*. Wiley and Sons, 2001.

[3] A. Cichocki, L. Zhang, and T. Rutkowski, "Blind separation and filtering using state space models," in *Proc. IEEE Int. Symp. Circuits and Systems ISCAS '99*, vol. 5, pp. 78–81, 1999.

[4] L.-Q. Zhang, A. Cichocki, and S. Amari, "Kalman filter and state-space approach to blind deconvolution," in *Proc. IEEE Signal Processing Society Workshop Neural Networks for Signal Processing X*, vol. 1, pp. 425–434, 2000.

[5] H. Valpola, A. Honkela, and J. Karhunen, "An ensemble learning approach to nonlinear dynamic blind source separation using state-space models," in *Proc. Int. Joint Conf. Neural Networks IJCNN '02*, vol. 1, pp. 460–465, 2002.

[6] L. Barrington, A. Chan, and G. Lanckriet, "Modeling music as a dynamic texture," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 602–612, Mar. 2010.

[7] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), pp. 253–256, Oct. 2011.

[8] E. Schmidt, R. Migneco, J. Scott, and Y. Kim, "Modeling instrument tones as dynamic textures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), pp. 253–256, Oct. 2011.

[9] C. Févotte, J. L. Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Vancouver, Canada), May 2013.

[10] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.

[11] A. Cemgil, P. Peeling, O. Dikmen, and S. Godsill, "Prior structures for Time-Frequency energy distributions," in *Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'07)*, (NY, USA), pp. 151–154, Oct. 2007.

[12] Q. K. N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1830–1840, July 2010.

[13] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3155 –3167, July 2011.

[14] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, pp. 2421–2456, Sep. 2011.

[15] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, Sept. 2009.

[16] O. Dikmen and A. Cemgil, "Unsupervised single-channel source separation using Bayesian NMF," in *Proc. of the 2009 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'09)*, (NY, USA), pp. 93–96, Oct. 2009.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, pp. 550–563, Mar. 2010.

[18] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[19] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, pp. 1180 –1191, oct. 2011.

[20] J. Durrieu and J. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. of International Conference on Latent Variable Analysis and Signal Separation*, (Tel-Aviv, Israel), March 12-15 2012.

[21] A. Bregman, *Auditory Scene Analysis, The perceptual Organization of Sound*. MIT Press, 1994.

[22] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE trans. on Audio, Speech and Language Proc. (TASLP)*, vol. 18(3), pp. 528–537, Mar. 2010.

[23] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with Gaussian process priors," *Computational Intelligence and Neuroscience*, vol. 2008, ID 361705.

[24] O. Dikmen and A. T. Cemgil, "Gamma markov random fields for audio source modelling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 589–601, Mar. 2010.

[25] G. Watson, "Smooth regression analysis," *Sankhya Ser.*, vol. A 26, pp. 359–372, 1964.

[26] W. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, pp. 829–836, 1979.

[27] C. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, pp. 595–620, 1977.

[28] W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, pp. 596–610, 1988.

[29] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical Science*, vol. 1, pp. 297–310, 1986.

[30] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, (Graz, Austria), Sept. 2010.

[31] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 221 –224, may 2011.

[32] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 53–56, mars 2012.

[33] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 1, pp. 71–82, 2013.

[34] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *ISMIR* (F. Gouyon, P. Herrera, L. G. Martins, and M. Müller, eds.), pp. 583–588, FEUP Edições, 2012.

[35] D. Ruppert and M. P. Wand, "Multivariate locally weighted least squares regression," *The Annals of Statistics*, vol. 22, pp. 1346–1370, 1994.

[36] E. Nadaraya, "On estimating regression," *Theor. Probab. Appl.*, vol. 9, pp. 141–142, 1964.

[37] L. Yang and R. Tschernig, "Multivariate bandwidth selection for local linear regression," *Journal Of The Royal Statistical Society Series B*, vol. 61, no. 4, pp. 793–815, 1999.

[38] I. Gijbels, R. Lambert, and P. Qiu, "Edge-preserving image denoising and estimation of discontinuous surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1075–1087, 2006.

[39] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, pp. 349–366, 2007.

[40] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, (CVPR 2005). IEEE Conference on*, vol. 2, pp. 60–65, 2005.

[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[42] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[43] J. Quiñonero-Candela, C. E. Rasmussen, and R. Herbrich, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, Dec. 2005.

[44] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. of Irish Sig. and Systems Conf. (ISSC'08)*, 2008.

[45] D. Fitzgerald and R. Jaiswal, "On the use of masking filters in sound source separation," in *Proc. of 15th International Conference on Digital Audio Effects, (DAFX12)*, 2012.

[46] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. London: Chapman & Hall, 1990.

[47] J. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, vol. 76, pp. 817–823, 1981.

[48] A. Dempster, N. Laird, and B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[49] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[50] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational Intelligence and Neuroscience special issue on Advances in Non-negative Matrix and Tensor Factorization*, vol. May, pp. 12–20, 2008.

[51] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *ICA '04: Proc. of the 8th Int. Conf. on Independent Component Analysis and Signal Separation*, 2004.
[52] J. Särelä, H. Valpola, and M. Jordan, "Denoising source separation," *Journal of Machine Learning Research*, vol. 6, no. 3, 2005.
[53] G. Arce, *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.
[54] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *23rd IET Irish Signals and Systems Conference (ISSC2012)*, (NUI Maynooth, Ireland), June 2012.
[55] R. Cleveland, W. Cleveland, J. Mcrae, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.
[56] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.
[57] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462 –1469, July 2006.
[58] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, pp. 1933 – 1950, Aug. 2007.

Antoine Liutkus received the State Engineering degree from Telecom ParisTech, France, in 2005, and the M.Sc. degree in acoustics, computer science and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005. He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and obtained his PhD in electrical engineering at Telecom ParisTech in 2012. He is currently researcher at Inria Nancy Grand Est in the speech processing team. His research interests include audio source separation and machine learning.

Derry Fitzgerald (PhD, M.A. B.Eng.) is a senior Researcher in the NIMBUS Centre at Cork Institute of Technology. He was a Stokes Lecturer in Sound Source Separation algorithms at the Audio Research Group in DIT from 2008-2013. Previous to this he worked as a post-doctoral researcher in the Dept. of Electronic Engineering at Cork Institute of Technology, having previously completed a Ph.D. and an M.A. at Dublin Institute of Technology. He has also worked as a Chemical Engineer in the pharmaceutical industry for some years. In the field of music and audio, he has also worked as a sound engineer and has written scores for theatre. He has recently utilised his sound source separation technologies to create the first ever officially released stereo mixes of several songs for the Beach Boys, including "Good Vibrations" and "I get around". His research interests are in the areas of sound source separation, tensor factorizations, and music information retrieval systems.

Zafar Rafii is a Ph.D. candidate in the department of Electrical Engineering and Computer Science at Northwestern University. He received a Master of Science in Electrical Engineering, Computer Science and Telecommunications from Ecole Nationale Sup´erieure de l'Electronique et de ses Applications (ENSEA) in France and a Master of Science in Electrical Engineering from Illinois Institute of Technology (IIT) in the US. He also worked as a research engineer at Audionamix in France and as a research intern at Gracenote in the US. His research interests are centered on audio analysis, at the intersection of signal processing, machine learning, and cognitive science.

Bryan Pardo, head of the Northwestern University Interactive Audio Lab, is an associate professor in the Northwestern University Department of Electrical Engineering and Computer Science with additional appointments in the Music Theory and Cognition Department, the Segal Design Institute and is an associate of the Northwestern Cognitive Science program. Prof. Pardo received a M. Mus. in Jazz Studies in 2001 and a Ph.D. in Computer Science in 2005, both from the University of Michigan. He has authored over 70 peer-reviewed publications. He is a past associate editor for IEEE Transactions on Audio Speech and Language Processing. He has developed speech analysis software for the Speech and Hearing department of the Ohio State University, statistical software for SPSS and worked as a machine learning researcher for General Dynamics. While finishing his doctorate, he taught in the Music Department of Madonna University. When he's not programming, writing or teaching, he performs throughout the United States on saxophone and clarinet at venues such as Albion College, the Chicago Cultural Center, the Detroit Concert of Colors, Bloomington Indiana's Lotus Festival and Tucson's Rialto Theatre.

Laurent Daudet (M'04–SM'10) studied at the Ecole Normale Supérieure in Paris, where he graduated in statistical and non-linear physics. In 2000, he received a PhD in mathematical modeling from the Université de Provence, Marseille, France. After a Marie Curie post-doctoral fellowship at the C4DM, Queen Mary University of London, UK, he worked as associate professor at UPMC (Paris 6 University) in the Musical Acoustics Lab. He is now Professor at Paris Diderot University – Paris 7, with research at the Langevin Institute for Waves and Images, where he currently holds a joint position with the Institut Universitaire de France. Laurent Daudet is author or co-author of over 150 publications (journal papers or conference proceedings) on various aspects of acoustics and audio signal processing, in particular using sparse representations.

Multichannel audio source separation with deep neural networks

Aditya Arie Nugraha, *Student Member, IEEE,* Antoine Liutkus, *Member, IEEE,* and Emmanuel Vincent, *Senior Member, IEEE*

***Abstract*—This article addresses the problem of multichannel audio source separation. We propose a framework where deep neural networks (DNNs) are used to model the source spectra and combined with the classical multichannel Gaussian model to exploit the spatial information. The parameters are estimated in an iterative expectation-maximization (EM) fashion and used to derive a multichannel Wiener filter. We present an extensive experimental study to show the impact of different design choices on the performance of the proposed technique. We consider different cost functions for the training of DNNs, namely the probabilistically motivated Itakura-Saito divergence, and also Kullback-Leibler, Cauchy, mean squared error, and phase-sensitive cost functions. We also study the number of EM iterations and the use of multiple DNNs, where each DNN aims to improve the spectra estimated by the preceding EM iteration. Finally, we present its application to a speech enhancement problem. The experimental results show the benefit of the proposed multichannel approach over a single-channel DNN-based approach and the conventional multichannel nonnegative matrix factorization based iterative EM algorithm.**

***Index Terms*—Audio source separation, speech enhancement, multichannel, deep neural network (DNN), expectation-maximization (EM).**

I. Introduction

AUDIO source separation aims to recover the signals of underlying sound sources from an observed mixture signal. Recent research on source separation can be divided into (1) speech separation, in which the speech signal is recovered from a mixture containing multiple background noise sources with possibly interfering speech; and (2) music separation, in which the singing voice and possibly certain instruments are recovered from a mixture containing multiple musical instruments. Speech separation is mainly used for speech enhancement in hearing aids or noise robust automatic speech recognition (ASR), while music separation has many interesting applications, including music editing/remixing, upmixing, music information retrieval, and karaoke [1]–[5].

Recent studies have shown that deep neural networks (DNNs) are able to model complex functions and perform well on various tasks, notably ASR [6], [7]. More recently, DNNs have been applied to single-channel speech enhancement and shown to provide a significant increase in ASR performance compared to earlier approaches based on beamforming or nonnegative matrix factorization (NMF) [8]. The DNNs typically operate on magnitude or log-magnitude spectra in the Mel domain or the short time Fourier transform (STFT) domain. Various other features have been studied in [9]–[11]. The DNNs can be used either to predict the source spectrograms [11]–[16] whose ratio yields a time-frequency mask or directly to predict a time-frequency mask [10], [17]–[21]. The estimated source signal is then obtained as the product of the input mixture signal and the estimated time-frequency mask. Various DNN architectures and training criteria have been investigated and compared [19], [21], [22]. Although the authors in [15] considered both speech and music separation, most studies focused either on speech separation [10]–[12], [14], [17]–[21] or on music separation [13], [16].

As shown in many works mentioned above, the use of DNNs for audio source separation by modeling the spectral information is extremely promising. However, a framework to exploit DNNs for multichannel audio source separation is lacking. Most of the approaches above considered single-channel source separation, where the input signal is either one of the channels of the original multichannel mixture signal or the result of delay-and-sum (DS) beamforming [19]. Efforts on exploiting multichannel data have been done by extracting multichannel features and using them to derive a single-channel mask [10], [11]. As a result, they do not fully exploit the benefits of multichannel data as achieved by multichannel filtering [1], [4].

In this article, we propose a DNN-based multichannel source separation framework where the source spectra are estimated using DNNs and used to derive a multichannel filter using an iterative EM algorithm. The framework is built upon the state-of-the-art iterative EM algorithm in [23] which integrates spatial and spectral models in a probabilistic fashion. This model was used up to some variants in [24]–[28]. We study the impact of different design choices on the performance, including the cost function used for the training of DNNs and the number of EM iterations. We also study the impact of the spatial information by varying the number of spatial parameter updates and the use of multiple DNNs to improve the spectra over the iterations. We present the application of the proposed framework to a speech enhancement problem.

This work extends our preliminary work in [29] by following the exact EM algorithm in [24], instead of its variant in [28] and by reporting extensive experiments to study the impact of different design choices not only on the speech

Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent are with Inria, Villers-lès-Nancy, F-54600, France; Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France; and CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France. E-mails: {aditya.nugraha, antoine.liutkus, emmanuel.vincent}@inria.fr.

Manuscript received XXX YY, 2016; revised XXX YY, 2016.

recognition performance, but also on the source separation performance.

The rest of this article is organized as follows. Section II describes the iterative EM algorithm for multichannel source separation, which is the basis for the proposed DNN-based iterative algorithm described in Section III. Section IV shows the effectiveness of the framework for a speech separation problem and the impact of different design choices. Finally, Section V concludes the article and presents future directions.

II. Background

In this section, we briefly describe the problem of multichannel source separation and the iterative EM algorithm in [23], [24], which is the basis for the proposed DNN-based multichannel source separation algorithm.

A. Problem formulation

Following classical source separation terminology [5], let I denote the number of channels, J the number of sources, $\mathbf{c}_j(t) \in \mathbb{R}^{I\times 1}$ the I-channel spatial image of source j, and $\mathbf{x}(t) \in \mathbb{R}^{I\times 1}$ the observed I-channel mixture signal. Both $\mathbf{c}_j(t)$ and $\mathbf{x}(t)$ are in the time domain and related by

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1}$$

Source separation aims to recover the source spatial images $\mathbf{c}_j(t)$ from the observed mixture signal $\mathbf{x}(t)$.

B. Model

Let $\mathbf{x}(f,n) \in \mathbb{C}^{I\times 1}$ and $\mathbf{c}_j(f,n) \in \mathbb{C}^{I\times 1}$ denote the short-time Fourier transform (STFT) coefficients of $\mathbf{x}(t)$ and $\mathbf{c}_j(t)$, respectively, for frequency bin f and time frame n. Also, let F be the number of frequency bins and N the number of time frames.

We assume that $\mathbf{c}_j(f,n)$ are independent of each other and follow a multivariate complex-valued zero-mean Gaussian distribution [23], [24], [27], [30]

$$\mathbf{c}_j(f,n) \sim \mathcal{N}_c\left(\mathbf{0}, v_j(f,n)\mathbf{R}_j(f)\right), \tag{2}$$

where $v_j(f,n) \in \mathbb{R}_+$ denotes the power spectral density (PSD) of source j for frequency bin f and time frame n, and $\mathbf{R}_j(f) \in \mathbb{C}^{I\times I}$ is the spatial covariance matrix of source j for frequency bin f. This $I \times I$ matrix represents spatial information by encoding the spatial position and the spatial width of the corresponding source [23]. Since the mixture $\mathbf{x}(f,n)$ is the sum of $\mathbf{c}_j(f,n)$, it is consequently distributed as

$$\mathbf{x}(f,n) \sim \mathcal{N}_c\left(\mathbf{0}, \sum_{j=1}^{J} v_j(f,n)\mathbf{R}_j(f)\right). \tag{3}$$

Given the PSDs $v_j(f,n)$ and the spatial covariance matrices $\mathbf{R}_j(f)$ of all sources, the spatial source images can be estimated in the minimum mean square error (MMSE) sense using multichannel Wiener filtering [23], [27]

$$\widehat{\mathbf{c}}_j(f,n) = \mathbf{W}_j(f,n)\mathbf{x}(f,n), \tag{4}$$

where the Wiener filter $\mathbf{W}_j(f,n)$ is given by

$$\mathbf{W}_j(f,n) = v_j(f,n)\mathbf{R}_j(f)\left(\sum_{j'=1}^{J} v_{j'}(f,n)\mathbf{R}_{j'}(f)\right)^{-1}. \tag{5}$$

Finally, the time-domain source estimates $\widehat{\mathbf{c}}_j(t)$ are recovered from $\widehat{\mathbf{c}}_j(f,n)$ by inverse STFT.

Following this formulation, source separation becomes the problem of estimating the PSD and the spatial covariance matrices of each source. This can be achieved using an EM algorithm.

C. General iterative EM framework

The general iterative EM algorithm is summarized in Algorithm 1. It can be divided into the E-step and the M-step. The estimated PSDs $v_j(f,n)$ are initialized in the *spectrogram initialization* step, for instance by computing the PSD of the mixture, while the estimated spatial covariance matrices $\mathbf{R}_j(f)$ can be initialized by $I \times I$ identity matrices. In the E-step, given the estimated parameters $v_j(f,n)$ and $\mathbf{R}_j(f)$ of each source, the source image estimates $\widehat{\mathbf{c}}_j(f,n)$ are obtained by multichannel Wiener filtering (4) and the posterior second-order raw moments of the spatial source images $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)$ are computed as

$$\begin{aligned}\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n) = \widehat{\mathbf{c}}_j(f,n)\widehat{\mathbf{c}}_j^H(f,n) \\ + \left(\mathbf{I} - \mathbf{W}_j(f,n)\right) v_j(f,n)\mathbf{R}_j(f),\end{aligned} \tag{6}$$

where $\mathbf{I}$ denotes the $I \times I$ identity matrix and $\cdot^H$ is the Hermitian transposition. In the M-step, the spatial covariance matrices $\mathbf{R}_j(f)$ are updated as

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{v_j(f,n)}\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n). \tag{7}$$

The source PSDs $v_j(f,n)$ are first estimated without constraints as

$$z_j(f,n) = \frac{1}{I}\operatorname{tr}\left(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)\right), \tag{8}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. Then, they are updated according to a given spectral model by fitting $v_j(f,n)$ from $z_j(f,n)$ in the *spectrogram fitting* step. The spectrogram initialization and the spectrogram fitting steps depend on how the spectral parameters are modeled. Spectral models used in this context may include NMF [24], which is a linear model with nonnegativity constraints, KAM [27], which relies on the local regularity of the sources, and continuity models [31]. In this study, we propose to use DNNs for this purpose.

III. DNN-based multichannel source separation

In this section, we propose a DNN-based multichannel source separation algorithm, which is based on the iterative algorithm presented in Section II. Theoretical arguments regarding the cost function for DNN training are also presented.

Algorithm 1 General iterative EM algorithm [23], [24]

Inputs:
The STFT of mixture $\mathbf{x}(f,n)$
The number of channels I
The number of sources J
The number of EM iterations L
The spectral models M_0, M_1, …, M_J

for each source j of J **do**
  Initialize the source spectrogram:
    $v_j(f,n) \leftarrow$ *spectrogram initialization*
  Initialize the source spatial covariance matrix:
    $\mathbf{R}_j(f) \leftarrow I \times I$ identity matrix
end for
for each EM iteration l of L **do**
  Compute the mixture covariance matrix:
    $\mathbf{R}_{\mathbf{x}}(f,n) \leftarrow \sum_{j=1}^{J} v_j(f,n)\,\mathbf{R}_j(f)$
  for each source j of J **do**
    Compute the Wiener filter gain:
      $\mathbf{W}_j(f,n) \leftarrow$ Eq. (5) given $v_j(f,n)$, $\mathbf{R}_j(f)$, $\mathbf{R}_{\mathbf{x}}(f,n)$
    Compute the spatial source image:
      $\widehat{\mathbf{c}}_j(f,n) \leftarrow$ Eq. (4) given $\mathbf{x}(f,n)$, $\mathbf{W}_j(f,n)$
    Compute the posterior second-order raw moment of the spatial source image:
      $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n) \leftarrow$ Eq. (6) given $v_j(f,n)$, $\mathbf{R}_j(f)$, $\mathbf{W}_j(f,n)$, $\widehat{\mathbf{c}}_j(f,n)$
    Update the source spatial covariance matrix:
      $\mathbf{R}_j(f) \leftarrow$ Eq. (7) given $v_j(f,n)$, $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)$
    Compute the uncontrained source spectrogram:
      $z_j(f,n) \leftarrow$ Eq. (8) given $\mathbf{R}_j(f)$, $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)$
    Update the source spectrogram:
      $v_j(f,n) \leftarrow$ *spectrogram fitting* given $z_j(f,n)$, M_j
  end for
end for
for each source j of J **do**
  Compute the final spatial source image:
    $\widehat{\mathbf{c}}_j(f,n) \leftarrow$ Eq. (4) given all $v_j(f,n)$, all $\mathbf{R}_j(f)$, $\mathbf{x}(f,n)$
end for

Outputs:
All spatial source images $[\widehat{\mathbf{c}}_1(f,n), \ldots, \widehat{\mathbf{c}}_J(f,n)]$

A. Algorithm

In our algorithm, DNNs are employed to model the source spectra $v_j(f,n)$. We use them to predict the source spectra instead of the time-frequency masks because our preliminary experiments showed that the performance of both approaches was similar on our dataset. Moreover, it is more convenient to integrate DNNs that estimate spectra into Algorithm 1 because the algorithm requires PSD and the power spectrum can be viewed as an estimate of the PSD [32].

