Contributions in Audio Modeling for Solving Inverse Problems: Source Separation, Compression and Inpainting

Alexey Ozerov

Habilitation à Diriger des Recherches

Université de Rennes 1

Spécialité "Traitement du Signal"

présentée et soutenue publiquement par

Alexey Ozerov

le 28 novembre 2019

Contributions in Audio Modeling for Solving Inverse Problems: Source Separation, Compression and Inpainting

Jury

M. Pierre Comon,	DR, CNRS	Rapporteur
M. Laurent Albera,	MCF, Université de Rennes 1	Examinateur
M. Jérôme Idier,	DR, CNRS	Examinateur
Mme Dorothea Kolossa,	Professeur, Ruhr-Universität Bochum	Rapporteur
M. Axel Roebel,	DR, IRCAM	Rapporteur

Préparée et soutenue chez InterDigital
Cesson-Sévigné, France

H
D
R

A ma mère

Acknowledgments

First of all, I would like to thank Laurent Albera, Pierre Comon, Jérôme Idier, Dorothea Kolossa and Axel Roebel for being in my jury.

I am also very grateful to all my collaborators (some of them are friends as well). I will not list all the names here, since there are too many and I am afraid forgetting someone. Without them the work I have accomplished would not be possible.

I thank all my friends in Russia and France without listing the names for the same reason as above. You are always here with me when I need during happy or sad moments.

Many many thanks to my family in Russia, my sisters, my aunt, my uncle, and my nephews. You are always with me too.

I also thank my 195 friends on Facebook :-)

I thank my loves.

I thank my balalaïkas, traditional 3-chords Russian folk music instruments I am playing regularly.

Finally, very special thanks to my children Sonia and Simon. Vous êtes mon bonheur !

Abstract

In this HDR manuscript I summarize some of my work concerning nonnegative matrix/tensor factorization (NMF/NTF) modeling of audio spectrograms to solve various ill-posed inverse problems in audio signal processing domain. Those inverse problems include audio source separation, audio inpainting (i.e., missing audio samples recovery) and audio compression. This summary is built based on four contributions published in four journal papers. As for the state of the art prior to this work, NMF/NTF decompositions were already used for a while to model audio and have successfully found applications in, e.g., audio source separation and music transcription. However, in those applications NMF/NTF models were applied to approximate some observed spectrograms of audio signals. In my opinion, the main qualitative change I have proposed that unifies all this work is as follows : instead of applying NMF/NTF to the spectrograms of observed signals, I proposed applying it to the spectrograms of latent signals whatever the observations. This became possible thanks to probabilistic Gaussian formulation of NMF/NTF with Itakura-Saito divergence. As a result, this allowed not only significantly improving audio source separation in the multichannel setting, but also applying NMF/NTF modeling to audio inpainting and compression, which lead to new approaches that are better than or on par with the state of the art.

Table of contents

Introduction **2**

1 NMF/NTF modeling for audio source separation **4**
1.1 Nonnegative matrix factorization 4
1.2 Nonnegative tensor factorization 9
1.3 Conclusion 12

2 Proposed multichannel NMF/NTF modeling **13**
2.1 Main idea 13
2.2 Assumptions and modeling 14
2.3 Estimation criteria and algorithms 16
2.4 Conclusion 18

3 Paper 1 : Multichannel NMF **19**
3.1 Audio source separation of multichannel mixtures 19
3.2 Multichannel NMF model 20
3.3 Results 21
3.4 Impact and followings 22
3.5 Conclusion 23

4 Paper 2 : A general flexible framework **24**
4.1 Local Gaussian modeling 24
4.2 Motivation 25
4.3 Formulation 27
4.4 Algorithm 28
4.5 Implementation 29
4.6 Conclusion 29

5 Paper 3 : Coding-based informed source separation **30**
5.1 Motivation 30
5.2 Coding-based ISS at glance 31
5.3 CISS based on multisource NTF 32
5.4 Results 34
5.5 Conclusion 34

6 Paper 4 : Time-domain inverse problems via NTF **35**
6.1 General framework formulation 35
6.2 Modeling and algorithms 36
6.3 Applications and results 37
6.4 Conclusion 42

7 Other work **44**
7.1 Flexible speech and audio coding 44
7.2 Learning from uncertain data 44
7.3 Source localization 45
7.4 Source separation evaluation 45
7.5 Informed source separation 45
7.6 Audio-visual scenes understanding 46
7.7 A bit of image/video processing : Faces 46
7.8 Audio style transfer 47
7.9 Tutorials, review paper and book chapters 48