A DNN is used for spectrogram initialization and one or more DNNs are used for spectrogram fitting. Let DNN_0 be the DNN used for spectrogram initialization and DNN_l the

Algorithm 2 DNN-based iterative algorithm

Inputs:
The STFT of mixture $\mathbf{x}(f,n)$
The number of channels I
The number of sources J
The number of spatial updates K
The number of EM iterations L
The DNN spectral models DNN_0, DNN_1, …, DNN_L

Compute a single-channel version of the mixture:
  $\widetilde{x}(f,n) \leftarrow$ DS beamforming given $\mathbf{x}(f,n)$
Initialize all source spectrograms simultaneously:
  $[v_1(f,n), \ldots, v_J(f,n)] \leftarrow \mathrm{DNN}_0\left(|\widetilde{x}(f,n)|\right)^2$
for each source j of J **do**
  Initialize the source spatial covariance matrix:
    $\mathbf{R}_j(f) \leftarrow I \times I$ identity matrix
end for
for each EM iteration l of L **do**
  for each spatial update k of K **do**
    Compute the mixture covariance matrix:
      $\mathbf{R}_{\mathbf{x}}(f,n) \leftarrow \sum_{j=1}^{J} v_j(f,n)\,\mathbf{R}_j(f)$
    for each source j of J **do**
      Compute the Wiener filter gain:
        $\mathbf{W}_j(f,n) \leftarrow$ Eq. (5) given $v_j(f,n)$, $\mathbf{R}_j(f)$, $\mathbf{R}_{\mathbf{x}}(f,n)$
      Compute the spatial source image:
        $\widehat{\mathbf{c}}_j(f,n) \leftarrow$ Eq. (4) given $\mathbf{x}(f,n)$, $\mathbf{W}_j(f,n)$
      Compute the posterior second-order raw moment of the spatial source image:
        $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n) \leftarrow$ Eq. (6) given $v_j(f,n)$, $\mathbf{R}_j(f)$, $\mathbf{W}_j(f,n)$, $\widehat{\mathbf{c}}_j(f,n)$
      Update the source spatial covariance matrix:
        $\mathbf{R}_j(f) \leftarrow$ Eq. (7) given $v_j(f,n)$, $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)$
    end for
  end for
  for each source j of J **do**
    Compute the uncontrained source spectrogram:
      $z_j(f,n) \leftarrow$ Eq. (8) given $\mathbf{R}_j(f)$, $\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)$
  end for
  Update all source spectrograms simultaneously:
    $[v_1(f,n), \ldots, v_J(f,n)] \leftarrow \mathrm{DNN}_l\left(\left[\sqrt{z_1(f,n)}, \ldots, \sqrt{z_J(f,n)}\right]\right)^2$
end for
for each source j of J **do**
  Compute the final spatial source image:
    $\widehat{\mathbf{c}}_j(f,n) \leftarrow$ Eq. (4) given all $v_j(f,n)$, all $\mathbf{R}_j(f)$, $\mathbf{x}(f,n)$
end for

Outputs:
All spatial source images $[\widehat{\mathbf{c}}_1(f,n), \ldots, \widehat{\mathbf{c}}_J(f,n)]$

ones used for spectrogram fitting. DNN_0 aims to estimate the source spectra simultaneously from the observed mixture. This usage of joint DNN is similar to the usage of DNNs in the context of single-channel source separation in [12], [14], [15]. Meanwhile, DNN_l aims to improve the source spectra estimated at iteration l. This usage of DNN to obtain clean spectra from noisy spectra is similar to the usage of DNNs in the context of single-channel speech enhancement in [33], [34]. Theoretically, we can train different DNNs for spectrogram fitting at different iterations. Thus, the maximum number of DNNs for spectrogram fitting is equal to the number of iterations L.

In this article, we consider magnitude STFT spectra as the input and output of DNNs. Following [19], the input and output spectra are computed from single-channel signals $\widetilde{x}(f,n)$ and $\widetilde{c}_j(f,n)$ obtained from the corresponding multichannel signals $\mathbf{x}(f,n)$ and $\mathbf{c}_j(f,n)$, respectively, by DS beamforming. All DNNs are trained with the magnitude spectra of the single-channel source images $|\widetilde{c}_j(f,n)|$ as the target.

The inputs of DNN_0 and DNN_l are denoted by $|\widetilde{x}(f,n)|$ and $\sqrt{z_j(f,n)}$, respectively. The outputs of both types of DNNs for source j, frequency bin f, and frame index n are denoted by $\sqrt{v_j(f,n)}$. DNN_0 takes the magnitude spectra $|\widetilde{x}(f,n)|$ and yields the initial magnitude spectra $\sqrt{v_j(f,n)}$ for all sources simultaneously. DNN_l takes the magnitude spectra $\sqrt{z_j(f,n)}$ of all sources and yields the improved magnitude spectra $\sqrt{v_j(f,n)}$ for all sources simultaneously.

The proposed DNN-based iterative algorithm is described in Algorithm 2.

B. Cost functions

We are interested in the use of different cost functions for training the DNNs.

1) The *Itakura-Saito (IS) divergence* [35] between the target $|\widetilde{c}_j(f,n)|$ and the estimate $\sqrt{v_j(f,n)}$ is expressed as

$$\mathcal{D}_{\text{IS}} = \frac{1}{JFN}\sum_{j,f,n}\left(\frac{|\widetilde{c}_j(f,n)|^2}{v_j(f,n)} - \log\frac{|\widetilde{c}_j(f,n)|^2}{v_j(f,n)} - 1\right). \tag{9}$$

It is a popular metric in the speech processing community because it yields signals with good perceptual quality. Moreover, it is desirable from a theoretical point of view because it results in maximum likelihood (ML) estimation of the spectra [35] and the whole Algorithm 2 then achieves ML estimation. While the IS divergence has become a popular choice for NMF-based audio source separation [35]–[37], its use as the cost function for DNN training is uncommon.

2) The *Kullback-Leibler (KL) divergence* [38] is expressed as

$$\mathcal{D}_{\text{KL}} = \frac{1}{JFN}\sum_{j,f,n}\left(|\widetilde{c}_j(f,n)|\log\frac{|\widetilde{c}_j(f,n)|}{\sqrt{v_j(f,n)}} - |\widetilde{c}_j(f,n)| + \sqrt{v_j(f,n)}\right). \tag{10}$$

It is also a popular choice for NMF-based audio source separation [35] and has been shown to be effective for DNN training [13].

3) The *Cauchy cost function* is expressed as

$$\mathcal{D}_{\text{Cau}} = \frac{1}{JFN}\sum_{j,f,n}\left(\frac{3}{2}\log\left(|\widetilde{c}_j(f,n)|^2 + v_j(f,n)\right) - \log\sqrt{v_j(f,n)}\right). \tag{11}$$

It has been proposed recently for NMF-based audio source separation and advocated as performing better than the IS divergence is some cases [39].

4) The *phase-sensitive (PS) cost function* is defined as

$$\mathcal{D}_{\text{PS}} = \frac{1}{2JFN}\sum_{j,f,n}|m_j(f,n)\widetilde{x}(f,n) - \widetilde{c}_j(f,n)|^2, \tag{12}$$

where $m_j(f,n) = v_j(f,n)/\sum_{j'} v_{j'}(f,n)$ is the single-channel Wiener filter [8], [22]. It minimizes the error in the complex-valued STFT domain, not in the magnitude STFT domain as the other cost functions considered here.

5) The *mean squared error (MSE)* [35] is expressed as

$$\mathcal{D}_{\text{MSE}} = \frac{1}{2JFN}\sum_{j,f,n}\left(|\widetilde{c}_j(f,n)| - \sqrt{v_j(f,n)}\right)^2. \tag{13}$$

It is the most widely used cost function for various optimization processes, including DNN training for regression tasks. Despite its simplicity, it works well in most cases.

IV. Experimental evaluation for speech enhancement

In this section, we present the application of the proposed framework for speech enhancement in the context of the CHiME-3 Challenge [40] and evaluate different design choices. We considered different cost functions, numbers of spatial updates, and numbers of spectral updates. We anticipated that these three parameters are important parameters for the proposed framework. Extensive experiments have been done to investigate the comparative importance of these three parameters. By presenting detailed descriptions, we want to boost the reproducibility of the experiments presented and the performance achieved in this article.

A. Task and dataset

The CHiME-3 Challenge is a speech separation and recognition challenge which considers the use of ASR for a multi-microphone tablet device. In this context, we consider two sources ($J = 2$), namely speech and noise. The challenge provides real and simulated 6-channel microphone array data in 4 varied noise settings (bus, cafe, pedestrian area, and street junction) divided into training, development, and test sets. The training set consists of 1,600 real and 7,138 simulated utterances (`tr05_real` and `tr05_simu`), the development

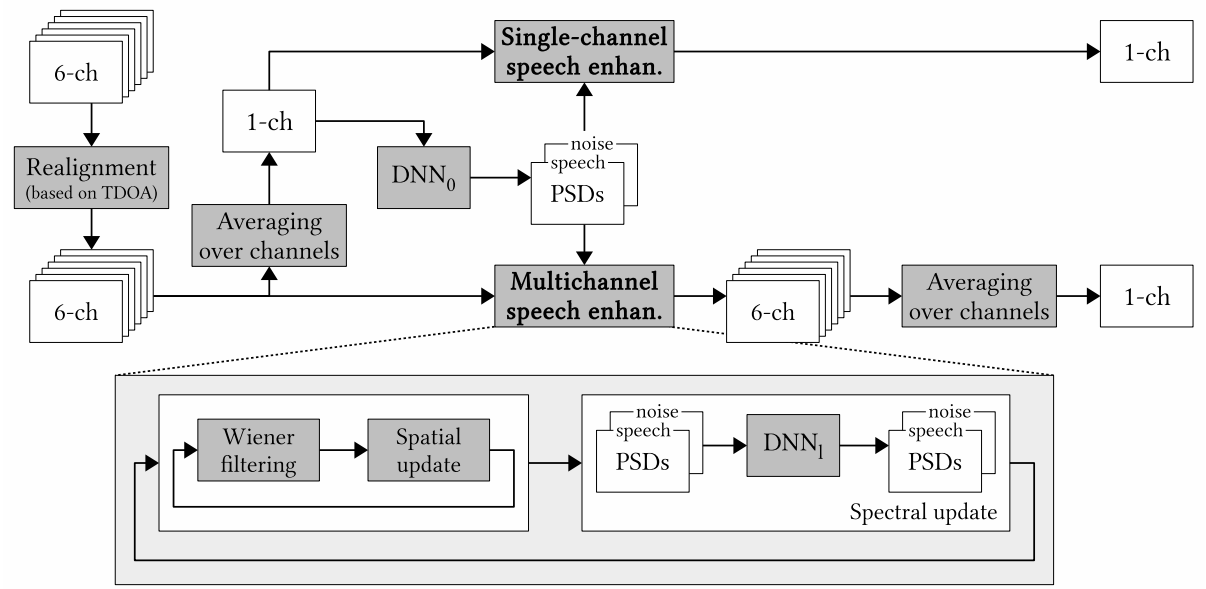


Fig. 1. Proposed DNN-based speech separation framework. Both the single-channel and the multichannel versions are shown.

set consists of 1,640 real and 1,640 simulated utterances (`dt05_real` and `dt05_simu`), while the test set consists of 1,320 real and 1,320 simulated utterances (`et05_real` and `et05_simu`). The utterances are taken from the 5k vocabulary subset of the Wall Street Journal corpus [41]. All data are sampled at 16 kHz. For further details, please refer to [40].

We used the source separation performance metrics defined in BSS Eval toolbox 3.0[1] [42] in most of the experiments presented in this section. The metrics include signal to distortion ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR), and signal to artifacts ratio (SAR). In addition, at the end of this section, we use the best speech separation system as the front-end, combine it with the best back-end in [29], and evaluate the ASR performance in terms of word error rate (WER).

The ground truth speech and noise signals, which are employed as training targets for DNN-based speech enhancement, were extracted using the baseline simulation tool provided by the challenge organizers [40]. The ground truth speech and noise signals for the real data are not perfect because they are extracted based on an estimation of the impulse responses (IRs) between the close-talking microphone and the microphones on the tablet device. Hence, the resulting source separation performance metrics for the real data are unreliable. Therefore, we evaluate the separation performance on the simulated data for studying the impact of the different design choices. By contrast, since the ground truth transcriptions for ASR are reliable, we evaluate the ASR performance on real data.

B. General system design

The proposed DNN-based speech separation framework is depicted in Fig. 1. A single-channel variant of this framework which boils down to the approach in [19] is also depicted for comparison.

The framework can be divided into three main successive steps, namely pre-processing, spectrogram initialization, and multichannel filtering. We describe each step in detail below and then provide further description of the DNNs in the following section.

1) Preprocessing: The STFT coefficients were extracted using a Hamming window of length 1024 and hopsize 512 resulting $F = 513$ frequency bins.

The time-varying time difference of arrivals (TDOAs) between the speaker's mouth and each of the microphones are first measured using the provided baseline speaker localization tool [40], which relies on a nonlinear variant of steered response power using the phase transform (SRP-PHAT) [43], [44]. All channels are then aligned with each other by shifting the phase of STFT of the input noisy signal $\mathbf{x}(f, n)$ in all time-frequency bins (f, n) by the opposite of the measured delay. This preprocessing is required to satisfy the model in (2) which assumes that the sources do not move over time.

In addition, we obtain a single-channel signal by averaging the realigned channels together. The combination of time alignment and channel averaging is known as DS beamforming in the microphone array literature [45], [46].

2) Spectrogram initialization: The initial PSDs of speech and noise are computed from the magnitude source spectra estimated by DNN_0.

3) Multichannel filtering: The PSDs and spatial covariance matrices of speech and noise are estimated and updated using the iterative algorithm (Algorithm 2), in which DNN_l is employed for spectrogram fitting at iteration l. In order to avoid numerical instabilities due to the use of single precision, the PSDs $v_j(f, n)$ are floored to 10^{-5} in the EM iteration.

In addition, the channels of estimated speech spatial image are averaged to obtain a single-channel signal for the ASR

[1] http://bass-db.gforge.inria.fr/bss_eval/

evaluation. Empirically, this provided better ASR performance than the use of one of the channels.

The number of spatial updates K is investigated in Section IV-E and the number of iterations L in Section IV-F.

C. DNN spectral models

Three design aspects are discussed below: the architecture, the input and output, and the training.

1) Architecture: The DNNs follow a multilayer perceptron (MLP) architecture. The number of hidden layers and the number of units in each input or hidden layer may vary. The number of units in the output layer equals the dimension of spectra multiplied by the number of sources. The activation functions of the hidden and output layers are rectified linear units (ReLUs) [47].

In this article, DNN_0 and DNN_l have a similar architecture. They have an input layer, three hidden layers, and an output layer. Both types of DNNs have hidden and output layers size of $F \times J = 1026$. DNN_0 has an input layer sizes of $F = 513$ and DNN_l of $F \times J = 1026$.

Other network architectures, e.g. recurrent neural network (RNN) and convolutional neural network (CNN), may be used instead of the one used here. The performance comparison with different architectures is beyond the scope of this article.

2) Inputs and outputs: In order to provide temporal context, the input frames are concatenated into *supervectors* consisting of a center frame, left context frames, and right context frames. In choosing the context frames, we use every second frame relative to the center frame in order to reduce the redundancies caused by the windowing of STFT. Although this causes some information loss, this enables the supervectors to represent a longer context [16], [48]. In addition, we do not use the magnitude spectra of the context frames directly, but the difference of magnitude between the context frames and the center frame. These differences act as complementary features similar to delta features. Preliminary experiments (not shown here) indicated that this improves DNN training and provides a minor improvement in terms of SDR.

Let $|\widetilde{x}(f,n)|$ be the input frames of DNN_0. The supervector can be expressed as

$$Z_0(f,n) = \begin{bmatrix} |\widetilde{x}(f,n-2c)| - |\widetilde{x}(f,n)| \\ \vdots \\ |\widetilde{x}(f,n)| \\ \vdots \\ |\widetilde{x}(f,n+2c)| - |\widetilde{x}(f,n)| \end{bmatrix} \tag{14}$$

where c is the one-sided context length in frames. The supervector for DNN_l, $Z_l(f,n)$, is constructed in a similar way where a stack of $\sqrt{z_j(f,n)}$ is used as input instead of $|\widetilde{x}(f,n)|$ (see Fig. 2 and 3). In this article, we considered $c = 2$, so the supervectors for the input of the DNNs were composed by 5 time frames (2 left context, 1 center, and 2 right context frames).

The dimension of the supervectors is reduced by principal component analysis (PCA) to the dimension of the DNN input. As shown in [49], dimensionality reduction by PCA significantly minimizes the computational cost of DNN training with a negligible effect on the performance of DNN. Standardization (zero mean, unit variance) is done element-wise before and after PCA over the training data as in [49]. The standardization factors and the PCA transformation matrix are then kept for pre-processing of any input. Thus, strictly speaking, the inputs of DNNs are not the supervectors of magnitude spectra $Z_0(f,n)$ and $Z_l(f,n)$, but their transformation into reduced dimension vectors.

Fig. 2 and 3 illustrates the inputs and outputs of the DNNs for spectrogram initialization and spectrogram fitting, respectively. F denotes the dimension of the spectra, $C = 2c + 1$ the context length, and J the number of sources.

3) Training criterion: The cost function used for DNN training is the sum of a primary cost function and an ℓ_2 regularization term. The ℓ_2 regularization term [50] is used to prevent overfitting and can be expressed as

$$\mathcal{D}_{\ell_2} = \frac{\lambda}{2} \sum_q w_q^2 \tag{15}$$

where w_q are the DNN weights and the regularization parameter is fixed to $\lambda = 10^{-5}$. No regularization is applied to the biases.

Table I summarizes the implementation of different cost functions for the experiments. In order to avoid numerical instabilities, instead of using the original formulation of IS divergence in (9), our implementation used a regularized formulation as shown in (16). It should be noted that the use of regularization in this case is a common practice to avoid instabilities [36], [51]. For the same reason, we used regularized formulations of KL divergence and Cauchy cost function as shown in (17) and (18), respectively. For these three divergences, the regularization parameter is set to $\delta = 10^{-3}$. In addition, geometric analysis on the PS cost function by considering that $m_j(f,n) \in \mathbb{R}_+^{F \times N}$ leads to a simplified formula shown in (19).

4) Training algorithm: The weights are initialized randomly from a zero-mean Gaussian distribution with standard deviation of $\sqrt{2/n_l}$, where n_l is the fan-in (the number of inputs to the neuron, which is equal to the size of the previous layer in our case) [52]. Finally, the biases are initialized to zero.

The DNNs are trained by greedy layer-wise supervised training [53] where the hidden layers are added incrementally. In the beginning, a NN with one hidden layer is trained after random weight initialization. The output layer of this trained NN is then substituted by new hidden and output layers to form a new NN, while the parameters of the existing hidden layer are kept. Thus, we can view this as a pre-training method for the training of a new NN. After random initialization for the parameters of new layers, the new NN is entirely trained. This procedure is done iteratively until the target number of hidden layers is reached.

Training is done by backpropagation with minibatch size of 100 and the ADADELTA parameter update algorithm [54]. Compared to standard stochastic gradient descent (SGD), ADADELTA employs adaptive per-dimension learning rates and does not require manual setting of the learning rate. The

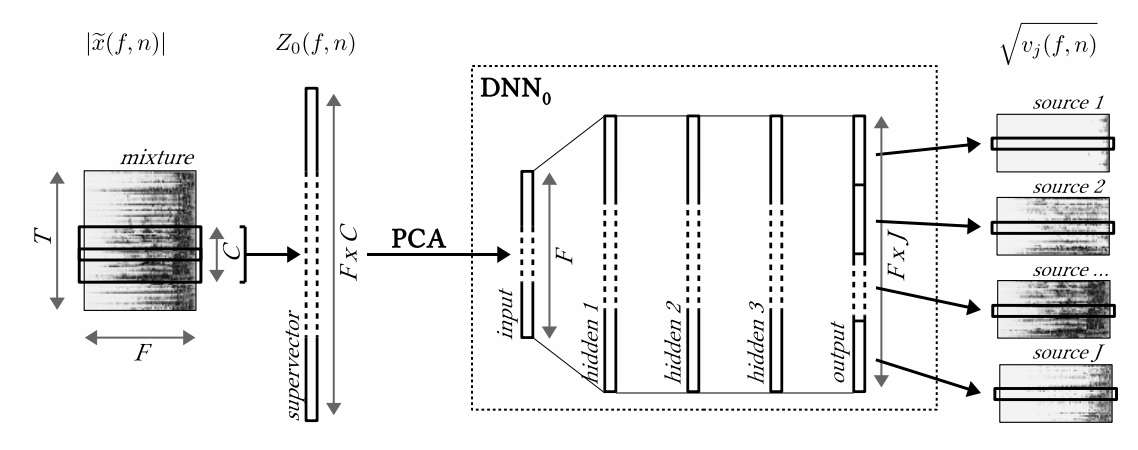


Fig. 2. Illustration of the inputs and outputs of the DNN for spectrogram initialization. Input: magnitude spectrum of the mixture (left). Output: magnitude spectra of the sources (right).

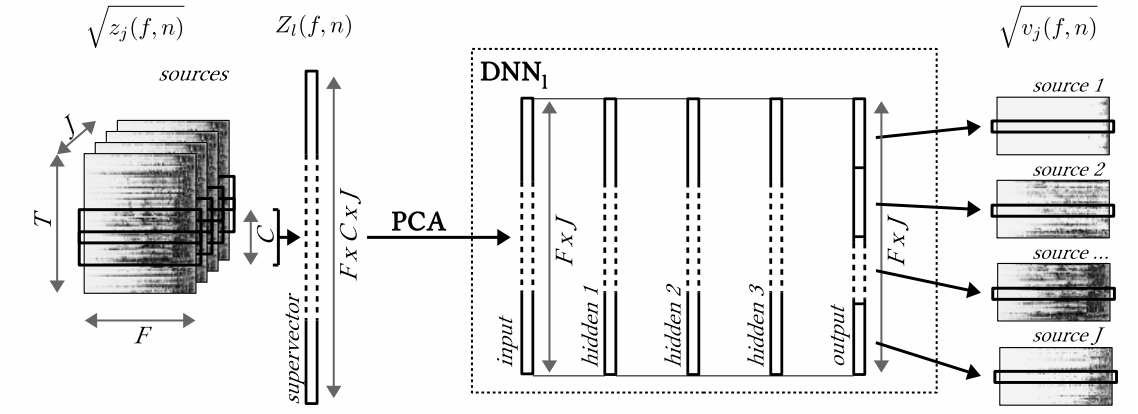


Fig. 3. Illustration of the inputs and outputs of the DNNs for spectrogram fitting. Input: stack of magnitude spectra of all sources (left). Output: magnitude spectra of the sources (right).

TABLE I
IMPLEMENTATION DETAILS OF THE DNN TRAINING COST FUNCTIONS.

Exp. label	Weight reg.	Primary cost function
IS	$\mathcal{D}_{\ell_2}$	$\mathcal{D}_{\overline{\text{IS}}} = \frac{1}{JFN}\sum_{j,f,n}\left(\frac{\left\|\widetilde{c}_j(f,n)\right\|^2+\delta}{v_j(f,n)+\delta} - \log\left(\left\|\widetilde{c}_j(f,n)\right\|^2+\delta\right) + \log\left(v_j(f,n)+\delta\right) - 1\right)$ (16)
KL	$\mathcal{D}_{\ell_2}$	$\mathcal{D}_{\overline{\text{KL}}} = \frac{1}{JFN}\sum_{j,f,n}\left(\left(\left\|\widetilde{c}_j(f,n)\right\|+\delta\right)\left(\log\left(\left\|\widetilde{c}_j(f,n)\right\|+\delta\right) - \log\left(\sqrt{v_j(f,n)}+\delta\right)\right) - \left\|\widetilde{c}_j(f,n)\right\| + \sqrt{v_j(f,n)}\right)$ (17)
Cau	$\mathcal{D}_{\ell_2}$	$\mathcal{D}_{\overline{\text{Cau}}} = \frac{1}{JFN}\sum_{j,f,n}\left(\frac{3}{2}\log\left(\left\|\widetilde{c}_j(f,n)\right\|^2 + v_j(f,n)+\delta\right) - \log\left(\sqrt{v_j(f,n)}+\delta\right)\right)$ (18)
PS	$\mathcal{D}_{\ell_2}$	$\mathcal{D}_{\overline{\text{PS}}} = \frac{1}{2JFN}\sum_{j,f,n}\left(\frac{v_j(f,n)}{\sum_{j'} v_{j'}(f,n)}\left\|\widetilde{x}(f,n)\right| - \left|\widetilde{c}_j(f,n)\right|\cos\left(\angle\widetilde{x}(f,n) - \angle\widetilde{c}_j(f,n)\right)\right)^2$ (19)
MSE	$\mathcal{D}_{\ell_2}$	$\mathcal{D}_{\text{MSE}} = \frac{1}{2JFN}\sum_{j,f,n}\left(\left\|\widetilde{c}_j(f,n)\right| - \sqrt{v_j(f,n)}\right)^2$ (13)

hyperparameters of ADADELTA are set to $\rho = 0.95$ and $\epsilon = 10^{-6}$ following [54]. The validation error is computed every epoch and the training is stopped after 10 consecutive epochs failed to obtain better validation error. The latest model which yields the best validation error is kept. Besides, the maximum number of training epochs is set to 100.

The DNNs for the source separation evaluation were trained on both the real and simulated training sets (`tr05_real` and `tr05_simu`) with the real and simulated development sets (`dt05_real` and `dt05_simu`) as validation data. Conversely, we trained the DNNs for the speech recognition evaluation on the real training set only (`tr05_real`) and validated them on the real development set only (`dt05_real`). The same DNNs were also used for the performance comparison to the general iterative EM algorithm. See [29] for the perfomance comparison between these two different settings.

D. Impact of cost functions

We first evaluated the impact of the cost function by setting $L = 0$ (see Algorithm 2) so that the separation relied on the PSD estimates $v_j(f,n)$ by letting the spatial covariance matrices $\mathbf{R}_j(f)$ be the identity matrix. This is equivalent to single-channel source separation for each channel.

Fig. 4 shows the evaluation results for the resulting 6-channel estimated speech signal on the simulated test set (`et05_simu`).

'KL', 'PS', and 'MSE' have comparable performance. Among these three cost functions, 'KL' is shown to have the best SDR and SIR properties, while 'PS' and 'MSE' whose performance is the same follow closely behind. 'MSE' is shown to have the best ISR property, while 'KL' and 'PS' follow behind. For the SAR, these three cost functions have almost the same performance. Among all of the cost functions used in this evaluation, 'IS' almost always has the worst performance. Interestingly, 'Cau' outperformed the others in terms of SIR, but it has a poor SAR property. Thus, in general, 'IS' and 'Cau' should be avoided for single-channel source separation with DNN model.

In addition, it is worth mentioning that the use of flooring function (e.g. ReLU activation function for the DNN outputs) during the training with 'IS', 'KL', 'Cau', 'PS' seems to be important. We found in additional experiments (not shown here) that training failed when a linear activation function was used for the output layer with these cost functions.

E. Impact of spatial parameters updates

In this subsection, we investigate the impact of the spatial parameters updates on the performance by setting the number of iterations to $L = 1$ and varying the number of spatial updates K, while ignoring the computation of $z_j(f,n)$ and the spectral parameters update (lines 16–19 of Algorithm 2). Thus, the spectral parameters $v_j(f,n)$ are only initialized by the first DNN (as in Section III-B) and kept fixed during the iterative procedure. We evaluate the different cost functions from Section III-B in this context again.

Fig. 5 shows the results for the resulting 6-channel estimated speech signal on the simulated test set (`et05_simu`). The x-axis of each chart corresponds to the number of spatial updates k. Thus, $k = 0$ is equivalent to single-channel source separation for each channel whose results are shown in Fig. 4.

In general, the performance of 'PS' saturated after a few updates, while the performance of other cost functions is increased with k in most metrics. Interestingly, after 20 iterations, each cost function showed its best property. Among all of the cost functions, 'KL' has the best SDR, 'Cau' the best SIR, and 'IS' the best SAR. While for the ISR, 'PS', 'MSE', and 'KL' performed almost identical and better than the other two cost functions.

In summary, the proposed multichannel approach outperformed single-channel DNN-based approach even when using DNN_0 only. The spatial parameters and their updates improved the enhancement performance. From the experiments using 20 spatial parameter updates, we can observe that each cost function has its own properties. 'KL' followed by 'MSE' are the most reasonable choices because they improved all of the metrics well. 'PS' is suitable for the tasks that put emphasis on the ISR. On the contrary, 'Cau' is suitable for the tasks in which the ISR is less important. Finally, 'IS' is suitable for the tasks that put emphasis on the SAR. Thus, the choice of the cost function should depend on the trade-off we want to achieve between these four metrics.

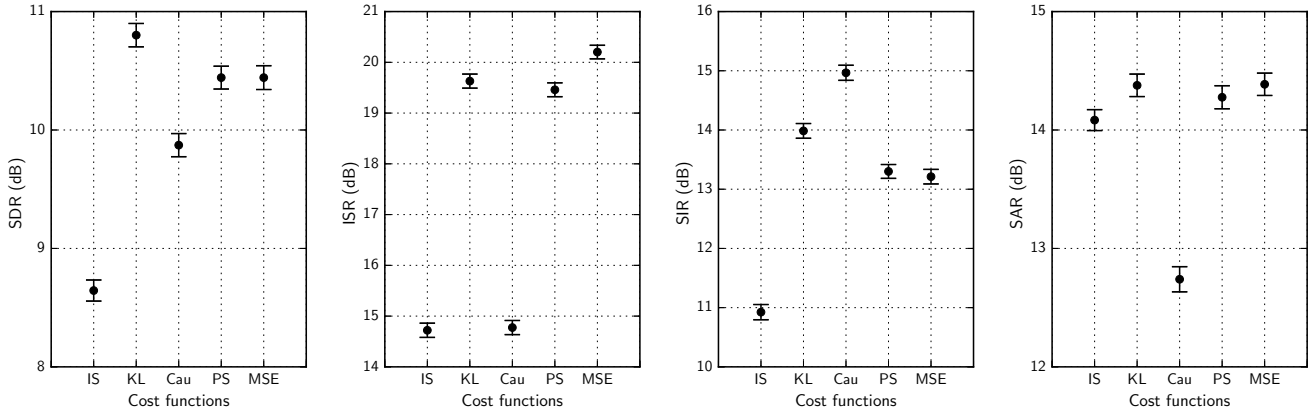
F. Impact of spectral parameters updates

In this subsection, we investigate the impact of spectral parameter updates (i.e. the spectrogram fitting) on the performance by setting the number of spatial updates to $K = 20$, varying the number of iterations L, and varying the DNN used for iteration l. We also evaluate different cost functions in this context, namely IS, KL, Cauchy, and MSE. We left the PS cost function because as shown previously, its SDR after 20 spatial updates was significantly lower than the others and the overall performance saturated already.

We trained two additional DNNs (DNN_1 and DNN_2) for spectrogram fitting. This allowed us to try different settings for the iterative procedure: (1) without spectral updates; (2) with spectral updates using only DNN_1; and (3) with spectral updates using DNN_1 and DNN_2.

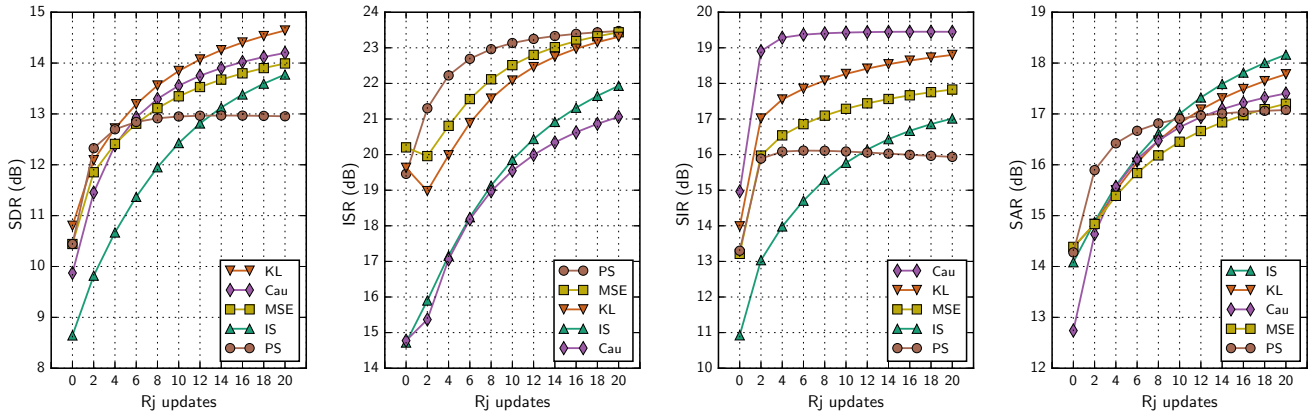
We present the comparison of these three settings using KL divergence as the cost function in Fig. 6. We then present the comparison of different cost functions using the third setting in Fig. 7. For both figures, the x-axis shows the index of EM iteration l, the update type (spatial or spectral), and the DNN index. Thus, $l = 0$ is equivalent to single-channel source separation for each channel whose results are shown in Fig. 4, while $l = 1$ with spatial updates is equivalent to the results shown in Fig. 5.

Fig. 6 shows that the use of a specific DNN for each iteration (here, DNN_1 for $l = 1$ and DNN_2 for $l = 2$) is beneficial. When a specific DNN is used, the spectral update provides a small improvement. Most importantly, this update allows the following spatial update to yield significant improvement. This behavior can be observed by comparing the performance of the spectral updates of EM iteration l and the spatial updates of the following iteration $l + 1$. Additionally,



(a) SDR (b) ISR (c) SIR (d) SAR

Fig. 4. Performance comparison for the DNNs trained with different cost functions. The PSDs $v_j(f,n)$ are estimated by DNN_0 and the spatial covariance matrices $\mathbf{R}_j(f)$ are the identity matrix. The SDR, ISR, SIR, and SAR measured on the observed 6-channel mixture signal are 3.8 dB, 18.7 dB, 4.0 dB, and 69.8 dB, respectively. The evaluation was done on the simulated test set (`et05_simu`). The figures show the mean value and the 95% confidence interval. Higher is better.



(a) SDR (b) ISR (c) SIR (d) SAR

Fig. 5. Performance comparison for various numbers of spatial updates with the DNNs trained with different cost functions. The PSDs $v_j(f,n)$ are estimated by the DNN_0 and the spatial covariance matrices $\mathbf{R}_j(f)$ are updated in the iterative procedure. The evaluation was done on the simulated test set (`et05_simu`). The figures show the mean value. The 95% confidence intervals are similar to those in Fig. 4. Higher is better. The legend is sorted by the final performance.

we can observe it by comparing the overall behavior of the "3 DNNs" curve to the "1 DNN" curve, in which no spectrogram fitting is done. Fig. 7 shows similar behavior for the other cost functions.

Fig. 6 also shows that the use of the same DNN for several iterations (here, DNN_1 for $l = 1$ and $l = 2$) did not improve the performance. Although the following spatial update recovered the performance, the use of a specific DNN for each iteration still provided better performance. We can observe this by comparing the "3 DNNs" curve to the "2 DNNs" curve for $l = 2$ and $l = 3$. It is understandable because there is a mismatch between the input and the training data of the DNN in this case.

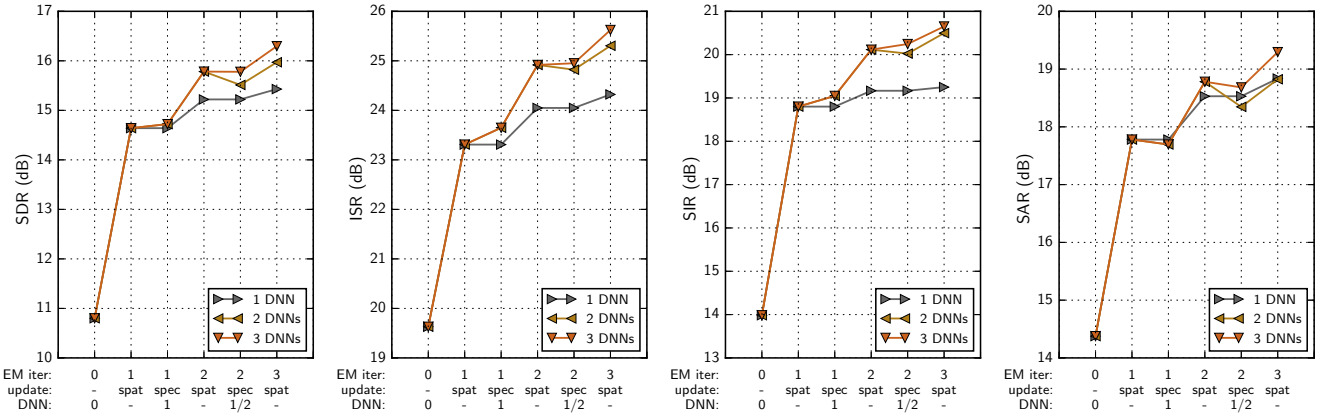
Fig. 7 shows that the performance of all cost functions improves with l. 'Cau' and 'IS' tend to saturate more quickly than the others.

In summary, the iterative spectral and spatial updates improve the enhancement performance. The performance saturates after few EM iteration. 'KL' and 'MSE' perform better than the other cost functions. Although the use of IS divergence for DNN training is theoretically motivated, the resulting performance is lower than the others for most metrics.

G. Comparison to NMF-based iterative EM algorithm

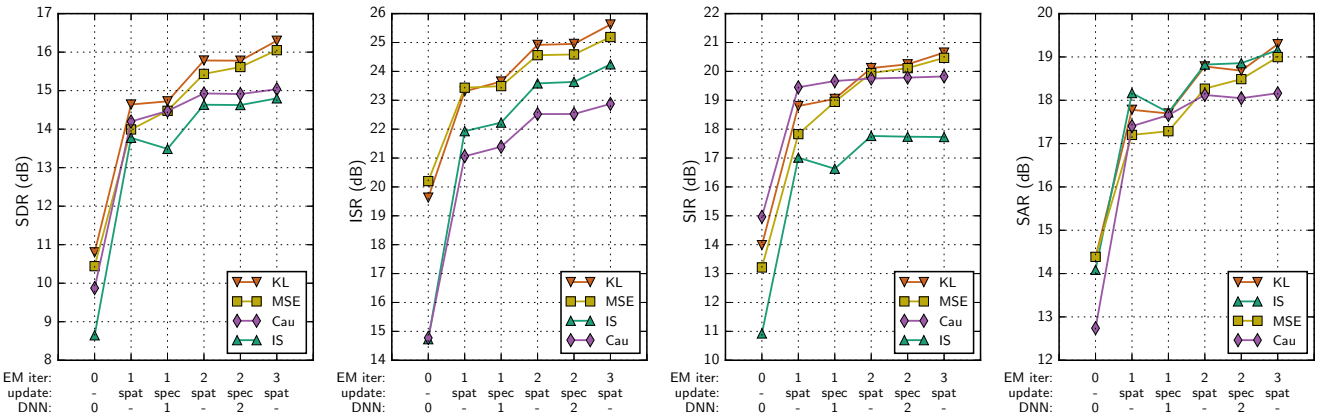
In this subsection, we compare the best system of the proposed framework to the NMF-based iterative EM algorithm [24] in terms of source separation performance. We used the algorithm implementation in the Flexible Audio Source Separation Toolbox (FASST)[2] and followed the settings used in [55]. The speech spectral and spatial models were trained on

[2]http://bass-db.gforge.inria.fr/fasst



(a) SDR (b) ISR (c) SIR (d) SAR

Fig. 6. Performance comparison for each update of the EM iterations in which different number of DNNs are used. In "1 DNN", there is no spectrogram fitting. In "2 DNNs", DNN_1 is used for spectrogram fitting of both $l = 1$ and $l = 2$. In "3 DNNs", DNN_1 and DNN_2 are used for spectrogram fitting of $l = 1$ and $l = 2$, respectively. Some markers and lines are not visible because they coincide. The DNNs are trained with KL divergence. The spatial covariance matrices $\mathbf{R}_j(f)$ are updated with $K = 20$. The evaluation was done on the simulated test set (`et05_simu`). The figures show the mean value. The 95% confidence intervals are similar to those in Fig. 4. Higher is better.



(a) SDR (b) ISR (c) SIR (d) SAR

Fig. 7. Performance comparison for each update of the EM iterations with the DNNs trained with different cost functions. Different DNNs are used for each EM iteration. The spatial covariance matrices $\mathbf{R}_j(f)$ are updated with $K = 20$. The evaluation was done on the simulated test set (`et05_simu`). The figures show the mean value. The 95% confidence interval for each cost function is similar to the interval of corresponding cost function in Fig. 4. Higher is better. The legend is sorted by the final performance.

the real training set (`tr05_real`). Meanwhile, the noise spectral and spatial models were initialized for each mixture using 5 seconds of background noise context based on its annotation. By doing so, the comparison is not completely fair since the proposed framework does not use this context information. However, this setting is favourable for the NMF-based iterative algorithm. As described in Section IV-C, the DNNs used in this evaluation were also trained on the real training set only. The separation results from this evaluation were then used for the speech recognition evaluation in Section IV-H.

Table II shows the performance of the NMF-based iterative EM algorithm after 50 EM iterations and the performance of the proposed framework after the spatial update of the EM iteration $l = 3$. The proposed framework was clearly better than the NMF-based iterative EM algorithm for all metrics.

TABLE II
PERFORMANCE COMPARISON IN TERMS OF SOURCE SEPARATION METRICS (IN DB). THE EVALUATION WAS DONE ON THE SIMULATED TEST SET (`ET05_SIMU`). THE TABLE SHOWS THE MEAN VALUE. HIGHER IS BETTER.

Enhancement method	SDR	ISR	SIR	SAR
NMF-based iterative EM [24]	7.72	10.77	13.29	12.29
Proposed: KL (3 DNNs)	13.25	24.25	15.58	18.23

This confirms that DNNs are able to model spectral parameters much better than NMF does.

H. Speech recognition

In this subsection, we evaluate the use of our best system as the front-end of a speech recognition system. We did a speech recognition evaluation by following the Kaldi setup distributed

by the CHiME-3 challenge organizers[3] [40], [56]. The evaluation includes the uses of (a) feature-space maximum likelihood regression (fMLLR) features [57]; (b) acoustic models based on Gaussian Mixture Model (GMM) and DNN trained with the cross entropy (CE) criterion followed by state-level minimum Bayes risk (sMBR) criterion [58]; and (c) language models with 5-gram Kneser-Ney (KN) smoothing [59] and rescoring using recurrent neural network-based language model (RNN-LM) [60]. The acoustic models are trained on enhanced multi-condition real and simulated data. The evaluation results are presented in terms of word error rate (WER). The optimization of the speech recognition back-end is beyond the scope of this article. Please refer to [56] for the further details of the methods used in the evaluation.

The evaluation results include the baseline performance (observed), DS beamforming, and NMF-based iterative EM algorithm [24]. The baseline performance was measured using only channel 5 of the observed signal. This channel is considered as the most useful channel because the corresponding microphone faces the user and is located at the bottom-center of the tablet device. DS beamforming was performed on the 6-channel observed signal as described in Section IV-B. For the NMF-based iterative EM algorithm and the proposed framework, we simply average over channels the separation results from the evaluation described in Section IV-G.

Table III shows the performance comparison using the GMM back-end retrained on enhanced multi-condition data. Table IV shows the performance comparison using the DNN+sMBR back-end trained with enhanced multi-condition data followed by 5-gram KN smoothing and RNN-LM rescoring. Both tables show the performance on the real development set (`dt05_real`) and the real test set (`et05_real`). Boldface numbers show the best performance for each dataset.

For the single-channel enhancement (see EM iteration $l = 0$), the WER on the real test set decreases by 22% and 21% relative using the GMM and the DNN+sMBR backends, respectively, w.r.t. the observed WER. Interestingly, this single-channel enhancement which is done after DS beamforming did not provide better performance compared to the DS beamforming alone. It indicates that proper exploitation of multichannel information is crucial.

The proposed multichannel enhancement then decreases the WER on the real test set up to 25% and 33% relative using the GMM and the DNN+sMBR backends, respectively, w.r.t. the corresponding single-channel enhancement. It decreases the WER up to 25% and 26% relative w.r.t. the DS beamforming alone. It also decreases the WER up to 16% and 24% relative w.r.t. the NMF-based iterative EM algorithm [24].

TABLE III
AVERAGE WERS (%) USING THE GMM BACK-END RETRAINED ON ENHANCED MULTI-CONDITION DATA. THE EVALUATION WAS DONE ON THE REAL SETS. LOWER IS BETTER.

Enhancement method	EM iter.	Update type	Dev	Test
Observed	-	-	18.32	33.02
DS beamforming	-	-	14.07	25.86
NMF-based iterative EM [24]	50	-	12.63	23.23
Proposed: KL (3 DNNs)	0	-	13.56	25.90
	1	spatial	11.17	20.42
		spectral	11.25	20.67
	2	spatial	10.80	19.96
		spectral	11.00	19.72
	3	spatial	**10.70**	**19.44**

TABLE IV
AVERAGE WERS (%) USING THE DNN+SMBR BACK-END TRAINED WITH ENHANCED MULTI-CONDITION DATA FOLLOWED BY 5-GRAM KN SMOOTHING AND RNN-LM RESCORING. THE EVALUATION WAS DONE ON THE REAL SETS. LOWER IS BETTER.