8 Conclusion **49**

Appendix **61**
Paper 1 (Ozerov & Févotte, *IEEE TASLP*, 2010) 62
Paper 2 (Ozerov, Vincent & Bimbot, *IEEE TASLP*, 2011) 77
Paper 3 (Ozerov, Liutkus, Badeau & Richard, *IEEE TASLP*, 2013) . . 94
Paper 4 (Bilen, Ozerov & Pérez, *IEEE TSP*, 2018) 109
Curriculum Vitae 123

List of figures

1.1 NMF with $K = 3$ as a matrix product and a sum of rank-1 matrices. . 4
1.2 An example of IS-NMF decomposition with $K = 5$ components of a piano expert available at https://www.irit.fr/~Cedric.Fevotte/extras/neco09/Piano.wav. 7
1.3 NTF (CANDECOMP / PARAFAC) with $K = 6$ as a sum of rank-1 tensors. 10

2.1 General multichannel multisource NTF modeling together with an example of linear transform $\mathbf{A}$, where 3 sources are convolutively mixed into 2 channels in time domain and then subsampled. 15

3.1 Representation of convolutive mixing system and formulation of Multichannel NMF problem (figure from [1]). 21
3.2 Results for "Under-determined speech and music mixtures - instantaneous mixtures" SiSEC 2008 [VAB09] task. The results are plotted in terms of source to distortion ratio (SDR) [VGF06] (higher is better). . 22

4.1 Current way of addressing a new source separation problem (top) and the way of addressing it using the proposed flexible framework (bottom) (figure from [2]). 27
4.2 A particular hierarchical spectral decomposition as applied to the spectral power of several xylophone notes (figure from [2]). 28
4.3 Overview of the proposed general MU-GEM algorithm for parameter estimation and source separation (figure from [2]). 29

5.1 Simplified visualization of the following probabilistic model-based methods applied in one TF point : (A) conventional ISS [LPB+12], (B) source coding and (C) the proposed coding-based ISS (CISS). Notations : x : mixture, $\mathbf{s} = [s_1, s_2]^T$: sources, $p(\mathbf{s}|\theta)$: *a priori* source distribution, $p(\mathbf{s}|x, \theta)$: *a posteriori* source distribution, $\mathbf{s}^*$: true sources, $\hat{\mathbf{s}}$: estimated sources. 32
5.2 CISS-NTF with different ways of optimizing parameters (solid lines), compared to state of the art [LPB+12] (dotted lines). δPSM and δSDR denote the improvements over the corresponding measures computed for the oracle Wiener filtering [VGP07] source estimates in the STFT domain (figure from [3]). 34

6.1 General formulation of time-domain inverse problems. 36
6.2 The average performance of all the audio declipping algorithms as a function of the clipping threshold. Lower threshold corresponds to more severe clipping (figure from [4]). 39

6.3 The declipping and source separation performance of joint optimization compared to sequential (figure from [4]). 40
6.4 The reconstruction performance measured in terms of SNR_m of a 4s long music signal from its random samples. The reconstruction results with our proposed algorithm (solid lines) are shown for different percentage of samples and different number of components, K, used in our approach. The results with shape preserving piecewise cubic interpolation are also shown for comparison (dashed lines), with the colors indicating corresponding percentage of samples (figure from [4]). 41
6.5 The rate-distortion performance of CS-ISS using different quantization levels of the encoded samples. The performance of the ISS algorithm from [LPB+12] and the coding-based ISS algorithm from [3] are also shown for comparison (figure from [4]). 42

7.1 Example of spatio-temporal face annotation in Hannah dataset [5]. . . 47
7.2 Example of image style transfer (from [6]). 48

Introduction

This document describes a part of my research work I have done since my PhD defense in 2006. I was interested in and working on various ill-posed inverse problems involving audio signals. Though the main focus of my research is on audio source separation, I was also interested in other inverse problems such as audio inpainting (or interpolation) and audio compression that may be sometimes seen as an inverse problem. As such, I start by first introducing audio source separation and its various scenarios.