Enhancement method	EM iter.	Update type	Dev	Test
Observed	-	-	9.65	19.28
DS beamforming	-	-	6.35	13.70
NMF-based iterative EM [24]	50	-	6.10	13.41
Proposed: KL (3 DNNs)	0	-	6.64	15.18
	1	spatial	5.37	11.46
		spectral	5.19	11.46
	2	spatial	**4.87**	10.79
		spectral	4.99	11.12
	3	spatial	4.88	**10.14**

V. CONCLUSION

In this article, we presented a DNN-based multichannel source separation framework where the multichannel filter is derived from the source spectra, which are estimated by DNNs, and the spatial covariance matrices, which are updated iteratively in an EM fashion. Evaluation has been done for a speech enhancement task. The experimental results show that the proposed framework works well. It outperforms single-channel DNN-based enhancement and the NMF-based iterative EM algorithm [24]. The use of a single DNN to estimate the source spectra from the mixture already suffices to observe an improvement. Spectral updates by employing additional DNNs moderately improve the performance themselves, but they allow the following spatial updates to provide further significant improvement. We also demonstrate that the use of a specific DNN for each iteration is beneficial. The use of KL divergence as the DNN training cost function is shown to provide the best performance. The widely used MSE is also shown to perform very well.

Future directions concern alternative training targets for DNNs, the use of spatial features [9]–[11] as additional inputs, the incorporation of prior information about the source position, the use of more advanced network architectures, such as RNN [8] and CNN, and the use of more advanced training techniques, such as dropout.

[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/chime3

ACKNOWLEDGMENT

The authors would like to thank the developers of Theano [61] and Kaldi [62]. Experiments presented in this article were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

REFERENCES

[1] S. Makino, H. Sawada, and T.-W. Lee, Eds., *Blind Speech Separation*, ser. Signals and Communication Technology. Dordrecht, The Netherlands: Springer, 2007.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Chichester, West Sussex, UK: Wiley, 2009.

[3] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Chicester, West Sussex, UK: Wiley, 2012.

[4] G. R. Naik and W. Wang, Eds., *Blind Source Separation: Advances in Theory, Algorithms and Applications*, ser. Signals and Communication Technology. Berlin, Germany: Springer, 2014.

[5] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.

[6] L. Deng and D. Yu, *Deep Learning: Methods and Applications*, ser. Found. Trends Signal Process. Hanover, MA, USA: Now Publishers Inc., Jun. 2014, vol. 7, no. 3-4.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[8] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int'l. Conf. Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.

[9] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[10] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[11] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 116–120.

[12] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proc. Int'l. Symp. Chinese Spoken Lang. Process. (ISCSLP)*, Singapore, Sept 2014, pp. 250–254.

[13] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Int'l. Soc. for Music Inf. Retrieval (ISMIR)*, Taipei, Taiwan, Oct. 2014, pp. 477–482.

[14] ——, "Deep learning for monaural speech separation," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1562–1566.

[15] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[16] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 2135–2139.

[17] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[18] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 7092–7096.

[19] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal and Information Process. (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 577–581.

[20] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 92–101, Jan. 2015.

[21] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 708–712.

[23] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Jul. 2010.

[24] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[25] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in *Proc. Int'l. Soc. for Music Inf. Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012, pp. 85–90.

[26] M. Togami and Y. Kawaguchi, "Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 11, pp. 1612–1623, Nov. 2014.

[27] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[28] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 76–80.

[29] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, I. Illina, and A. Liutkus, "Robust ASR using neural network based speech enhancement and feature simulation," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 482–489.

[30] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. Hershey, PA, USA: IGI Global, 2011, ch. 7, pp. 162–185.

[31] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 205–208.

[32] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 266–270.

[33] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. ISCA INTERSPEECH*, Singapore, Sep. 2014, pp. 2685–2688.

[34] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[35] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[36] A. Lefèvre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 313–316.

[37] N. Bertin, C. Févotte, and R. Badeau, "A tempering approach for Itakura-Saito non-negative matrix factorization. with application to music transcription," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 1545–1548.

[38] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," in *Proc. Int'l. Symp. on Comput. Music Modeling and Retrieval*, Málaga, Spain, Jun. 2010.

[39] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2015.

[40] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 504–511.

[41] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," 2007, Linguistic Data Consortium, Philadelphia.

[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[43] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *Proc. Int'l. Conf. Latent Variable Analysis and Signal Separation*, Saint-Malo, France, 2010, pp. 41–48.
[44] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.
[45] J. McDonough and K. Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. Chicester, West Sussex, UK: Wiley, 2012, ch. 6.
[46] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, 2012.
[47] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proc. Int'l. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 15, Fort Lauderdale, FL, USA, Apr. 2011, pp. 315–323.
[48] A. A. Nugraha, K. Yamamoto, and S. Nakagawa, "Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition," *EURASIP J. Audio, Speech and Music Process.*, vol. 2014, no. 13, 2014.
[49] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1266–1279, Jul. 2016.
[50] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science, G. Montavon, G. Orr, and K.-R. Müller, Eds. Berlin, Germany: Springer, 2012, vol. 7700, ch. 19, pp. 437–478.
[51] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-negative matrix factorization for audio source separation," in *Excursions in Harmonic Analysis, Volume 4*, ser. Applied and Numerical Harmonic Analysis, R. Balan, M. Begué, J. J. Benedetto, W. Czaja, and K. A. Okoudjou, Eds. Switzerland: Springer, 2015, pp. 407–420.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv e-prints*, Feb. 2015. [Online]. Available: http://arxiv.org/abs/1502.01852
[53] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Conf. on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 2006, pp. 153–160.
[54] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *ArXiv e-prints*, Dec. 2012. [Online]. Available: http://arxiv.org/abs/1212.5701
[55] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The Flexible Audio Source Separation Toolbox Version 2.0," IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP), Florence, Italy, May 2014, Show & Tell. [Online]. Available: https://hal.inria.fr/hal-00957412
[56] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL/SRI system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 475–481.
[57] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.
[58] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. ISCA INTERSPEECH*, Lyon, France, Aug. 2013, pp. 2345–2349.
[59] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, Detroit, MI, USA, May 1995, pp. 181–184.
[60] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. ISCA INTERSPEECH*, Chiba, Japan, Sep. 2010, pp. 1045–1048.
[61] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688
[62] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop (ASRU)*, Hawaii, USA, Dec. 2011.

Aditya Arie Nugraha received the B.S. and M.S. degrees in electrical engineering from Institut Teknologi Bandung, Indonesia, and the M.Eng. degree in computer science and engineering from Toyohashi University of Technology, Japan, in 2008, 2011 and 2013, respectively. He is currently a Ph.D. student at the Université de Lorraine, France and Inria Nancy - Grand-Est, France. His research focuses on deep neural networks based audio source separation and noise-robust speech recognition.

Antoine Liutkus received the State Engineering degree from Telecom ParisTech, France, in 2005, and the M.Sc. degree in acoustics, computer science and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005. He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and obtained his PhD in electrical engineering at Telecom ParisTech in 2012. He is currently researcher at Inria Nancy Grand Est in the speech processing team. His research interests include audio source separation and machine learning.

Emmanuel Vincent is a Research Scientist with Inria (Nancy, France). He received the Ph.D. degree in music signal processing from the Institut de Recherche et Coordination Acoustique/Musique (Paris, France) in 2004 and worked as a Research Assistant with the Centre for Digital Music at Queen Mary, University of London (United Kingdom), from 2004 to 2006. His research focuses on probabilistic machine learning for speech and audio signal processing, with application to real-world audio source localization and separation, noise-robust speech recognition, and music information retrieval. He is a founder of the series of Signal Separation Evaluation Campaigns and CHiME Speech Separation and Recognition Challenges. He was an associate editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

The Journal of Open Source Software

Open-Unmix - A Reference Implementation for Music Source Separation

Fabian-Robert Stöter[1], **Stefan Uhlich**[2], **Antoine Liutkus**[1], **and Yuki Mitsufuji**[3]

1 Inria and LIRMM, University of Montpellier, France **2** Sony Europe B.V., Germany **3** Sony Corporation, Japan

DOI: 10.21105/joss.01667

Software

- Review
- Repository
- Archive

Submitted: 17 August 2019
Published: 08 September 2019

Summary

Music source separation is the task of decomposing music into its constitutive components, e.g., yielding separated stems for the vocals, bass, and drums. Such a separation has many applications ranging from rearranging/repurposing the stems (remixing, repanning, upmixing) to full extraction (karaoke, sample creation, audio restoration). Music separation has a long history of scientific activity as it is known to be a very challenging problem. In recent years, deep learning-based systems - for the first time - yielded high-quality separations that also lead to increased commercial interest. However, until now, no open-source implementation that achieves state-of-the-art results is available. *Open-Unmix* closes this gap by providing a reference implementation based on deep neural networks. It serves two main purposes. Firstly, to accelerate academic research as *Open-Unmix* provides implementations for the most popular deep learning frameworks, giving researchers a flexible way to reproduce results. Secondly, we provide a pre-trained model for end users and even artists to try and use source separation. Furthermore, we designed *Open-Unmix* to be one core component in an open ecosystem on music separation, where we already provide open datasets, software utilities, and open evaluation to foster reproducible research as the basis of future development.

Background

Music separation is a problem which has fascinated researchers for over 50 years. This is partly because, mathematically, there exists no closed-form solution when many sources (instruments) are recorded in a mono or stereo signal. To address the problem, researchers exploited additional knowledge about the way the signals were recorded and mixed. A large number of these methods are centered around "classical" signal processing methods. For a more detailed overview see (Rafii, Liutkus, Stöter, Mimilakis, & Bittner, 2017) and (Cano, FitzGerald, Liutkus, Plumbley, & Stöter, 2019). Many of these methods were hand-crafted and tuned to a small number of music recordings (Araki et al., 2012; Ono, Koldovsky, Miyabe, & Ito, 2013; Vincent et al., 2012). Systematic objective evaluation of these methods, however, was hardly feasible as freely available datasets did not exist at that time. In fact, for a meaningful evaluation, the ground truth separated stems are necessary. However, because commercial music is usually subject to copyright protection, and the separated stems are considered to be valuable assets in the music recording industry, they are usually unavailable.

Nonetheless, thanks to some artists who choose licenses like Creative Commons, that allow sharing of the stems, freely available datasets were released in the past five years and have enabled the development of data-driven methods. Since then, progress in performance has

been closely linked to the availability of more data that allowed the use of machine-learning-based methods. This led to a large performance boost similar to other audio tasks such as automatic speech recognition (ASR) where a large amount of data was available. In fact, in 2016 the speech recognition community had access to datasets with more than 10000 hours of speech (Amodei et al., 2016). In contrast, at the same time, the *MUSDB18* dataset was released (Rafii et al., 2017) which comprises 150 full-length music tracks – a total of just 10 hours of music. To date, this is still the largest freely available dataset for source separation. Nonetheless, even with this small amount of data, deep neural networks (DNNs) were not only successfully used for music separation but they are now setting the state-of-the-art in this domain as can be seen by the results of the community-based signal separation evaluation campaign (SiSEC) (Liutkus et al., 2017; Ono, Rafii, Kitamura, Ito, & Liutkus, 2015; Stöter, Liutkus, & Ito, 2018). In these challenges, the proposed systems are compared to other methods. Among the systems under test, classical signal processing based methods were clearly outperformed by machine learning methods. However they were still useful as a *fast* and often *simple to understand* baseline.

In the following, we will describe a number of these reference implementations for source separation. While there are some commercial systems available, such as *Audionamix XTRAX STEMS*, *IZOTOPE RX 7* or *AudioSourceRE*, we only considered tools that are available as open-source software, and are suitable for research.

The first publicly available software for source separation was *openBlissart*, released in 2011 (Weninger, Lehmann, & Schuller, 2011). It is written in C++ and accounts for the class of systems that are based on non-negative matrix factorization (NMF). In 2012, the *Flexible Audio Source Separation Toolbox (FASST)* was presented in (Ozerov, Vincent, & Bimbot, 2011; Salaün et al., 2014). It is written in MATLAB/C++ and is also based on NMF methods, but also includes other model-based methods. In 2016, the *untwist* library was proposed in (Roma, Grais, Simpson, Sobieraj, & Plumbley, 2016). It comprises several methods, ranging from classical signal-processing-based methods to feed-forward neural networks. The library is written in Python 2.7. Unfortunately, it has not been updated since 2017 and many of its methods are not subjected to automated testing. *Nussl* is a very recent framework, presented in (Manilow, Seetharaman, & Pardo, 2018). It includes a large number of methods and generally focuses on classical signal processing methods rather than machine-learning-based techniques. It has built-in interfaces for common evaluation metrics and data sets. The library offers great modularity and a good level of abstraction. However, this also means that it is challenging for beginners who might only want to focus on changing the machine learning parts of the techniques.

The main problem with these implementations is that they do not deliver state-of-the-art results. No open-source system is available today that matches the performance of the best system proposed more than four years ago by (Uhlich, Giron, & Mitsufuji, 2015). We believe that the lack of such a baseline has a serious negative impact on future research on source separation. Many new methods that were published in the last few years are usually compared to their own baseline implementations, thus showing relative instead of absolute performance gains, so that other researchers cannot assess if a method performs as well as state-of-the-art. Also, the lack of a common reference for the community potentially misguides young researchers and students who enter the field of music separation. The result of this can be observed by looking at the popularity of the above-mentioned music separation frameworks on GitHub: all of the frameworks mentioned above, combined, are less popular than two recent deep learning papers that were accompanied by code such as `MTG/DeepConvSep` from (Chandna, Miron, Janer, & Gómez, 2017) and `f90/Wave-U-Net` from (Stoller, Ewert, & Dixon, 2018). Thus, users might be confused regarding which of these implementations can be considered state-of-the-art.

Open-Unmix

We propose to close this gap with *Open-Unmix*, which applies machine learning to the specific tasks of music separation. With the rise of simple to use machine learning frameworks such as *Pytorch*, *Keras*, *Tensorflow* or *NNabla*, the technical challenge of developing a music separation system appears to be very low at first glance. However, the lack of domain knowledge about the specifics of music signals often results in poor performance where issues are difficult to track using learning-based algorithms. We therefore designed *Open-Unmix* to address these issues by relying on procedures that were verified by the community or have proven to work well in the literature.

Design Choices

The design choices made for *Open-Unmix* have sought to reach two somewhat contradictory objectives. Its first aim is to have state-of-the-art performance, and its second aim is to still be easily understandable, so that it can serve as a basis for research to allow improved performance in the future. In the past, many researchers faced difficulties in pre- and post-processing that could be avoided by sharing domain knowledge. Our aim was thus to design a system that allows researchers to focus on A) new representations and B) new architectures.

Framework specific vs. framework agnostic

We choose *PyTorch* to serve as a reference implementation due to its balance between simplicity and modularity (Stöter & Liutkus, 2019a). Furthermore, we already ported the core model to NNabla and plan to release a port for Tensorflow 2.0, once the framework is released. Note that the ports will not include pre-trained models as we cannot make sure the ports would yield identical results, thus leaving a single baseline model for researchers to compare with.

"MNIST-like"

Keeping in mind that the learning curve can be quite steep in audio processing, we did our best for *Open-unmix* to be:

- **simple to extend**: The pre/post-processing, data-loading, training and models part of the code is isolated and easy to replace/update. In particular, a specific effort was done to make it easy to replace the model.
- **not a package**: The software is composed of largely independent and self-containing parts, keeping it easy to use and easy to change.
- **hackable (MNIST like)**: Due to our objective of making it easier for machine-learning experts to try out music separation, we did our best to stick to the philosophy of baseline implementations for this community. In particular, *Open-unmix* mimics the famous MNIST example, including the ability to instantly start training on a dataset that is automatically downloaded.

Reproducible

Releasing *Open-Unmix* is first and foremost an attempt to provide a reliable implementation sticking to established programming practice as were also proposed in (McFee et al., 2018). In particular:

- **reproducible code**: everything is provided to exactly reproduce our experiments and display our results.

- **pre-trained models**: we provide pre-trained weights that allow a user to use the model right away or fine-tune it on user-provided data (Stöter & Liutkus, 2019b, 2019c).
- **tests**: the release includes unit and regression tests, useful to organize future open collaboration using pull requests.

Results

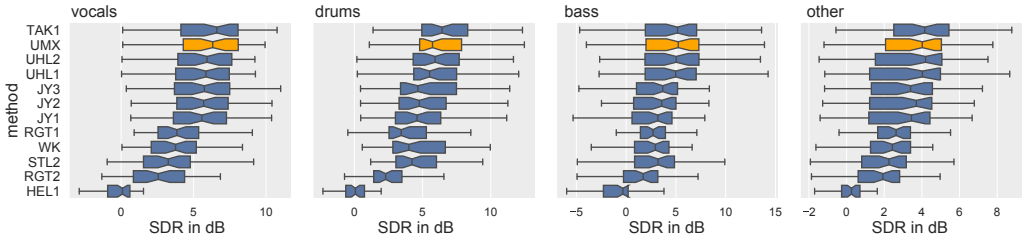


Figure 1: Boxplots of evaluation results of the `UMX` model compared with other methods from (Stöter et al., 2018) (methods that did not only use MUSDB18 for training were omitted)

Open-Unmix is based on the bi-directional LSTM model from (Uhlich et al., 2017) and we compared it to other separation models that were submitted to the last SiSEC contest (Stöter et al., 2018). The results of `UMX` are depicted in 1. It can be seen that our proposed model reaches state-of-the-art results. There is no statistically significant difference between the best method `TAK1` and `UMX`. Because `TAK1` is not released as open-source, this indicates that *Open-Unmix* is the current state-of-the-art open-source source separation system.

Community

Open-Unmix was developed by Fabian-Robert Stöter and Antoine Liutkus at Inria Montpellier. The research concerning the deep neural network architecture as well as the training process was done in close collaboration with Stefan Uhlich and Yuki Mitsufuji from Sony Corporation.

In the future, we hope the software will be well received by the community. *Open-Unmix* is part of an ecosystem of software, datasets, and online resources: the **sigsep** community.

First, we provide MUSDB18 (Rafii et al., 2017) and MUSDB18-HQ (Rafii, Liutkus, Stöter, Mimilakis, & Bittner, 2019) which are the largest freely available datasets; this comes with a complete toolchain to easily parse and read the datasets (Stöter & Liutkus, 2019a). We maintain *museval*, the most used evaluation package for source separation (Stöter & Liutkus, 2019b). We also are the organizers of the largest source separation evaluation campaign such as (Stöter et al., 2018). In addition, we implemented a reference implementation using a multi-channel Wiener filter, released in (Liutkus & Stöter, 2019). The `sigsep` community is organized and presented on its own website. *Open-Unmix* itself can be found on https://open.unmix.app, which links to all other relevant sites and provides further information, such as audio demos.

Outlook

Open-Unmix is a community-focused project. We therefore encourage the community to submit bug-fixes and comments and improve the computational performance. However, we are not looking for changes that only focus on improving the separation performance as this would be out of scope for a baseline implementation. Instead, we expect many researchers

will fork the software as a basis for their research and the documentation explicates several custom options to extend the code (shown here).

References

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *ICML* (pp. 173–182).

Araki, S., Nesta, F., Vincent, E., Koldovsky, Z., Nolte, G., Ziehe, A., & Benichoux, A. (2012). The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation -. In *10th international conference on latent variable analysis and signal separation*. doi:10.1007/978-3-642-28551-6_51

Cano, E., FitzGerald, D., Liutkus, A., Plumbley, M. D., & Stöter, F. (2019). Musical source separation: An introduction. *IEEE Signal Processing Magazine*, *36*(1), 31–40. doi:10.1109/MSP.2018.2874719

Chandna, P., Miron, M., Janer, J., & Gómez, E. (2017). Monoaural audio source separation using deep convolutional neural networks. In *Latent variable analysis and signal separation* (pp. 258–266). doi:10.1007/978-3-319-53547-0_25

Liutkus, A., & Stöter, F.-R. (2019, September). sigsep/norbert: v0.2.1. doi:10.5281/zenodo.3386463

Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., et al. (2017). The 2016 signal separation evaluation campaign. In *Proc. Intl. Conference on latent variable analysis and signal separation (lva/ica)* (pp. 323–332). Springer International Publishing. doi:10.1007/978-3-319-53547-0_31

Manilow, E., Seetharaman, P., & Pardo, B. (2018). The northwestern university source separation library. In *ISMIR* (pp. 297–305).

McFee, B., Kim, J. W., Cartwright, M., Salamon, J., Bittner, R. M., & Bello, J. P. (2018). Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research. *IEEE Signal Processing Magazine*, *36*(1), 128–137. doi:10.1109/MSP.2018.2875349

Ono, N., Koldovsky, Z., Miyabe, S., & Ito, N. (2013). The 2013 signal separation evaluation campaign. In *Proc. IEEE international workshop on machine learning for signal processing (MLSP)* (pp. 1–6). doi:10.1109/MLSP.2013.6661988

Ono, N., Rafii, Z., Kitamura, D., Ito, N., & Liutkus, A. (2015). The 2015 signal separation evaluation campaign. In *Proc. Intl. Conference on latent variable analysis and signal separation (lva/ica)*. Liberec, Czech Republic, doi:10.1007/978-3-319-22482-4_45

Ozerov, A., Vincent, E., & Bimbot, F. (2011). A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1118–1133. doi:10.1109/TASL.2011.2172425

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2017, December). MUSDB18, a corpus for audio source separation. doi:10.5281/zenodo.1117372

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019, August). MUSDB18-hq - an uncompressed version of musdb18. doi:10.5281/zenodo.3338373

Roma, G., Grais, E. M., Simpson, A., Sobieraj, I., & Plumbley, M. D. (2016). Untwist: A new toolbox for audio source separation. In *Extended abstracts for the late-breaking demo session of the 17th international society for music information retrieval conference, ismir* (pp. 7–11).

Salaün, Y., Vincent, E., Bertin, N., Souviraà-Labastie, N., Jaureguiberry, X., Tran, D. T., & Bimbot, F. (2014, May). The Flexible Audio Source Separation Toolbox Version 2.0. ICASSP. Retrieved from https://hal.inria.fr/hal-00957412

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*.

Stöter, F.-R., & Liutkus, A. (2019a, August). sigsep/open-unmix-pytorch: Initial release of Open-Unmix. doi:10.5281/zenodo.3382104

Stöter, F.-R., & Liutkus, A. (2019b, August). Open-unmix-pytorch umx. doi:10.5281/zenodo.3370486

Stöter, F.-R., & Liutkus, A. (2019c, August). Open-unmix-pytorch umx-hq. doi:10.5281/zenodo.3370489

Stöter, F.-R., & Liutkus, A. (2019a, July). sigsep/sigsep-mus-db: v0.1.7. doi:10.5281/zenodo.3271451

Stöter, F.-R., & Liutkus, A. (2019b, June). sigsep/sigsep-mus-eval: v0.3.0. doi:10.5281/zenodo.3261102

Stöter, F.-R., Liutkus, A., & Ito, N. (2018). The 2018 signal separation evaluation campaign. In *Latent variable analysis and signal separation: 14th international conference, lva/ica 2018, surrey, uk* (pp. 293–305). doi:10.1007/978-3-319-93764-9_28

Uhlich, S., Giron, F., & Mitsufuji, Y. (2015). Deep neural network based instrument extraction from music. In *Icassp* (pp. 2135–2139). doi:10.1109/ICASSP.2015.7178348

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., & Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. In *Icassp*. New Orleans, LA, USA. doi:10.1109/ICASSP.2017.7952158

Vincent, E., Araki, S., Theis, F. J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., et al. (2012). The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges, *92*(8), 1928–1936. doi:10.1016/j.sigpro.2011.10.007

Weninger, F., Lehmann, A., & Schuller, B. (2011). OpenBliSSART: Design and evaluation of a research toolkit for blind source separation in audio recognition tasks. In *Proc. IEEE intl. Conf. On acoustics, speech and signal processing (icassp)* (pp. 1625–1628). doi:10.1109/ICASSP.2011.5946809

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
OPTICAL SENSORS
APPLIED MATHEMATICS
INFORMATION THEORY AND COMPUTATION

Received
28 January 2014

Accepted
13 June 2014

Published
9 July 2014

Correspondence and requests for materials should be addressed to S.G. (sylvain.gigan@espci.fr)

Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium

Antoine Liutkus[1,5], David Martina[1,4], Sébastien Popoff[1], Gilles Chardon[1,2], Ori Katz[1,4], Geoffroy Lerosey[1], Sylvain Gigan[1,4], Laurent Daudet[1] & Igor Carron[3]

[1]Institut Langevin, ESPCI ParisTech, Paris Diderot Univ., UPMC Univ. Paris 6, CNRS UMR 7587, Paris, France, [2]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, [3]TEES SERC, Texas A&M University, [4]Laboratoire Kastler-Brossel, UMR8552 CNRS, Ecole Normale Supérieure, Univ. Paris 6, Collège de France, 24 rue Lhomond, 75005 PARIS, [5]Inria, CNRS, Loria UMR 7503 Villers-lès-Nancy, France.

The recent theory of compressive sensing leverages upon the structure of signals to acquire them with much fewer measurements than was previously thought necessary, and certainly well below the traditional Nyquist-Shannon sampling rate. However, most implementations developed to take advantage of this framework revolve around controlling the measurements with carefully engineered material or acquisition sequences. Instead, we use the natural randomness of wave propagation through multiply scattering media as an optimal and instantaneous compressive imaging mechanism. Waves reflected from an object are detected after propagation through a well-characterized complex medium. Each local measurement thus contains global information about the object, yielding a purely analog compressive sensing method. We experimentally demonstrate the effectiveness of the proposed approach for optical imaging by using a 300-micrometer thick layer of white paint as the compressive imaging device. Scattering media are thus promising candidates for designing efficient and compact compressive imagers.

Acquiring digital representations of physical objects - in other words, *sampling* them - was, for the last half of the 20th century, mostly governed by the Shannon-Nyquist theorem. In this framework, depicted in Fig. 1(a), a signal is acquired by N regularly-spaced samples whose sampling rate is equal to at least twice its bandwidth. However, this line of thought is thoroughly pessimistic since most signals and objects of interest are not only of limited bandwidth but also generally possess some additional *structure*[1]. For instance, images of natural scenes are composed of smooth surfaces and/or textures, separated by sharp edges.

Recently, new mathematical results have emerged in the field of Compressive Sensing (or Compressed Sensing, CS in short) that introduce a paradigm shift in signal acquisition. It was indeed demonstrated by Donoho, Candès, Tao and Romberg[2–4] that this additional structure could actually be exploited *directly at the acquisition stage* so as to provide a drastic reduction in the number of measurements without loss of reconstruction fidelity.

For CS to be efficient, the sampling must fulfill specific technical conditions that are hard to translate into practical design guidelines. In this respect, the most interesting argument featured very early on in[2–4] is that a *randomized* sensing mechanism yields perfect reconstruction with high probability. As a matter of convenience, hardware designers have created physical systems that *emulate* this property. This way, each measurement gathers information from all parts of the object, in a controlled but pseudo-random fashion. Once this is achieved, CS theory provides good reconstruction guarantees.

In the past few years, several hardware implementations capable of performing such random compressive sampling were introduced[5–13]. In optics, these include the single pixel camera[6], which is depicted in Fig. 1(b), and uses a digital array of micromirrors (abbreviated DMD) to sequentially reflect different random portions of the object onto a single photodetector. Other approaches include phase modulation with a spatial light modulator[10], or a rotating optical diffuser[13]. The idea of random multiplexing for imaging has also been considered in other domains of wave propagation. CS holds much promise in areas where detectors are rather complicated and expensive such as the THz or far infrared. In this regards, there have been proposals to implement CS imaging procedures in the THz using random pre-fabricated masks[5], DMD or SLM photo-generated contrast masks on semi-conductors slabs[14] and efforts are also pursued on tunable metamaterial reflectors[15]. Recently, a carefully engineered metamaterial aperture was used to generate complex RF beams at different frequencies[8].

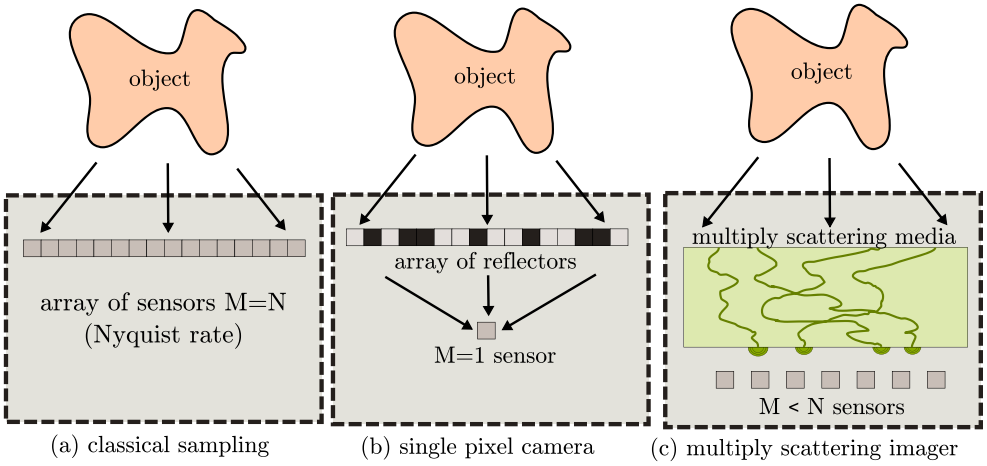


Figure 1 | **Concept.** (a) Classical Nyquist-Shannon sampling, where the waves originating from the object, of size N, are captured by a dense array of M = N sensors. (b) The "Single Pixel Camera" concept, where the object is sampled by M successive random projections onto a single sensor using a digital multiplexer. (c) Imaging with a multiply scattering medium. The M sensors capture, in a parallel fashion, several random projections of the original object. In cases (b) and (c), sparse objects can be acquired with a low sensor density M/N < 1.

However, these CS implementations come with some limitations. First, these devices include carefully engineered hardware designed to achieve randomization, via a DMD[6], a metamaterial[8] or a coded aperture[11]. Second, the acquisition time of most implementations can be large because they require the *sequential* generation of a large number of random patterns.

In this work, we replace such man-made *emulated* randomization by a natural multiply scattering material, as depicted in Fig. 1(c). Whereas scattering is usually seen as a time-varying nuisance, for instance when imaging through turbid media[16], the recent results of wave control in stable complex material have largely demonstrated that it could also be exploited, for example so as to build focusing systems that beat their coherent counterparts in terms of resolution[17,18]. Such complex and stable materials are readily available in several frequency ranges -they were even coined in as one-way physical functions for hardware cryptography[19]. In the context of CS, such materials perform an efficient randomized multiplexing of the object into several sensors and hence appear as *analog* randomizers. The approach is applicable in a broad wavelength range and in many domains of wave propagation where scattering media are available. As such, this study is close in spirit to earlier approaches such as the random reference structure[20], the random lens imager[7], the metamaterial imager[8], or the CS filters proposed in[21] for microwave imaging. They all abandoned digitally controlled multiplexors as randomizers. Still, we go further in this direction and even drop the need for a designer to *craft* the randomizer.

Compressive sampling with multiply scattering material has several advantages. First, it has recently been shown that they have an optimal multiplexing power for coherent waves[22], which consequently makes them optimal sensors within the CS paradigm. Second, these materials are often readily available and require very few engineering. In the domain of optics for example, we demonstrate one successful implementation using a 300 μm layer of Zinc Oxide (ZnO), which is essentially white paint. Third, contrarily to most aforementioned approaches, this sensing method provides the somewhat unique ability to take a scalable number of measurements in parallel, thus with a potential of strongly reducing acquisition time. In practice, if 500 samples are required to reconstruct a given image using CS principles, this imaging framework allows their acquisition at once on 500 independent sensors, whereas state-of-the-art systems such as the single pixel camera require a sequence of 500 random patterns on the DMD.

On practical grounds, the use of a multiply scattering material in CS raises several ideas that we consider in this study. First, the random multiplexing achieved through multiple scattering must be measured *a posteriori*, since it is no longer known *a priori* as in engineered random sensing. This calibration problem has been the topic of recent studies in the context of CS[23] and we propose here a simple least squares calibration procedure that extends our previous work[24,25]. Second, the use of such a measured Transmission Matrix (TM) induces an inherent uncertainty in the sensing mechanism, that can be modeled as noise in the observations. As we show both through extensive simulations and actual experiments, this uncertainty is largely compensated by the use of adequate reconstruction techniques. In effect, the imager we propose almost matches the performance of idealized sub-Nyquist random sensing.

Theoretical background

In its simplest form, CS may be understood as a way to solve an underdetermined linear inverse problem. Let x be the object to image, understood as a $N \times 1$ vector, and let us suppose that x is only observed through its multiplication y by a known measurement matrix H, of dimension $M \times N$, we have $y = Hx$. Each one of the M entries of y is thus the scalar product of the object with the corresponding row of H. When there are fewer measurements than the size of the object, i.e. $M < N$, it is impossible to recover x perfectly without further assumptions, since the problem has infinitely many solutions. However, if x is known to be *sparse*, meaning that only a few of its coefficients are nonzero (such as stars in astronomical images), and provided H is sufficiently random, x can still be recovered uniquely through sparse optimization techniques[1].

In a signal processing framework, the notion of *structure* may also be embodied as sparsity in a known representation[1]. For example, most natural images are not sparse, yet often yield near-sparse representations in the wavelet domain. If the object x is known to have some sparse or near-sparse representation s in a known basis B ($x = Bs$), then it may again be possible to recover it from a few samples, by solving $y = HBs$, provided H and B obey some technical conditions such as *incoherence*[1–4,26].

In practice, when trying to implement Compressive Sensing in a hardware device, fulfilling this *incoherence* requirement is nontrivial. It requires a way to deterministically *scramble* the information somewhere between the object and the sensors. Theory shows that an efficient way to do this is by using *random measurement matrices* H *or* HB[2–4]. Using such matrices, it can indeed be shown[26] that the number of samples required to recover the object is mostly governed by its sparsity k, i.e. the number of its nonzero coefficients in the given basis. If the coefficients of the $M \times N$ measurement matrix are

independent and identically distributed (i.i.d.) with respect to a Gaussian distribution, perfect reconstruction can be achieved with only $\mathcal{O}(k \log(N/k))$ measurements[27]. Furthermore, many algorithms are available, for instance Orthogonal Matching Pursuit (OMP) or Lasso[26,28], which can efficiently perform such reconstruction under sparsity constraints.

Using natural complex media as random sensing devices

Our approach is summarized in Fig. 1(c) and its implementation in an optical experiment is depicted in Fig. 2. The coherent waves originating from the object and entering the imaging system propagate through a multiply scattering medium. Within the imager, propagation produces a seemingly random and wavelength-dependent interference pattern called speckle on the sensors plane. The speckle figure is the result of the random phase variations imposed on the waves by the propagation within the multiply scattering sample[29]. Scattering, although the realization of a random process, is deterministic: for a given input, and as long as the medium is stable, the interference speckle figure is fully determined and remains constant. In essence, the complex medium acts as a highly efficient analog multiplexer for light, with an input-output response characterized by its transmission-matrix[24,25]. We highlight the fact that the multiple scattering material is not understood here as a nuisance occurring between the object and the sensors, but rather as a desirable component of the imaging system itself. After propagation, sensing takes place using a limited number $M < N$ of sensors.

Let x and y denote the $N \times 1$ and $M \times 1$ vectors gathering the value of the complex optical field at discrete positions before and after, respectively, the scattering material. It was confirmed experimentally[24,25] that any particular output y_m can be efficiently modeled as a linear function of the N complex values x_n of the input optical field:

$$y_m = \sum_{n=1}^{N} h_{mn} x_n,$$

where the mixing factor $h_{mn} \in \mathbb{C}$ corresponds to the overall contribution of the input field x_n into the output field y_m. All these factors can be gathered into a complex matrix $[H]_{mn} = h_{mn}$ called the *Transmission Matrix* (TM), which characterizes the action of the scattering material on the propagating waves between input and output. The medium hence produces a very complex but deterministic mixing of the input to the output, that can be understood as spatial multiplexing. This linear model, in the ideal noiseless case, can be written more concisely as:

$$y = Hx.$$

As can be seen, each of the M measurements of the output complex field may hence be considered as a scalar product between the input and the corresponding row of the TM. If multiply scattering materials have already been considered for the purpose of *focusing*, thus serving as perfect "opaque lenses"[17,18], the main idea of the present study is to exploit them for compressive imaging. In other wavelength domains than optics, analogous configurations may be designed to achieve CS through multiple scattering. For instance, a collection of randomly packed metallic scatterers could be used as a multiply scattering media from the microwave domain up to the far infrared, and the method proposed here could allow imaging at these frequencies with only a few sensors. A similar approach could be used to lower the number of sensors in 3D ultrasound imaging using CS through multiple scattering media.

In our optical experimental setup, we used a Spatial Light Modulator (an array of $N = 1024$ micromirrors, abbreviated as SLM) to calibrate the system and also to display various objects, using a monochromatic continuous wave laser as light source.

During a first *calibration* phase, which lasts a few minutes and needs to be performed only once, a series of controlled inputs x are emitted and the corresponding outputs y are measured. The TM can be estimated through a simple least-squares error procedure, which generalizes the method proposed in[24,25], as detailed in the supplementary material below. In short, this calibration procedure benefits from an arbitrarily high number of measurements for calibration, which permits to better estimate the TM. It is important to note here that the need for calibration is the main disadvantage of this

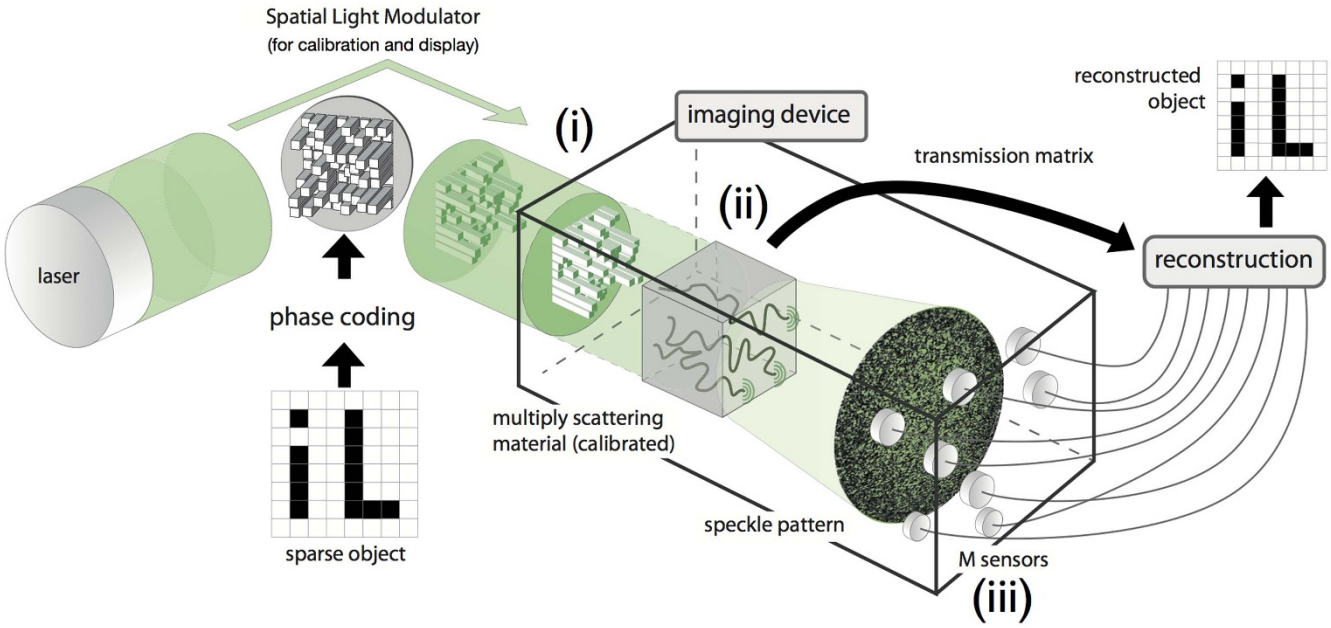


Figure 2 | Experimental setup for compressive imaging using multiply scattering medium. Within the imaging device, waves coming from the object (i) go through a scattering material (ii) that efficiently multiplexes the information to all M sensors (iii). Provided the transmission matrix of the material has been estimated beforehand, reconstruction can be performed using only a limited number of sensors, potentially much lower than without the scattering material. In our optical scenario, the light coming from the object is displayed using a spatial light modulator.

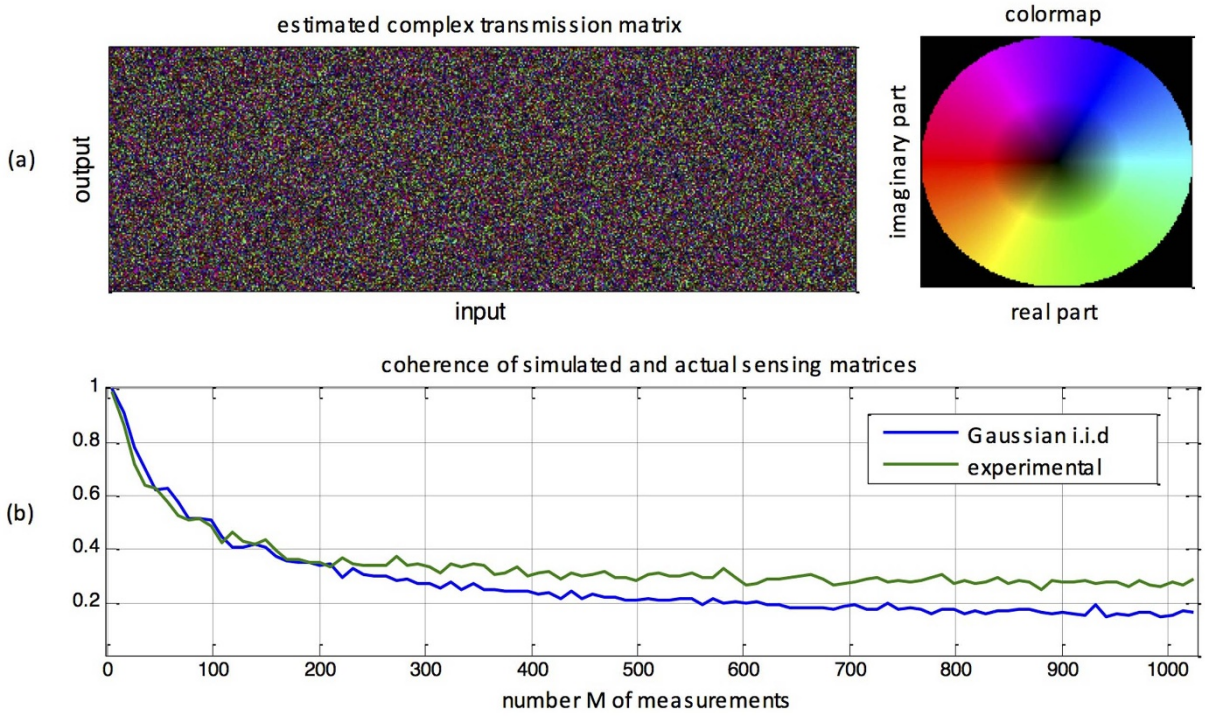


Figure 3 | Experimentally measured Transmission Matrix (TM). (a) TM for a multiply scattering material as obtained in our experimental study. (b) Coherence of sensing matrices as a function of their number M of rows, for both a randomly generated Gaussian i.i.d. matrix, and an actual experimental TM. Coherence gives the maximal colinearity between the columns of a matrix. The lower, the better is the matrix for CS.

technique, compared to the more classical CS imagers based on pseudo-random projections, which have direct control on the TM. However, this calibration step here involves only standard least-squares estimation of the linear mapping between input and output of the scattering material[24,25]. In our experimental setup, the whole calibration is performed in less than 1 minute. While we here rely on optical holography to extract complex amplitude from intensity measurements, the TM measurement can be implemented in a simplified way for other types of waves (RF, acoustics, Terahertz), where direct access to the field amplitude is possible. It may not be so straightforward in practical situations when only the intensity of the output is available, and where more sophisticated methods[30] would be required.

After calibration, the scattering medium can be used to perform CS, using this estimated TM as a measurement matrix. Note that, in our experiment, the same SLM used for calibration is then used as a display to generate the sparse objects. This approach is not restrictive as any sparse optical field or other device capable of modulating light could equivalently be used at this stage. As demonstrated in our results section, using such an estimated TM instead of a perfectly controlled one does yield very good results all the same, while bringing important advantages such as ease of implementation and acquisition speed. Hence, even if the proposed methodology does require the introduction of a supplementary calibration step, this step comes at the cost of a few mandatory supplementary calibration measurements rather than at the cost of performance. This claim is further developed in our results and methods sections.

For a TM to be efficient in a CS setup, it has to correctly scramble the information from all of its inputs to each of its outputs. It is known that a matrix with i.i.d Gaussian entries is an excellent candidate for CS[31] and the TM of optical multiple scattering materials were recently shown to be well approximated by such matrices[22]. The rationale for this fact is that the transmission of light through an opaque lens leads to a very large number of independent scattering events. Even if the total transmission matrix that links the whole input field to the transmitted field shows some non-trivial mesoscopic correlations[32], recent studies proved that these correlations vanish when controlling/measuring only a random partition of input/output channels[22]. In our experimental setup, the number of sensors is very small compared to the total number of output speckle grains and we can hence safely disregard any mesoscopic correlation.