Most audio signals are *mixtures* of several *sources*. For example a music recording may be a mixture of several instruments. The goal of audio source separation consists in estimating the sources from their mixtures. This may be useful for various applications. First, breaking audio into its elementary parts may facilitate its analysis (e.g., speech recognition or audio events detection). Second, extracting individual sources is useful for upmixing/re-mixing applications and other audio editing tasks. Audio source separation being in general a very ill-posed inverse problem, it still remains a very challenging and it is extensively studied. Moreover, the success of audio source separation depends strongly on the amount of available prior information about the sources. As such, various source separation scenarios might be considered :

— *Blind (or non-supervised) source separation* - One cannot assume anything about the sources, except that they are audio signals.

— *Supervised and semi-supervised source separation* - One can describe all or a part of sources, e.g., by providing examples of similar sources.

— *Informed source separation* - Some complementary information about the sources is available, i.e., this may be music score in case of music source separation or some information provided by a user via a dedicated interface.

— *Audio objects compression* - It is assumed that at a so-called *encoding stage* the original sources and the mixture are available, and the goal is to extract some compact information that will allow reconstructing the sources at a so-called *decoding stage*, where only the mixture is given.

Since my PhD defended in 2006 I was working on various cases from all these scenarios, as well as on some other problems including audio inpainting (reconstruction of missing parts of audio signals).

Most of my post-PhD research being conducted in the pre-deep learning era, I was mostly concentrated on non-negative matrix factorization (NMF) and non-negative tensor factorization (NTF) which were and still remain very popular and successful approaches for audio source separation. In my opinion my main methodological contribution consists in introducing a new probabilistic structured Gaussian modeling of multichannel and multisource audio. This modeling is based on the assumption that the short-time Fourier transform (STFT) coefficients of latent sources are zero-mean Gaussians with variances structured via an NMF or an NTF decompositions. As such,

this modeling has been baptized as *multichannel NMF* or *multichannel NTF*. I have used it within several audio source separation scenarios described above, as well as for audio inpainting. Multichannel NMF/NTF modeling in my research gave rise to a series of four journal papers [1, 2, 3, 4] and to several conference papers. The modeling was initially proposed in [1] and then extended to or used for different applications/contexts in [2, 3, 4]. In summary :

- *Paper 1, [1] :* Multichannel NMF modeling is introduced and applied to blind source separation. It is then extended to multichannel NTF in [7].
- *Paper 2, [2] :* Multichannel NMF modeling is extended to a much more general framework generalizing several other state-of-the-art source separation methods and allowing implementing new ones.
- *Paper 3, [3] :* Multichannel NTF modeling is applied to audio objects compression.
- *Paper 4, [4] :* Multichannel NTF modeling is revisited within a framework allowing reconstructing missing audio samples in time domain. This allows various applications including audio declipping, audio objects compression and compressive sampling recovery.

I have decided to speak in this document mostly about the work presented in these four publications. This is because there is a very common and consistent story regrouping them, and because multichannel NMF/NTF modeling had a quite significant scientific impact in the community.

This document is structured as follows. Chapter 1 introduces NMF/NTF modeling and its applications in audio with a strong emphasis on audio source separation. In Chapter 2 I give a very general and high level presentation of multichannel NMF/NTF modeling. Chapters 3 to 6 are devoted to the presentation of the four journal papers mentioned above. In Chapter 7 I speak briefly about other work I've done. The conclusions are drawn in Chapter 8. The four papers [1, 2, 3, 4] together with a long version of my Curriculum Vitae are annexed at the end of this document.

I must acknowledge that I have not done this work alone, but with many collaborators including colleagues and students I have supervised. I am very grateful to all these people and without them this work would not be possible. As such, from now on in this document, except if I am expressing my personal opinion, I am switching from saying "I" and "my" to saying "we" and "our", while speaking about the work.

Chapitre 1

NMF/NTF modeling for audio source separation

1.1 Nonnegative matrix factorization

1.1.1 Problem statement

Nonnegative matrix factorization (NMF) [LS99, FBD09] is a dimensionality reduction technique that approximates an $F \times N$ data matrix $\mathbf{B}$ with nonnegative entries as a product of two matrices with nonnegative entries (see Fig. 1.1) such as

$$\mathbf{B} \approx \mathbf{WH}, \tag{1.1}$$

where $\mathbf{B} \in \mathbb{R}_+^{F \times N}$, $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$. It is also usually assumed that K is much smaller than F and N, e.g., $K << \min(F, N)$, so as to achieve dimensionality reduction. Matrix product in approximation (1.1) can also be rewritten as a sum of matrices of rank 1 as follows :

$$\mathbf{B} \approx \sum_{k=1}^{K} \mathbf{w}_k \mathbf{h}_k, \tag{1.2}$$

where $\mathbf{w}_k$ denote the columns of matrix $\mathbf{W}$ and $\mathbf{h}_k$ denote the rows of matrix $\mathbf{H}$. This decomposition is represented on Figure 1.1.