Several previous studies[24,25] have shown on experimental grounds that TMs were close to i.i.d. Gaussians by considering their spectral behavior, i.e. the distribution of their eigenvalues. As a consistency check, we also verified that our experimentally-obtained TMs are close to Gaussian i.i.d., through a complementary study of their *coherence*, which is the maximal correlation between their columns with values between 0 and 1. Among all the features that were proposed to characterize a matrix as a good candidate for CS[33–35], coherence plays a special role because it is easily computed and because a low coherence is sufficient for good recovery performance in CS applications[36–39], even if it is not necessary[40]. In Fig. 3(a), we display one actual TM obtained in our experiments. In Fig. 3(b), we compare its coherence with the one of randomly generated i.i.d. Gaussian matrices. The similar behavior confirms the results and discussions given in[22,24], but also suggests that TMs are good candidates in a CS setup, as will be demonstrated below.

Results and discussion

During our experiments, we measured the reconstruction performance of the imaging system, when the image to reconstruct is composed of $N = 32 \times 32 = 1024$ pixels, using a varying number M of measurements. In practice, we use a CCD array, out of which we select M pixels. These are chosen at random in the array, with an exclusion distance equal to the coherence length of the speckle, in order to ensure uncorrelated measurements. Details of the experiments can be found in the methods section below. For each sparsity level k between 1 and N, a sparse object with only k nonzero coefficients was displayed under $P = 3$ different random phase illuminations [Since our SLM can only do phase modulation, we used a simple trick as in[41] to simulate actual amplitude objects, based on two phase-

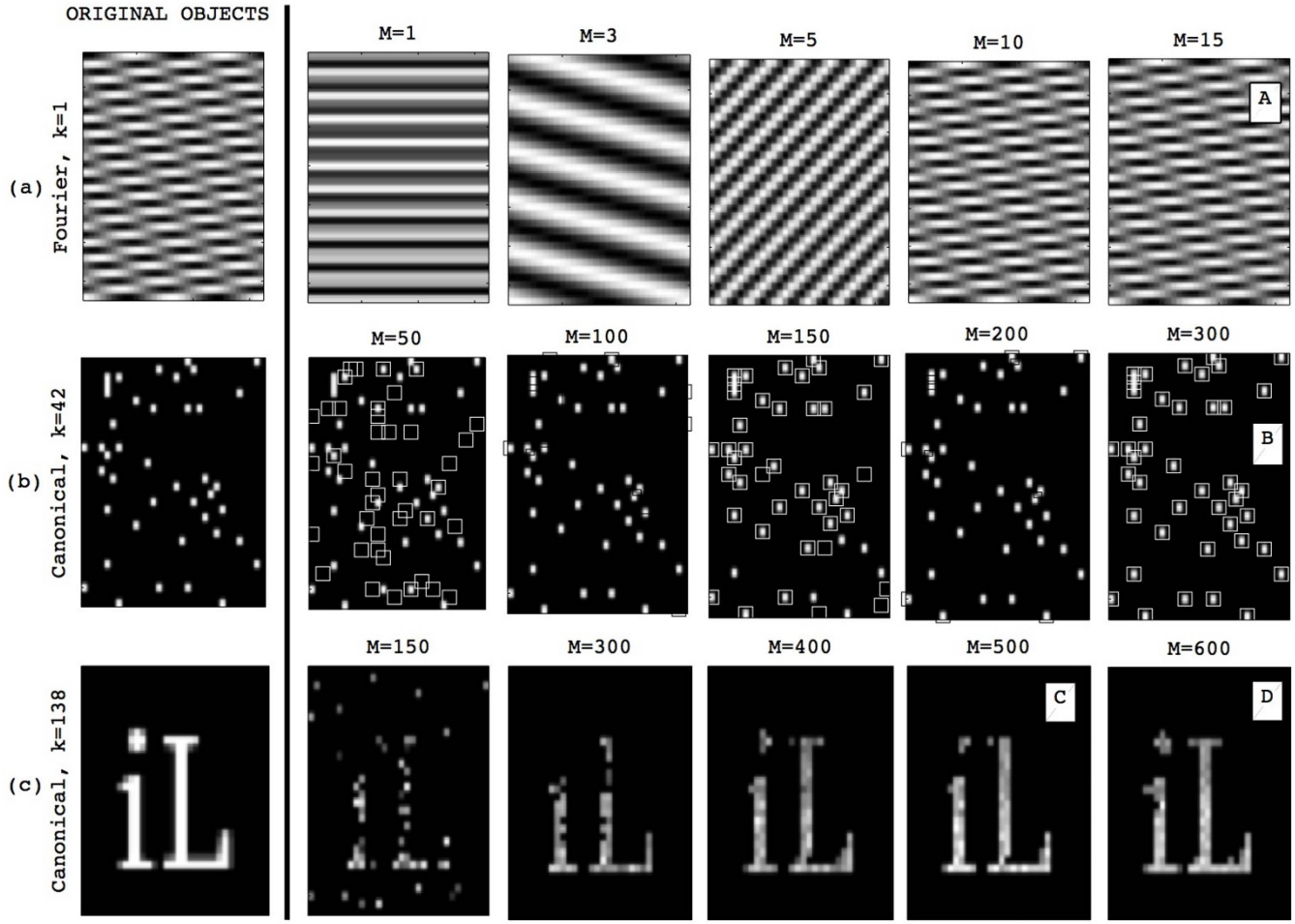


Figure 4 | **Imaging results.** Examples of signals, which are sparse either in the Fourier or canonical domain (left), along with their actual experimental reconstruction using a varying number of measurements. (a) Fourier-sparse object (b–c) canonical sparse objects. In (b), small squares are the original object and large squares are the reconstruction. In all cases, the original object contains 1024 pixels and is thus sampled with a number M of sensors much smaller than N. A, B, C and D images are correspondingly represented in the phase transition diagram of Fig. 5.

modulated measurements. See the supplementary material on this point.]. These virtual measurements may, without loss of generality, be replaced by the use of an amplitude light modulator and are anyways replaced by the actual object to image in a real use-case. The corresponding outputs were then measured and fed into a Multiple Measurement Vector (MMV) sparse recovery algorithm[20]. For each sparsity level, 32 such independent experiments were performed.

Reconstruction of the sparse objects was then achieved numerically using the $M \times P$ measurements only. The TM used for reconstruction is the one estimated in the calibration phase. In order to demonstrate the efficiency and the simplicity of the proposed system, we used the simple Multichannel Orthogonal Matching Pursuit algorithm[42] for MMV reconstruction. It should be noted that more specialized algorithms may lead to better performance and should be considered in the future.

Examples of actual reconstructions performed by our analog compressive sampler are shown on Fig. 4. As can be seen, near-perfect reconstruction of complex sparse patterns occur with sensor density ratios M/N that are much smaller than in classical Shannon-Nyquist sampling ($M = N$). An important feature of the approach is its universality: reconstruction is also efficient for objects that are sparse in the Fourier domain.

The performance of the proposed compressive sampler for all sampling and sparsity rates of interest is summarized on Fig. 5, which is the main result of this paper. It gives the probability of successful reconstruction displayed as a function of the sensor density M/N and relative sparsity k/M. Each point of this surface is the average reconstruction performance for real measurements over approximately 50 independent trials. As can be seen, this *experimental* diagram exhibits a clear "phase transition" from complete failure to systematic success. This thorough experimental study largely confirms that the proposed methodology for sampling using scattering media indeed reaches very competitive sampling rates that are far below the Shannon-Nyquist traditional scheme.

The phase transition observed on Fig. 5 appears to be slightly different from the ones described in the literature[31,43]. The main reason for this fact is that this diagram concerns reconstruction under $P = 3$ Multiple Measurement Vectors (MMV) instead of the classical Single Measurement Vector (SMV) case. This choice, which proves important in practice, is motivated by the fact that MMV is much more robust to noise than SMV[44]. In order to compare our experimental performance to its numerical counterpart, we performed a numerical experiment whose 50% success-rate transition curve is represented by the dashed green line. The transmission matrix is taken as i.i.d Gaussian. The measurement matrix is estimated with the same calibration procedure as in the physical experiment. Each measurement, during calibration and imaging, is contaminated by additive Gaussian noise of variance 3%. Performance obtained in this idealized situation is close to that obtained in our practical setup, for this level of additive noise.

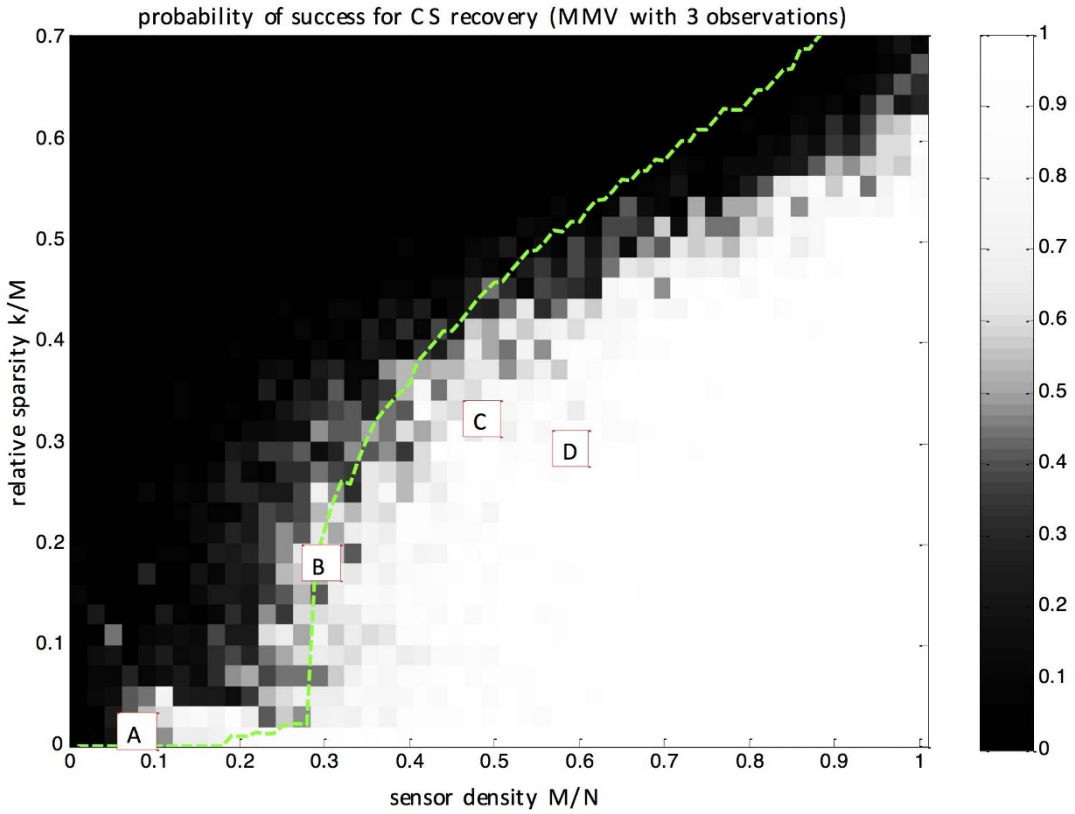


Figure 5 | Probability of success for CS recovery. Experimental probability of successful recovery (between 0 and 1) for a k-sparse image of N pixels via M measurements. On the x-axis is displayed the sensor density ratio M/N. A ratio of 1 corresponds to the Nyquist rate, meaning that all correct reconstructions found in this figure beat traditional sampling. On the y-axis is displayed the relative sparsity ratio k/M. A clear phase transition between failure and success is observable, which is close to that obtained by simulations (dashed line), where exactly the same experimental protocol was conducted with simulated noisy observations both for calibration and imaging. Boxes A, B, C and D locate the corresponding examples of Fig. 4. Each point in this 50×50 grid is the average performance over approximately 50 independent measurements. This figure hence summarizes the results of more than 10^5 actual physical experiments.

Conclusion

In this study, we have demonstrated that a simple natural layer of multiply scattering material can be used to successfully perform compressive sensing. The compressive imager relies on scattering theory to optimally dispatch information from the object to all measurement sensors, shifting the complexity of devising CS hardware from the design, fabrication and electronic control to a simple calibration procedure.

As in any hardware implementation of CS, experimental noise is an important issue limiting the performance, especially since it impacts the measurement matrix. Using baseline sparse reconstruction algorithms along with standard least-squares calibration techniques, we demonstrated that successful reconstruction exhibits a clear phase transition between failure and success even at very competitive sampling rates. The proposed methodology can be considered to be a truly analog compressive sampler and as such, benefits from both theoretical elegance and ease of implementation.

The imaging system we introduced has many advantageous features. First, it enables the implementation of an extremely flat imaging device with few detectors. Second, this imaging methodology can be implemented in practice with very few conventional lenses, as in[45] for instance. This is a strong point for implementation in domains outside optics where it is hard to fabricate lenses. Indeed, the concept presented here can directly be used in other domains of optics such as holography, but also in other disciplines such as THz, RF or ultrasound imaging. Third, similarly to recent work on metamaterials apertures, non-resonant scattering materials work over a wide frequency range and have a strongly frequency-dependent response. Fourth, unlike most current compressive sensing hardware, this system gives access to many compressive measurements in a parallel fashion, potentially speeding up acquisition. These advantages come at the simple cost of a calibration step, which amounts to estimate the Transmission Matrix of the scattering material considered. As we demonstrated, this can be achieved by simple input/output mapping techniques such as linear least-squares and needs to be done only once.

While conventional direct imaging can be thought as an embarrassingly parallel process that does not exploit the structure of the scene, in contrast most current CS hardware (such as the single pixel camera) require a heavily sequential process that does take into account the structure of the scene. Our approach borrows from the best of both acquisition processes, in that it is both embarrassingly parallel and takes into account the structure of the scene.

1. Elad, M. *Sparse and redundant representations: from theory to applications in signal and image processing.* (Springer, 2010).
2. Candès, E. J. Compressive sampling. *Proceedings of the International Congress of Mathematicians: Madrid, August 22–30, 2006: invited lectures.* 2006.
3. Candès, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE T. Inform. Theory* **52**, 489 (2006).
4. Donoho, D. L. Compressed sensing. *IEEE T. Inform. Theory* **52**, 1289 (2006).
5. Chan, W. L. *et al.* A single-pixel terahertz imaging system based on compressed sensing. *Appl. Phys. Lett.* **93**, 12105 (2008).
6. Duarte, M. F. *et al.* Single-pixel imaging via compressive sampling. *IEEE T. Signal Process* **25**, 83–91 (2008).
7. Fergus, R., Torralba, A. & Freeman, W. T. Random lens imaging (Tech. Rep. Massachusetts Institute of Technology, 2006).

8. Hunt, J. *et al*. Metamaterial Apertures for Computational Imaging. *Science* **339**, 310–313 (2013).
9. Jacques, L. *et al*. CMOS compressed imaging by random convolution. *Int. Conf. Acoust. Spee.* 1113–1116 (2009).
10. Katz, O. *et al*. Compressive ghost imaging. *Appl. Phys. Lett.* **95**, 131110 (2009).
11. Levin, A. *et al*. Image and depth from a conventional camera with a coded aperture. *ACM T. Graphics* **26**, 70 (2007).
12. Potuluri, P., Xu, M. & Brady, D. Imaging with random 3D reference structures. *Opt. Express.* **11**, 2134–2141 (2003).
13. Zhao, C. *et al*. Ghost imaging lidar via sparsity constraints. *Appl. Phys. Lett.* **101**, 141123 (2012).
14. Shrekenhamer, D., Watts, C. M. & Padilla, W. J. Terahertz single pixel imaging with an optically controlled dynamic spatial light modulator. *Opt. Express* **21**, 12507–12518 (2013).
15. Chen, H. T. *et al*. Active terahertz metamaterial devices. *Nature* **444**, 597–600 (2006).
16. Tyson, R. *Principles of adaptive optics.* (CRC Press, 2010).
17. Vellekoop, I. M. & Mosk, A. P. Focusing coherent light through opaque strongly scattering media. *Opt. lett* **32**, 2309–2311 (2007).
18. Vellekoop, I. M., Lagendijk, A. & Mosk, A. P. Exploiting disorder for perfect focusing. *Nat. Photon.* **4**, 320–322 (2010).
19. Pappu, R., Recht, R., Taylor, J. & Gershenfeld, N. Physical one-way functions. *Science* **297**, 2026–2030 (2002).
20. Cotter, S. F. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE T. Signal Process* **53**, 2477–2488 (2005).
21. Li, L. & Li, F. Compressive Sensing Based Robust Signal Sampling. *Appl. Phys. Research* **4**, 30 (2012).
22. Goetschy, A. & Stone, A. D. Filtering Random Matrices: The Effect of Incomplete Channel Control in Multiple Scattering. *Phys. Rev. Lett.* **111**, 063901 (2013).
23. Gribonval, R., Chardon, G. & Daudet, L. Blind calibration for compressed sensing by convex optimization. *Int. Conf. Acoust. Spee.* 2713–2716 (2012).
24. Popoff, S. M. *et al*. Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media. *Phys. Rev. Lett.* **104**, 100601 (2010).
25. Popoff, S. M. *et al*. Controlling light through optical disordered media: transmission matrix approach. *New J. Phys.* **13**, 123021 (2011).
26. Eldar, Y. C. & Kutyniok, G. eds. *Compressed sensing: theory and applications.* (Cambridge University Press, 2012).
27. Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Process.* **25**, 21–30 (2008).
28. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B Met.* 267–288 (1996).
29. Goodman, J. W. *Introduction to Fourier optics.* (Roberts and Company Publishers, 2005).
30. Ohlsson, H. *et al*. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv, 1111.6323* (2011).
31. Donoho, D. & Tanner, J. Precise undersampling theorems. *Proc. IEEE* **98**, 913–924 (2010).
32. Akkermans, E. *Mesoscopic physics of electrons and photons* (Cambridge University Press, 2007).
33. Candès, E. J. The restricted isometry property and its implications for compressed sensing. *C. R. Math.* **346**, 589–592 (2008).
34. Cohen, A., Dahmen, W. & DeVore, R. Compressed sensing and best k-term approximation. *J. Am. Math. Soc.* **22**, 211–231 (2009).
35. Yin, W. & Zhang, Y. Extracting salient features from less data via l1-minimization. *SIAG/OPT Views-and-News* **19**, 11–19 (2008).
36. Donoho, D. L. & Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE T. Inf. Theory* **47**, 2845–2862 (2001).
37. Tropp, J. A. Greed is good: Algorithmic results for sparse approximation. *IEEE T. Inf. Theory* **50**, 2231–2242 (2004).
38. Tropp, J. A. & Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE T. Inf. Theory* **53**, 4655–4666 (2007).
39. Wang, J. & Shim, B. A Simple Proof of the Mutual Incoherence Condition for Orthogonal Matching Pursuit. *arXiv, 1105.4408* (2011).
40. Candès, E. J. *et al*. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. A.* **31**, 59–73 (2011).
41. Popoff, S. M. *et al*. "Image transmission through an opaque material." *Nat. Comms.* **1**, 81 (2010).
42. Gribonval, R. *et al*. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Four. Anal. Appl.* **14**, 655–687 (2008).
43. Donoho, D. & Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. T Roy. S. A* **367**, 4273–4293 (2009).
44. Eldar, Y. & Rauhut, H. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE T. inform. Theory* **56**, 505–519 (2010).
45. Katz, O. *et al*. Non-invasive real-time imaging through scattering layers and around corners via speckle correlations. *arXiv, 1403.3316* (2014).

Acknowledgments

This work was supported by the European Research Council (Grant N°278025), the Emergence(s) program from the City of Paris, and LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under references ANR-10-LABX-24 and ANR-10-IDEX-0001-02 PSL*. G.C. is supported by the Austrian Science Fund (FWF) START-project FLAME (Y 551-N13). O.K. is supported by the Marie Curie intra-European fellowship for career development (IEF) and the Rothschild fellowship. I.C. would like to thank the Physics arXiv Blog for drawing his attention to opaque lenses and Ms. Iris Carron for her typesetting support.

Author contributions

L.D., S.G., I.C. proposed the use of a multiply scattering material for compressive sensing. S.P., G.L. and S.G. designed the initial experimental setup. G.C. performed initial numerical analysis. D.M., O.K. and S.G. discussed the experimental implementation. D.M. and A.L. performed the experiments and A.L. performed the numerical analysis with the help of L.D. All authors contributed to discussing the results and writing the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Liutkus, A. *et al*. Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium. *Sci. Rep.* **4**, 5552; DOI:10.1038/srep05552 (2014).

Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions

Antoine Liutkus [1] **Umut Şimşekli** [2] **Szymon Majewski** [3] **Alain Durmus** [4] **Fabian-Robert Stöter** [1]

Abstract

By building upon the recent theory that established the connection between implicit generative modeling (IGM) and optimal transport, in this study, we propose a novel parameter-free algorithm for learning the underlying distributions of complicated datasets and sampling from them. The proposed algorithm is based on a functional optimization problem, which aims at finding a measure that is close to the data distribution as much as possible and also expressive enough for generative modeling purposes. We formulate the problem as a gradient flow in the space of probability measures. The connections between gradient flows and stochastic differential equations let us develop a computationally efficient algorithm for solving the optimization problem. We provide formal theoretical analysis where we prove finite-time error guarantees for the proposed algorithm. To the best of our knowledge, the proposed algorithm is the first nonparametric IGM algorithm with explicit theoretical guarantees. Our experimental results support our theory and show that our algorithm is able to successfully capture the structure of different types of data distributions.

1. Introduction

Implicit generative modeling (IGM) (Diggle & Gratton, 1984; Mohamed & Lakshminarayanan, 2016) has become very popular recently and has proven successful in various fields; variational auto-encoders (VAE) (Kingma & Welling, 2013) and generative adversarial networks (GAN) (Goodfellow et al., 2014) being its two well-known examples. The goal in IGM can be briefly described as learning the underlying probability measure of a given dataset, denoted as $\nu \in \mathcal{P}(\Omega)$, where $\mathcal{P}$ is the space of probability measures on the measurable space $(\Omega, \mathcal{A})$, $\Omega \subset \mathbb{R}^d$ is a domain and $\mathcal{A}$ is the associated Borel σ-field.

Given a set of data points $\{y_1, \dots, y_P\}$ that are assumed to be independent and identically distributed (i.i.d.) samples drawn from ν, the implicit generative framework models them as the output of a measurable map, i.e. $y = T(x)$, with $T : \Omega_\mu \mapsto \Omega$. Here, the inputs x are generated from a known and easy to sample source measure μ on Ω_μ (e.g. Gaussian or uniform measures), and the outputs $T(x)$ should match the unknown target measure ν on Ω.

Learning generative networks have witnessed several groundbreaking contributions in recent years. Motivated by this fact, there has been an interest in illuminating the theoretical foundations of VAEs and GANs (Bousquet et al., 2017; Liu et al., 2017). It has been shown that these implicit models have close connections with the theory of Optimal Transport (OT) (Villani, 2008). As it turns out, OT brings new light on the generative modeling problem: there have been several extensions of VAEs (Tolstikhin et al., 2017; Kolouri et al., 2018) and GANs (Arjovsky et al., 2017; Gulrajani et al., 2017; Guo et al., 2017; Lei et al., 2017), which exploit the links between OT and IGM.

OT studies whether it is possible to transform samples from a source distribution μ to a target distribution ν. From this perspective, an ideal generative model is simply a transport map from μ to ν. This can be written by using some 'push-forward operators': we seek a mapping T that 'pushes μ onto ν', and is formally defined as $\nu(A) = \mu(T^{-1}(A))$ for all Borel sets $A \subset \mathcal{A}$. If this relation holds, we denote the push-forward operator $T_\#$, such that $T_\#\mu = \nu$. Provided mild conditions on these distributions hold (notably μ is non-atomic (Villani, 2008)), existence of such a transport map is guaranteed; however, it remains a challenge to construct it in practice.

One common point between VAE and GAN is to adopt an approximate strategy and consider transport maps that

[1]Inria and LIRMM, Univ. of Montpellier, France [2]LTCI, Télécom Paristech, Université Paris-Saclay, Paris, France [3]Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland [4]CNRS, ENS Paris-Saclay,Universit Paris-Saclay, Cachan, France. Correspondence to: Antoine Liutkus <antoine.liutkus@inria.fr>, Umut Şimşekli <umut.simsekli@telecom-paristech.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

belong to a *parametric* family T_ϕ with $\phi \in \Phi$. Then, they aim at finding the best parameter $\phi^\star$ that would give $T_{\phi^\star}\#\mu \approx \nu$. This is typically achieved by attempting to minimize the following optimization problem: $\phi^\star = \arg\min_{\phi\in\Phi} \mathcal{W}_2(T_{\phi}\#\mu, \nu)$, where $\mathcal{W}_2$ denotes the Wasserstein distance that will be properly defined in Section 2. It has been shown that (Genevay et al., 2017) OT-based GANs (Arjovsky et al., 2017) and VAEs (Tolstikhin et al., 2017) both use this formulation with different parameterizations and different equivalent definitions of $\mathcal{W}_2$. However, their resulting algorithms still lack theoretical understanding.

In this study, we follow a completely different approach for IGM, where we aim at developing an algorithm with explicit theoretical guarantees for estimating a transport map between source μ and target ν. The generated transport map will be *nonparametric* (in the sense that it does not belong to some family of functions, like a neural network), and it will be iteratively augmented: always increasing the quality of the fit along iterations. Formally, we take T_t as the constructed transport map at time $t \in [0, \infty)$, and define $\mu_t = T_t\#\mu$ as the corresponding output distribution. Our objective is to build the maps so that μ_t will converge to the solution of a functional optimization problem, defined through a gradient flow in the Wasserstein space. Informally, we will consider a gradient flow that has the following form:

$$\partial_t \mu_t = -\nabla_{\mathcal{W}_2}\Big\{\mathrm{Cost}(\mu_t, \nu) + \mathrm{Reg}(\mu_t)\Big\}, \quad \mu_0 = \mu, \quad (1)$$

where the functional Cost computes a discrepancy between μ_t and ν, Reg denotes a regularization functional, and $\nabla_{\mathcal{W}_2}$ denotes a notion of gradient with respect to a probability measure in the $\mathcal{W}_2$ metric for probability measures[1]. If this flow can be simulated, one would hope for $\mu_t = (T_t)_\#\mu$ to converge to the minimum of the functional optimization problem: $\min_\mu(\mathrm{Cost}(\mu, \nu) + \mathrm{Reg}(\mu))$ (Ambrosio et al., 2008; Santambrogio, 2017).

We construct a gradient flow where we choose the Cost functional as the *sliced Wasserstein distance* ($\mathcal{SW}_2$) (Rabin et al., 2012; Bonneel et al., 2015) and the Reg functional as the negative entropy. The $\mathcal{SW}_2$ distance is equivalent to the $\mathcal{W}_2$ distance (Bonnotte, 2013) and has important computational implications since it can be expressed as an average of (one-dimensional) projected optimal transportation costs whose analytical expressions are available.

We first show that, with the choice of $\mathcal{SW}_2$ and the negative-entropy functionals as the overall objective, we obtain a valid gradient flow that has a solution path $(\mu_t)_t$, and the probability density functions of this path solve a particular partial differential equation, which has close connections with stochastic differential equations. Even though gradient flows in Wasserstein spaces cannot be solved in general, by exploiting this connection, we are able to develop a practical algorithm that provides approximate solutions to the gradient flow and is algorithmically similar to stochastic gradient Markov Chain Monte Carlo (MCMC) methods[2] (Welling & Teh, 2011; Ma et al., 2015; Durmus et al., 2016; Şimşekli, 2017; Şimşekli et al., 2018). We provide finite-time error guarantees for the proposed algorithm and show explicit dependence of the error to the algorithm parameters.

To the best of our knowledge, the proposed algorithm is the first nonparametric IGM algorithm that has explicit theoretical guarantees. In addition to its nice theoretical properties, the proposed algorithm has also significant practical importance: it has low computational requirements and can be easily run on an everyday laptop CPU.Our experiments on both synthetic and real datasets support our theory and illustrate the advantages of the algorithm in several scenarios.

2. Technical Background

2.1. Wasserstein distance, optimal transport maps and Kantorovich potentials

For two probability measures $\mu, \nu \in \mathcal{P}_2(\Omega)$, $\mathcal{P}_2(\Omega) = \{\mu \in \mathcal{P}(\Omega) : \int_\Omega \|x\|^2 \mu(\mathrm{d}x) < +\infty\}$, the 2-Wasserstein distance is defined as follows:

$$\mathcal{W}_2(\mu, \nu) \triangleq \Big\{\inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int_{\Omega\times\Omega} \|x - y\|^2 \gamma(dx, dy)\Big\}^{1/2}, \quad (2)$$

where $\mathcal{C}(\mu, \nu)$ is called the set of *transportation plans* and defined as the set of probability measures γ on $\Omega\times\Omega$ satisfying for all $A \in \mathcal{A}$, $\gamma(A\times\Omega) = \mu(A)$ and $\gamma(\Omega\times A) = \nu(A)$, i.e. the marginals of γ coincide with μ and ν. From now on, we will assume that Ω is a compact subset of $\mathbb{R}^d$.

In the case where Ω is finite, computing the Wasserstein distance between two probability measures turns out to be a linear program with linear constraints, and has therefore a dual formulation. Since Ω is a Polish space (i.e. a complete and separable metric space), this dual formulation can be generalized as follows (Villani, 2008)[Theorem 5.10]:

$$\mathcal{W}_2(\mu, \nu) = \sup_{\psi\in \mathrm{L}^1(\mu)} \Big\{\int_\Omega \psi(x)\mu(dx) + \int_\Omega \psi^c(x)\nu(dx)\Big\}^{1/2} \quad (3)$$

where $\mathrm{L}^1(\mu)$ denotes the class of functions that are absolutely integrable under μ and ψ^c denotes the c-conjugate of ψ and is defined as follows: $\psi^c(y) \triangleq \{\inf_{x\in\Omega} \|x -$

[1] This gradient flow is similar to the usual Euclidean gradient flows, i.e. $\partial_t x_t = -\nabla(f(x_t) + r(x_t))$, where f is typically the data-dependent cost function and r is a regularization term. The (explicit) Euler discretization of this flow results in the well-known gradient descent algorithm for solving $\min_x(f(x) + r(x))$.

[2] We note that, despite the algorithmic similarities, the proposed algorithm is not a Bayesian posterior sampling algorithm.

$y\|^2 - \psi(x)\}$. The functions ψ that realize the supremum in (3) are called the Kantorovich potentials between μ and ν. Provided that μ satisfies a mild condition, we have the following uniqueness result.

Theorem 1 ((Santambrogio, 2014)[Theorem 1.4]). *Assume that $\mu \in \mathcal{P}_2(\Omega)$ is absolutely continuous with respect to the Lebesgue measure. Then, there exists a unique optimal transport plan $\gamma^\star$ that realizes the infimum in (2) and it is of the form $(Id \times T)_{\#}\mu$, for a measurable function $T : \Omega \to \Omega$. Furthermore, there exists at least a Kantorovich potential ψ whose gradient $\nabla\psi$ is uniquely determined μ-almost everywhere. The function T and the potential ψ are linked by $T(x) = x - \nabla\psi(x)$.*

The measurable function $T : \Omega \to \Omega$ is referred to as the optimal transport map from μ to ν. This result implies that there exists a solution for transporting samples from μ to samples from ν and this solution is optimal in the sense that it minimizes the ℓ_2 displacement. However, identifying this solution is highly non-trivial. In the discrete case, effective solutions have been proposed (Cuturi, 2013). However, for continuous and high-dimensional probability measures, constructing an actual transport plan remains a challenge. Even if recent contributions (Genevay et al., 2016) have made it possible to rapidly compute $\mathcal{W}_2$, they do so without constructing the optimal map T, which is our objective here.

2.2. Wasserstein spaces and gradient flows

By (Ambrosio et al., 2008)[Proposition 7.1.5], $\mathcal{W}_2$ is a distance over $\mathcal{P}(\Omega)$. In addition, if $\Omega \subset \mathbb{R}^d$ is compact, the topology associated with $\mathcal{W}_2$ is equivalent to the weak convergence of probability measures and $(\mathcal{P}(\Omega), \mathcal{W}_2)$[3] is compact. The metric space $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$ is called the *Wasserstein space*.

In this study, we are interested in functional optimization problems in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$, such as $\min_{\mu\in\mathcal{P}_2(\Omega)} \mathcal{F}(\mu)$, where $\mathcal{F}$ is the functional that we would like to minimize. Similar to Euclidean spaces, one way to formulate this optimization problem is to construct a gradient flow of the form $\partial_t\mu_t = -\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$ (Benamou & Brenier, 2000; Lavenant et al., 2018), where $\nabla_{\mathcal{W}_2}$ denotes a notion of gradient in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$. If such a flow can be constructed, one can utilize it both for practical algorithms and theoretical analysis.

Gradient flows $\partial_t\mu_t = \nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$ with respect to a functional $\mathcal{F}$ in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$ have strong connections with partial differential equations (PDE) that are of the form of a *continuity equation* (Santambrogio, 2017). Indeed, it is shown than under appropriate conditions on $\mathcal{F}$ (see e.g.(Ambrosio et al., 2008)), $(\mu_t)_t$ is a solution of the gradient flow if and only if it admits a density ρ_t with respect to the Lebesgue measure for all $t \geq 0$, and solves the continuity equation given by: $\partial_t\rho_t + \operatorname{div}(v\rho_t) = 0$, where v denotes a vector field and div denotes the divergence operator. Then, for a given gradient flow in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$, we are interested in the evolution of the densities ρ_t, i.e. the PDEs which they solve. Such PDEs are of our particular interest since they have a key role for building practical algorithms.

[3]Note that in that case, $\mathcal{P}_2(\Omega) = \mathcal{P}(\Omega)$

2.3. Sliced-Wasserstein distance

In the one-dimensional case, i.e. $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, $\mathcal{W}_2$ has an analytical form, given as follows: $\mathcal{W}_2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(\tau) - F_\nu^{-1}(\tau)|^2 \, d\tau$, where F_μ and F_ν denote the cumulative distribution functions (CDF) of μ and ν, respectively, and F_μ^{-1}, F_ν^{-1} denote the inverse CDFs, also called quantile functions (QF). In this case, the optimal transport map from μ to ν has a closed-form formula as well, given as follows: $T(x) = (F_\nu^{-1} \circ F_\mu)(x)$ (Villani, 2008). The optimal map T is also known as the *increasing arrangement*, which maps each quantile of μ to the same quantile of ν, e.g. minimum to minimum, median to median, maximum to maximum (Villani, 2008). Due to Theorem 1, the derivative of the corresponding Kantorovich potential is given as:

$$\psi'(x) \triangleq \partial_x\psi(x) = x - (F_\nu^{-1} \circ F_\mu)(x).$$

In the multidimensional case $d > 1$, building a transport map is much more difficult. The nice properties of the one-dimensional Wasserstein distance motivate the usage of *sliced-Wasserstein distance* ($\mathcal{SW}_2$) for practical applications. Before formally defining $\mathcal{SW}_2$, let us first define the orthogonal projection $\theta^*(x) \triangleq \langle\theta, x\rangle$ for any direction $\theta \in \mathbb{S}^{d-1}$ and $x \in \mathbb{R}^d$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner-product and $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ denotes the d-dimensional unit sphere. Then, the $\mathcal{SW}_2$ distance is formally defined as follows:

$$\mathcal{SW}_2(\mu, \nu) \triangleq \int_{\mathbb{S}^{d-1}} \mathcal{W}_2(\theta^*_{\#}\mu, \theta^*_{\#}\nu) \, d\theta, \qquad (4)$$

where $d\theta$ represents the uniform probability measure on $\mathbb{S}^{d-1}$. As shown in (Bonnotte, 2013), $\mathcal{SW}_2$ is indeed a distance metric and induces the same topology as $\mathcal{W}_2$ for compact domains.

The $\mathcal{SW}_2$ distance has important practical implications: provided that the projected distributions $\theta^*_{\#}\mu$ and $\theta^*_{\#}\nu$ can be computed, then for any $\theta \in \mathbb{S}^{d-1}$, the distance $\mathcal{W}_2(\theta^*_{\#}\mu, \theta^*_{\#}\nu)$, as well as its optimal transport map and the corresponding Kantorovich potential can be analytically computed (since the projected measures are one-dimensional). Therefore, one can easily approximate (4) by using a simple Monte Carlo scheme that draws uniform random samples from $\mathbb{S}^{d-1}$ and replaces the integral in (4) with a finite-sample average. Thanks to its computational benefits, $\mathcal{SW}_2$ was very recently considered for OT-based VAEs and GANs (Deshpande et al., 2018; Wu et al., 2018;

Kolouri et al., 2018), appearing as a stable alternative to the adversarial methods.

3. Regularized Sliced-Wasserstein Flows for Generative Modeling

3.1. Construction of the gradient flow

In this paper, we propose the following functional minimization problem on $\mathcal{P}_2(\Omega)$ for implicit generative modeling:

$$\min_{\mu}\Big\{\mathcal{F}_\lambda^\nu(\mu) \triangleq \frac{1}{2}\mathcal{SW}_2^2(\mu,\nu) + \lambda\mathcal{H}(\mu)\Big\}, \qquad (5)$$

where $\lambda > 0$ is a regularization parameter and $\mathcal{H}$ denotes the negative entropy defined by $\mathcal{H}(\mu) \triangleq \int_\Omega \rho(x)\log\rho(x)dx$ if μ has density ρ with respect to the Lebesgue measure and $\mathcal{H}(\mu) = +\infty$ otherwise. Note that the case $\lambda = 0$ has been already proposed and studied in (Bonnotte, 2013) in a more general OT context. Here, in order to introduce the necessary noise inherent to generative model, we suggest to penalize the slice-Wasserstein distance using $\mathcal{H}$. In other words, the main idea is to find a measure $\mu^\star$ that is close to ν as much as possible and also has a certain amount of entropy to make sure that it is sufficiently expressive for generative modeling purposes. The importance of the entropy regularization becomes prominent in practical applications where we have finitely many data samples that are assumed to be drawn from ν. In such a circumstance, the regularization would prevent $\mu^\star$ to collapse on the data points and therefore avoid 'over-fitting' to the data distribution. Note that this regularization is fundamentally different from the one used in Sinkhorn distances (Genevay et al., 2018).

In our first result, we show that there exists a flow $(\mu_t)_{t\geq 0}$ in $(\mathcal{P}(\overline{\mathrm{B}}(0,r)), \mathcal{W}_2)$ which decreases along $\mathcal{F}_\lambda^\nu$, where $\overline{\mathrm{B}}(0,a)$ denotes the closed unit ball centered at 0 and radius a. This flow will be referred to as a generalized minimizing movement scheme (see Definition 1 in the supplementary document). In addition, the flow $(\mu_t)_{t\geq 0}$ admits a density ρ_t with respect to the Lebesgue measure for all $t > 0$ and $(\rho_t)_{t\geq 0}$ is solution of a non-linear PDE (in the weak sense).

Theorem 2. *Let ν be a probability measure on $\overline{\mathrm{B}}(0,1)$ with a strictly positive smooth density. Choose a regularization constant $\lambda > 0$ and radius $r > \sqrt{d}$, where d is the data dimension. Assume that $\mu_0 \in \mathcal{P}(\overline{\mathrm{B}}(0,r))$ is absolutely continuous with respect to the Lebesgue measure with density $\rho_0 \in \mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$. There exists a generalized minimizing movement scheme $(\mu_t)_{t\geq 0}$ associated to (5) and if ρ_t stands for the density of μ_t for all $t \geq 0$, then $(\rho_t)_t$ satisfies the following continuity equation:*

$$\frac{\partial\rho_t}{\partial t} = -\operatorname{div}(v_t\rho_t) + \lambda\Delta\rho_t, \qquad (6)$$

$$v_t(x) \triangleq v(x,\mu_t) = -\int_{\mathbb{S}^{d-1}} \psi'_{t,\theta}(\langle x,\theta\rangle)\theta d\theta \qquad (7)$$

in a weak sense. Here, Δ denotes the Laplacian operator, div the divergence operator, and $\psi_{t,\theta}$ denotes the Kantorovich potential between $\theta^*_\#\mu_t$ and $\theta^*_\#\nu$.

The precise statement of this Theorem, related results and its proof are postponed to the supplementary document. For its proof, we use the technique introduced in (Jordan et al., 1998): we first prove the existence of a generalized minimizing movement scheme by showing that the solution curve $(\mu_t)_t$ is a limit of the solution of a time-discretized problem. Then we prove that the curve $(\rho_t)_t$ solves the PDE given in (6).

3.2. Connection with stochastic differential equations

As a consequence of the entropy regularization, we obtain the Laplacian operator Δ in the PDE given in (6). We therefore observe that the overall PDE is a Fokker-Planck-type equation (Bogachev et al., 2015) that has a well-known probabilistic counterpart, which can be expressed as a stochastic differential equation (SDE). More precisely, let us consider a stochastic process $(X_t)_t$, that is the solution of the following SDE starting at $X_0 \sim \mu_0$:

$$dX_t = v(X_t,\mu_t)dt + \sqrt{2\lambda}dW_t, \qquad (8)$$

where $(W_t)_t$ denotes a standard Brownian motion. Then, the probability distribution of X_t at time t solves the PDE given in (6) (Bogachev et al., 2015). This informally means that, if we could simulate (8), then the distribution of X_t would converge to the solution of (5), therefore, we could use the sample paths $(X_t)_t$ as samples drawn from $(\mu_t)_t$. However, in practice this is not possible due to two reasons: (i) the drift v_t cannot be computed analytically since it depends on the probability distribution of X_t, (ii) the SDE (8) is a continuous-time process, it needs to be discretized.

We now focus on the first issue. We observe that the SDE (8) is similar to McKean-Vlasov SDEs (Veretennikov, 2006; Mishura & Veretennikov, 2016), a family of SDEs whose drift depends on the distribution of X_t. By using this connection, we can borrow tools from the relevant SDE literature (Malrieu, 2003; Cattiaux et al., 2008) for developing an approximate simulation method for (8).

Our approach is based on defining a *particle system* that serves as an approximation to the original SDE (8). The particle system can be written as a collection of SDEs, given as follows (Bossy & Talay, 1997):

$$dX_t^i = v(X_t^i,\mu_t^N)dt + \sqrt{2\lambda}dW_t^i\,, \quad i = 1,\dots,N, \qquad (9)$$

where i denotes the particle index, $N \in \mathbb{N}_+$ denotes the total number of particles, and $\mu_t^N = (1/N)\sum_{j=1}^N \delta_{X_t^j}$ denotes the empirical distribution of the particles $\{X_t^j\}_{j=1}^N$. This particle system is particularly interesting, since (i) one

typically has $\lim_{N\to\infty} \mu_t^N = \mu_t$ with a rate of convergence of order $\mathcal{O}(1/\sqrt{N})$ for all t (Malrieu, 2003; Cattiaux et al., 2008), and (ii) each of the particle systems in (9) can be simulated by using an Euler-Maruyama discretization scheme. We note that the existing theoretical results in (Veretennikov, 2006; Mishura & Veretennikov, 2016) do not directly apply to our case due to the non-standard form of our drift. However, we conjecture that a similar result holds for our problem as well. Such a result would be proven by using the techniques given in (Zhang et al., 2018); however, it is out of the scope of this study.

3.3. Approximate Euler-Maruyama discretization

In order to be able to simulate the particle SDEs (9) in practice, we propose an approximate Euler-Maruyama discretization for each particle SDE. The algorithm iteratively applies the following update equation: $(\forall i \in \{1,\dots,N\})$

$$\bar{X}_0^i \overset{\text{i.i.d.}}{\sim} \mu_0, \quad \bar{X}_{k+1}^i = \bar{X}_k^i + h\hat{v}_k(\bar{X}_k^i) + \sqrt{2\lambda h} Z_{k+1}^i, \quad (10)$$

where $k \in \mathbb{N}_+$ denotes the iteration number, Z_k^i is a standard Gaussian random vector in $\mathbb{R}^d$, h denotes the step-size, and $\hat{v}_k$ is a short-hand notation for a computationally tractable estimator of the original drift $v(\cdot, \bar{\mu}_{kh}^N)$, with $\bar{\mu}_{kh}^N = (1/N)\sum_{j=1}^N \delta_{\bar{X}_k^j}$ being the empirical distribution of $\{\bar{X}_k^j\}_{j=1}^N$. A question of fundamental practical importance is how to compute this function $\hat{v}$.