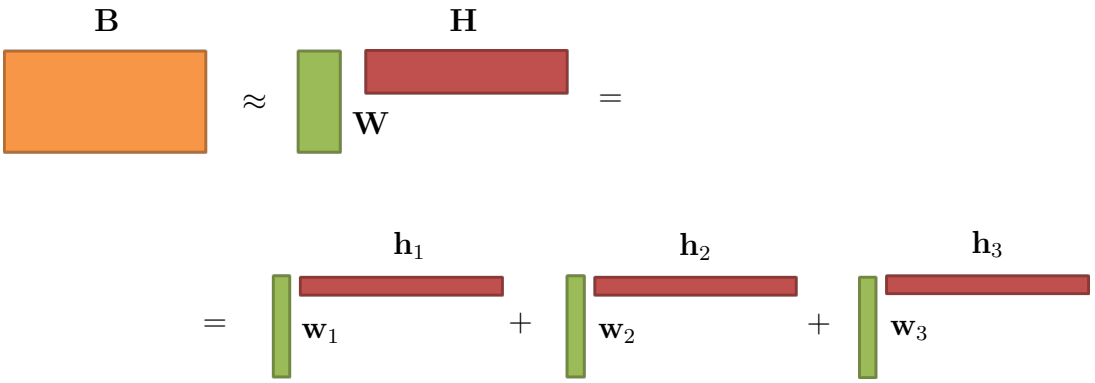


FIGURE 1.1 – NMF with $K = 3$ as a matrix product and a sum of rank-1 matrices.

Approximation (1.1) being mathematically poorly defined, one usually looks for a pair ($\mathbf{W}$, $\mathbf{H}$), while optimizing some measure of fit between $\mathbf{B}$ and its approximation

WH. More precisely **W** and **H** are usually found by optimizing

$$(\mathbf{W}, \mathbf{H}) = \arg \min_{\mathbf{W}' \geq 0, \mathbf{H}' \geq 0} C(\mathbf{W}', \mathbf{H}'), \tag{1.3}$$

where $C(\mathbf{W}, \mathbf{H})$ is defined as

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{B} \| \mathbf{WH}), \tag{1.4}$$

and $D(\mathbf{B}\|\mathbf{A})$ is some divergence between nonnegative matrices $\mathbf{B} = [b_{fn}]_{f,n=1}^{F,N} \in \mathbb{R}_+^{F \times N}$ and $\mathbf{A} = [a_{fn}]_{f,n=1}^{F,N} \in \mathbb{R}_+^{F \times N}$ specified as

$$D(\mathbf{B}\|\mathbf{A}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(b_{fn}|a_{fn}), \tag{1.5}$$

with $d(b_{fn}|a_{fn})$ being a scalar divergence. Many different scalar divergences were proposed [LS01, FBD09, CZPA09], but among the most popular and mostly used in audio processing there are the following three :

— Euclidean (EUC) distance :

$$d_{EUC}(b|a) = \frac{1}{2}(b-a)^2.$$

— Kullback-Leibler (KL) divergence :

$$d_{KL}(b|a) = b \log \frac{b}{a} - b + a.$$

— Itakura-Saito (IS) divergence :

$$d_{IS}(b|a) = \frac{b}{a} - \log \frac{b}{a} - 1.$$

Note that the uniqueness of solution of (1.3) is in general not assured [LCP+08]. First of all, there are obvious scaling and permutation ambiguities, i.e., any column of **W** and the corresponding row of **H** may be multiplied by $z > 0$ and $1/z$, respectively, without changing the value of **WH**, and those columns and rows might be altogether arbitrary permuted. However, besides those obvious ambiguities there are ambiguities that are less evident (see [LCP+08] for details).