We propose to approximate the integral in (7) via a simple Monte Carlo estimate. This is done by first drawing N_θ uniform i.i.d. samples from the sphere $\mathbb{S}^{d-1}$, $\{\theta_n\}_{n=1}^{N_\theta}$. Then, at each iteration k, we compute:

$$\hat{v}_k(x) \triangleq -(1/N_\theta)\sum_{n=1}^{N_\theta} \psi'_{k,\theta_n}(\langle\theta_n, x\rangle)\theta_n, \quad (11)$$

where for any θ, $\psi'_{k,\theta}$ is the derivative of the Kantorovich potential (cf. Section 2) that is applied to the OT problem from $\theta^*_\#\bar{\mu}_{kh}^N$ to $\theta^*_\#\nu$: i.e.

$$\psi'_{k,\theta}(z) = \left[z - (F^{-1}_{\theta^*_\#\nu} \circ F_{\theta^*_\#\bar{\mu}_{kh}^N})(z)\right]. \quad (12)$$

For any particular $\theta \in \mathbb{S}^{d-1}$, the QF, $F^{-1}_{\theta^*_\#\nu}$ for the projection of the target distribution ν on θ can be easily computed from the data. This is done by first computing the projections $\langle\theta, y_i\rangle$ for all data points y_i, and then computing the empirical quantile function for this set of P scalars. Similarly, $F_{\theta^*_\#\bar{\mu}_{kh}^N}$, the CDF of the particles at iteration k, is easy to compute: we first project all particles $\bar{X}_k^i$ to get $\langle\theta, \bar{X}_k^i\rangle$, and then compute the empirical CDF of this set of N scalar values.

In both cases, the true CDF and quantile functions are approximated as a linear interpolation between a set of the computed $Q \in \mathbb{N}_+$ empirical quantiles. Another source of approximation here comes from the fact that the target ν will in practice be a collection of Dirac measures on the observations y_i. Since it is currently common to have a very large dataset, we believe this approximation to be accurate in practice for the target. Finally, yet another source of approximation comes from the error induced by using a finite number of θ_n instead of a sum over $\mathbb{S}^{d-1}$ in (12).

Even though the error induced by these approximation schemes can be incorporated into our current analysis framework, we choose to neglect it for now, because (i) all of these one-dimensional computations can be done very accurately and (ii) the quantization of the empirical CDF and QF can be modeled as additive Gaussian noise that enters our discretization scheme (10) (Van der Vaart, 1998). Therefore, we will assume that $\hat{v}_k$ is an *unbiased* estimator of v, i.e. $\mathbb{E}[\hat{v}(x,\mu)] = v(x,\mu)$, for any x and μ, where the expectation is taken over θ_n.

The overall algorithm is illustrated in Algorithm 1. It is remarkable that the updates of the particles only involves the learning data $\{y_i\}$ through the CDFs of its projections on the many $\theta_n \in \mathbb{S}^{d-1}$. This has a fundamental consequence of high practical interest: these CDF may be computed beforehand in a massively distributed manner that is independent of the sliced Wasserstein flow. This aspect is reminiscent of the *compressive learning* methodology (Gribonval et al., 2017), except we exploit quantiles of random projections here, instead of random generalized moments as done there.

Besides, we can obtain further reductions in the computing time if the CDF, $F_{\theta^*_\#\nu}$ for the target is computed on random

Algorithm 1: Sliced-Wasserstein Flow (SWF)

input : $\mathcal{D} \equiv \{y_i\}_{i=1}^P$, μ_0, N, N_θ, h, λ
output : $\{\bar{X}_K^i\}_{i=1}^N$

// Initialize the particles
$\bar{X}_0^i \overset{\text{i.i.d.}}{\sim} \mu_0, \qquad i = 1,\dots,N$
// Generate random directions
$\theta_n \sim \text{Uniform}(\mathbb{S}^{d-1}), \qquad n = 1,\dots,N_\theta$
// Quantiles of projected target
for $\theta \in \{\theta_n\}_{n=1}^{N_\theta}$ **do**
  $F^{-1}_{\theta^*_\#\nu} = \text{QF}\{\langle\theta, y_i\rangle\}_{i=1}^P$
// Iterations
for $k = 0,\dots K-1$ **do**
  for $\theta \in \{\theta_n\}_{n=1}^{N_\theta}$ **do**
    // CDF of projected particles
    $F_{\theta^*_\#\bar{\mu}_{kh}^N} = \text{CDF}\{\langle\theta, \bar{X}_k^i\rangle\}_{i=1}^N$
  // Update the particles
  $\bar{X}_{k+1}^i = \bar{X}_k^i - h\hat{v}_k(\bar{X}_k^i) + \sqrt{2\lambda h} Z_{k+1}^i$
  $i = 1,\dots,N$

mini-batches of the data, instead of the whole dataset of size P. This simplified procedure might also have some interesting consequences in privacy-preserving settings: since we can vary the number of projection directions N_θ for each data point y_i, we may guarantee that y_i cannot be recovered via these projections, by picking fewer than necessary for reconstruction using, e.g. compressed sensing (Donoho & Tanner, 2009).

3.4. Finite-time analysis for the infinite particle regime

In this section we will analyze the behavior of the proposed algorithm in the asymptotic regime where the number of particles $N \to \infty$. Within this regime, we will assume that the original SDE (8) can be directly simulated by using an approximate Euler-Maruyama scheme, defined starting at $\bar{X}_0 \overset{\text{i.i.d.}}{\sim} \mu_0$ as follows:

$$\bar{X}_{k+1} = \bar{X}_k + h\hat{v}(\bar{X}_k^i, \bar{\mu}_{kh}) + \sqrt{2\lambda h} Z_{k+1}, \tag{13}$$

where $\bar{\mu}_{kh}$ denotes the law of $\bar{X}_k$ with step size h and $\{Z_k\}_k$ denotes a collection of standard Gaussian random variables. Apart from its theoretical significance, this scheme is also practically relevant, since one would expect that it captures the behavior of the particle method (10) with large number of particles.

In practice, we would like to approximate the measure sequence $(\mu_t)_t$ as accurate as possible, where μ_t denotes the law of X_t. Therefore, we are interested in analyzing the distance $\|\bar{\mu}_{Kh} - \mu_T\|_{\text{TV}}$, where K denotes the total number of iterations, $T = Kh$ is called the horizon, and $\|\mu - \nu\|_{\text{TV}}$ denotes the total variation distance between two probability measures μ and ν: $\|\mu - \nu\|_{\text{TV}} \triangleq \sup_{A \in \mathcal{B}(\Omega)} |\mu(A) - \nu(A)|$.

In order to analyze this distance, we exploit the algorithmic similarities between (13) and the stochastic gradient Langevin dynamics (SGLD) algorithm (Welling & Teh, 2011), which is a Bayesian posterior sampling method having a completely different goal, and is obtained as a discretization of an SDE whose drift has a much simpler form. We then bound the distance by extending the recent results on SGLD (Raginsky et al., 2017) to time- and measure-dependent drifts, that are of our interest in the paper.

We now present our second main theoretical result. We present all our assumptions and the explicit forms of the constants in the supplementary document.

Theorem 3. *Assume that the conditions given in the supplementary document hold. Then, the following bound holds for $T = Kh$:*

$$\|\bar{\mu}_{Kh} - \mu_T\|_{\text{TV}}^2 \leq \delta_\lambda \left\{ \frac{L^2 K}{2\lambda} \left(\frac{C_1 h^3}{3} + 3\lambda d h^2 \right) + \frac{C_2 \delta K h}{4\lambda} \right\}, \tag{14}$$

for some $C_1, C_2, L > 0$, $\delta \in (0, 1)$, and $\delta_\lambda > 1$.

Here, the constants C_1, C_2, L are related to the regularity and smoothness of the functions v and $\hat{v}$; δ is directly proportional to the variance of $\hat{v}$, and δ_λ is inversely proportional to λ. The theorem shows that if we choose h small enough, we can have a non-asymptotic error guarantee, which is formally shown in the following corollary.

Corollary 1. *Assume that the conditions of Theorem 3 hold. Then for all $\varepsilon > 0$, $K \in \mathbb{N}_+$, setting*

$$h = (3/C_1) \wedge \left(\frac{2\varepsilon^2 \lambda}{\delta_\lambda L^2 T} (1 + 3\lambda d)^{-1} \right)^{1/2}, \tag{15}$$

we have

$$\|\bar{\mu}_{Kh} - \mu_T\|_{\text{TV}} \leq \varepsilon + \left(\frac{C_2 \delta_\lambda \delta T}{4\lambda} \right)^{1/2} \tag{16}$$

for $T = Kh$.

This corollary shows that for a large horizon T, the approximate drift $\hat{v}$ should have a small variance in order to obtain accurate estimations. This result is similar to (Raginsky et al., 2017) and (Nguyen et al., 2019): for small ε the variance of the approximate drift should be small as well. On the other hand, we observe that the error decreases as λ increases. This behavior is expected since for large λ, the Brownian term in (8) dominates the drift, which makes the simulation easier.

We note that these results establish the explicit dependency of the error with respect to the algorithm parameters (e.g. step-size, gradient noise) for a fixed number of iterations, rather than explaining the asymptotic behavior of the algorithm when K goes to infinity.

4. Experiments

In this section, we evaluate the SWF algorithm on a synthetic and a real data setting. Our primary goal is to validate our theory and illustrate the behavior of our non-standard approach, rather than to obtain the state-of-the-art results in IGM. In all our experiments, the initial distribution μ_0 is selected as the standard Gaussian distribution on $\mathbb{R}^d$, we take $Q = 100$ quantiles and $N = 5000$ particles, which proved sufficient to approximate the quantile functions accurately.

4.1. Gaussian Mixture Model

We perform the first set of experiments on synthetic data where we consider a standard Gaussian mixture model (GMM) with 10 components and random parameters. Centroids are taken as sufficiently distant from each other to make the problem more challenging. We generate $P = 50000$ data samples in each experiment.

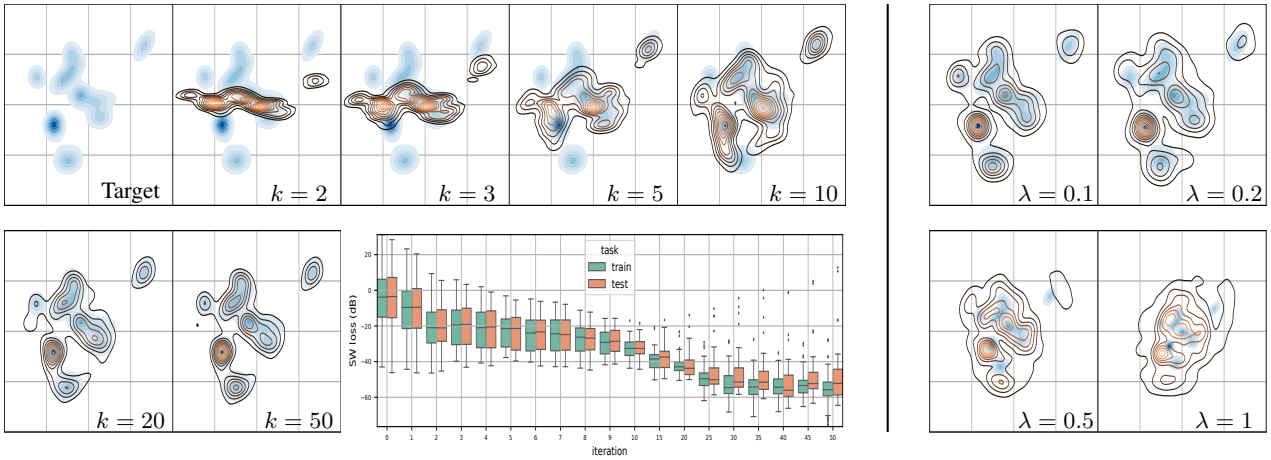


Figure 1. SWF on toy 2D data. **Left:** Target distribution (shaded contour plot) and distribution of particles (lines) during SWF. (bottom) SW cost over iterations during training (left) and test (right) stages. **Right:** Influence of the regularization parameter λ.

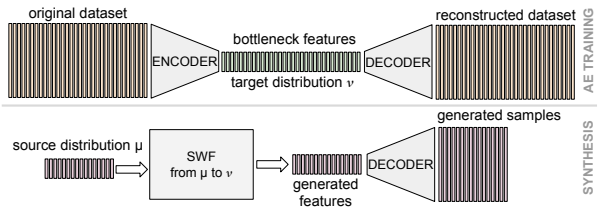


Figure 2. First, we learn an autoencoder (AE). Then, we use SWF to transport random vectors to the distribution of the bottleneck features of the training set. The trained decoder is used for visualization.

In our first experiment, we set $d = 2$ for visualization purposes and illustrate the general behavior of the algorithm. Figure 1 shows the evolution of the particles through the iterations. Here, we set $N_\theta = 30$, $h = 1$ and $\lambda = 10^{-4}$. We first observe that the SW cost between the empirical distributions of training data and particles is steadily decreasing along the SW flow. Furthermore, we see that the QFs, $F^{-1}_{\theta^*_\# \bar{\mu}^N_{kh}}$ that are computed with the initial set of particles (the *training* stage) can be perfectly re-used for new unseen particles in a subsequent *test* stage, yielding similar — yet slightly higher — SW cost.

In our second experiment on Figure 1, we investigate the effect of the level of the regularization λ. The distribution of the particles becomes more spread with increasing λ. This is due to the increment of the entropy, as expected.

4.2. Experiments on real data

In the second set of experiments, we test the SWF algorithm on two real datasets. (i) The traditional MNIST dataset that contains 70K binary images corresponding to different digits. (ii) The popular CelebA dataset (Liu et al., 2015), that contains 202K color-scale images. This dataset is advocated as more challenging than MNIST. Images were interpolated as 32×32 for MNIST, and 64×64 for CelebA.

Figure 3. Samples generated after 200 iterations of SWF to match the distribution of bottleneck features for the training dataset. Visualization is done with the pre-trained decoder.

In experiments reported in the supplementary document, we found out that directly applying SWF to such high-dimensional data yielded noisy results, possibly due to the insufficient sampling of $\mathbb{S}^{d-1}$. To reduce the dimensionality, we trained a standard convolutional autoencoder (AE) on the training set of both datasets (see Figure 2 and the supplementary document), and the target distribution ν considered becomes the distribution of the resulting bottleneck features, with dimension d. Particles can be visualized with the pre-trained decoder. Our goal is to show that SWF permits to directly sample from the distribution of bottleneck features, as an alternative to enforcing this distribution to

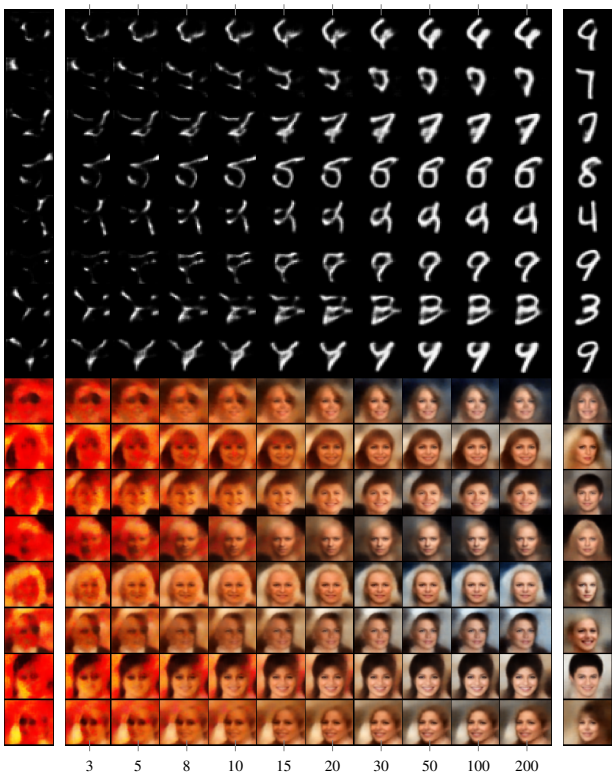


Figure 4. Initial random particles (left), particles through iterations (middle, from 1 to 200 iterations) and closest sample from the training dataset (right), for both MNIST and CelebA.

match some prior, as in VAE. In the following, we set $\lambda = 0$, $N_\theta = 40000$, $d = 32$ for MNIST and $d = 64$ for CelebA.

Assessing the validity of IGM algorithms is generally done by visualizing the generated samples. Figure 3 shows some particles after 500 iterations of SWF. We can observe they are considerably accurate. Interestingly, the generated samples gradually take the form of either digits or faces along the iterations, as seen on Figure 4. In this figure, we also display the closest sample from the original database to check we are not just reproducing training data.

For a visual comparison, we provide the results presented in (Deshpande et al., 2018) in Figure 5. These results are obtained by running different IGM approaches on the MNIST dataset, namely GAN (Goodfellow et al., 2014), Wasserstein GAN (W-GAN) (Arjovsky et al., 2017) and the Sliced-Wasserstein Generator (SWG) (Deshpande et al., 2018). The visual comparison suggests that the samples generated by SWF are of slightly better quality than those, although research must still be undertaken to scale up to high dimensions without an AE.

Figure 5. Performance of GAN (left), W-GAN (middle), SWG (right) on MNIST. (The figure is directly taken from (Deshpande et al., 2018).)

Figure 6. Applying a pre-trained SWF on new samples located in-between the ones used for training. Visualization is done with the pre-trained decoder.

We also provide the outcome of the pre-trained SWF with samples that are regularly spaced in between those used for training. The result is shown in Figure 4.2. This plot suggests that SWF is a way to interpolate non-parametrically in between latent spaces of regular AE.

5. Conclusion and Future Directions

In this study, we proposed SWF, an efficient, nonparametric IGM algorithm. SWF is based on formulating IGM as a functional optimization problem in Wasserstein spaces, where the aim is to find a probability measure that is close to the data distribution as much as possible while maintaining the expressiveness at a certain level. SWF lies in the intersection of OT, gradient flows, and SDEs, which allowed us to convert the IGM problem to an SDE simulation problem. We provided finite-time bounds for the infinite-particle regime and established explicit links between the algorithm parameters and the overall error. We conducted several experiments, where we showed that the results support our theory: SWF is able to generate samples from non-trivial distributions with low computational requirements.

The SWF algorithm opens up interesting future directions: (i) extension to differentially private settings (Dwork & Roth, 2014) by exploiting the fact that it only requires random projections of the data, (ii) showing the convergence scheme of the particle system (9) to the original SDE (8), (iii) providing bounds directly for the particle scheme (10).

Acknowledgments

This work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014) and KAMoulox (ANR-15-CE38-0003-01) projects. Szymon Majewski is partially supported by Polish National Science Center grant number 2016/23/B/ST1/00454.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Bogachev, V. I., Krylov, N. V., Röckner, M., and Shaposhnikov, S. V. *Fokker-Planck-Kolmogorov Equations*, volume 207. American Mathematical Soc., 2015.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Bossy, M. and Talay, D. A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Mathematics of Computation of the American Mathematical Society*, 66(217):157–192, 1997.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Cattiaux, P., Guillin, A., and Malrieu, F. Probabilistic approach for granular media equations in the non uniformly convex case. *Prob. Theor. Rel. Fields*, 140(1-2):19–40, 2008.

Şimşekli, U., Yildiz, C., Nguyen, T. H., Cemgil, A. T., and Richard, G. Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. In *ICML*, pp. 4674–4683, 2018.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Deshpande, I., Zhang, Z., and Schwing, A. Generative modeling using the sliced wasserstein distance. *arXiv preprint arXiv:1803.11188*, 2018.

Diggle, P. J. and Gratton, R. J. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 193–227, 1984.

Donoho, D. and Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

Durmus, A., Şimşekli, U., Moulines, E., Badeau, R., and Richard, G. Stochastic gradient Richardson-Romberg Markov Chain Monte Carlo. In *NIPS*, 2016.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Gan and vae from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gribonval, R., Blanchard, G., Keriven, N., and Traonmilin, Y. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, 2017.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5769–5779, 2017.

Guo, X., Hong, J., Lin, T., and Yang, N. Relaxed Wasserstein with applications to GANs. *arXiv preprint arXiv:1705.07164*, 2017.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kolouri, S., Martin, C. E., and Rohde, G. K. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

Lavenant, H., Claici, S., Chien, E., and Solomon, J. Dynamical optimal transport on discrete surfaces. In *SIGGRAPH Asia 2018 Technical Papers*, pp. 250. ACM, 2018.

Lei, N., Su, K., Cui, L., Yau, S.-T., and Gu, D. X. A geometric view of optimal transportation and generative model. *arXiv preprint arXiv:1710.05488*, 2017.

Liu, S., Bousquet, O., and Chaudhuri, K. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pp. 5551–5559, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Ma, Y. A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2899–2907, 2015.

Malrieu, F. Convergence to equilibrium for granular media equations and their Euler schemes. *Ann. Appl. Probab.*, 13(2):540–560, 2003.

Mishura, Y. S. and Veretennikov, A. Y. Existence and uniqueness theorems for solutions of McKean–Vlasov stochastic equations. *arXiv preprint arXiv:1603.02212*, 2016.

Mohamed, S. and Lakshminarayanan, B. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Nguyen, T. H., Şimşekli, U., , and Richard, G. Non-asymptotic analysis of fractional Langevin Monte Carlo for non-convex optimization. In *ICML*, 2019.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M. (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1674–1703, 2017.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

Santambrogio, F. Introduction to optimal transport theory. In Pajot, H., Ollivier, Y., and Villani, C. (eds.), *Optimal Transportation: Theory and Applications*, chapter 1. Cambridge University Press, 2014.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

Şimşekli, U. Fractional Langevin Monte Carlo: Exploring Lévy Driven Stochastic Differential Equations for Markov Chain Monte Carlo. In *International Conference on Machine Learning*, 2017.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

Veretennikov, A. Y. On ergodic measures for McKean-Vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pp. 471–486. Springer, 2006.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pp. 681–688, 2011.

Wu, J., Huang, Z., Li, W., and Gool, L. V. Sliced wasserstein generative models. *arXiv preprint arXiv:1706.02631*, abs/1706.02631, 2018.

Zhang, J., Zhang, R., and Chen, C. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. *arXiv preprint arXiv:1809.01293*, 2018.

This document is my thesis for getting the *habilitation à diriger des recherches*, which is the french diploma that is required to fully supervise Ph.D. students. It summarizes the research I did in the last 15 years and also provides the short-term research directions and applications I want to investigate.
Regarding my past research, I first describe the work I did on probabilistic audio modeling, including the separation of Gaussian and α-stable stochastic processes. Then, I mention my work on deep learning applied to audio, which rapidly turned into a large effort for community service. Finally, I present my contributions in machine learning, with some works on hardware compressed sensing and probabilistic generative models.
My research programme involves a theoretical part that revolves around probabilistic machine learning, and an applied part that concerns the processing of time series arising in both audio and life sciences.

Laboratoire d'informatique
de robotique et de microelectronique
de Montpellier

161 Rue Ada
34095 Montpellier
www.lirmm.fr