1.1.2 Algorithms

To optimize criterion (1.3) various algorithms exist [CZPA09]. However, the so-called *multiplicative update (MU)* rules [LS01, FBD09, FI11] are among the most popular and the most widely used. As such, we describe these rules below. Let $\nabla_{\mathbf{W}} C(\mathbf{W}, \mathbf{H})$ and $\nabla_{\mathbf{H}} C(\mathbf{W}, \mathbf{H})$ partial derivatives of the cost function $C(\mathbf{W}, \mathbf{H})$ with respect to **W** and **H**, respectively, the MU rules consist in alternating between the following two steps [FI11]

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{[\nabla_{\mathbf{W}} C(\mathbf{W}, \mathbf{H})]_-}{[\nabla_{\mathbf{W}} C(\mathbf{W}, \mathbf{H})]_+} \right)^{.\eta}, \tag{1.6}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{[\nabla_{\mathbf{H}} C(\mathbf{W}, \mathbf{H})]_-}{[\nabla_{\mathbf{H}} C(\mathbf{W}, \mathbf{H})]_+} \right)^{.\eta}, \tag{1.7}$$

where $\eta > 0$, "$\odot$" denotes element-wise matrix product, "$.p$" denotes element-wise matrix power, matrix division is element-wise as well, and $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_-$ and $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_+$ are both nonnegative and such that

$$\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H}) = [\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_+ - [\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_- , \tag{1.8}$$

and similarly for $[\nabla_{\mathbf{H}} C(\mathbf{W},\mathbf{H})]_-$ and $[\nabla_{\mathbf{H}} C(\mathbf{W},\mathbf{H})]_+$. Note that decomposition (1.8) is arbitrary, since it still holds with any positive constant matrix of suitable size added to both $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_+$ and $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_-$. As such, decomposition (1.8) is usually chosen so as $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_+$ and $[\nabla_{\mathbf{W}} C(\mathbf{W},\mathbf{H})]_-$ are both entry-wise minimal, while keeping closed-form expressions. In order to avoid numerical issues (e.g., over or underflow) a suitable re-scaling should be applied after each iteration or from time to time [FBD09]. Note also that with this approach the nonnegativity constraints are respected by construction.

For example, in case of NMF with IS divergence the MU rules become [FBD09]

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\mathbf{H}^T((\mathbf{WH})^{\cdot -2} \odot \mathbf{B})}{\mathbf{H}^T(\mathbf{WH})^{\cdot -1}} \right)^{\cdot \eta}, \tag{1.9}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T((\mathbf{WH})^{\cdot -2} \odot \mathbf{B})}{\mathbf{W}^T(\mathbf{WH})^{\cdot -1}} \right)^{\cdot \eta}. \tag{1.10}$$

The MU rules were initially discovered based on some heuristics [LS01]. They can be also interpreted as a diagonally rescaled gradient descent [LS01]. However, such a property does not directly bring any light on the algorithm's properties such as for example *monotonicity*, i.e., whether the updates guarantee the cost function to be non-increasing after each iteration ? In some cases the monotonicity was investigated and proven a-posteriori [LS01, FI11] by interpreting each particular variant of MU rules as a majorisation-minimization (MM) procedure [HL04]. In particular, it was proven [FI11] that the monotonicity is guaranteed for the EUC and KL divergences when $0 < \eta \leq 1$, and for the IS divergence when $0 < \eta \leq 1/2$. However, as for the IS divergence, the monotonicity is usually observed in practice for $0 < \eta \leq 1$. As such, hereafter we will use MU rules for NMF with IS divergence without exponent η, i.e., we assume $\eta = 1$.

While the MU rules are far from being the most efficient algorithm in terms of convergence speed [CZPA09], it is probably one of the most popular for the following reasons. The nonnegativity constraints are respected by construction, the formulation is very compact, and, by consequence, the implementation is usually easy (e.g., just two lines of code in a loop in Matlab).

1.1.3 Application in audio

Let $\mathbf{X} = [x_{fn}]_{f,n=1}^{F,N}$ complex-valued short-time Fourier transform (STFT) of an audio signal. In case of application of NMF to audio signals one usually considers the magnitude spectrogram $|\mathbf{X}|$ as nonnegative data matrix $\mathbf{B}$ ($\mathbf{B} = |\mathbf{X}|$), e.g., in case of EUC distance or KL divergence ; or the power spectrogram ($\mathbf{B} = |\mathbf{X}|^{.2}$), e.g., in case of IS divergence. Since in this work we are mostly using the IS divergence, we assume hereafter $\mathbf{B} = |\mathbf{X}|^{.2}$.

Let us first look what happens when applying NMF to a power spectrogram of a music audio sample. We took a piano expert, where four different notes are played in different combinations, and we applied to it an IS-NMF decomposition with $K = 5$ components. The result is shown on Figure 1.2. One can see that the columns of matrix

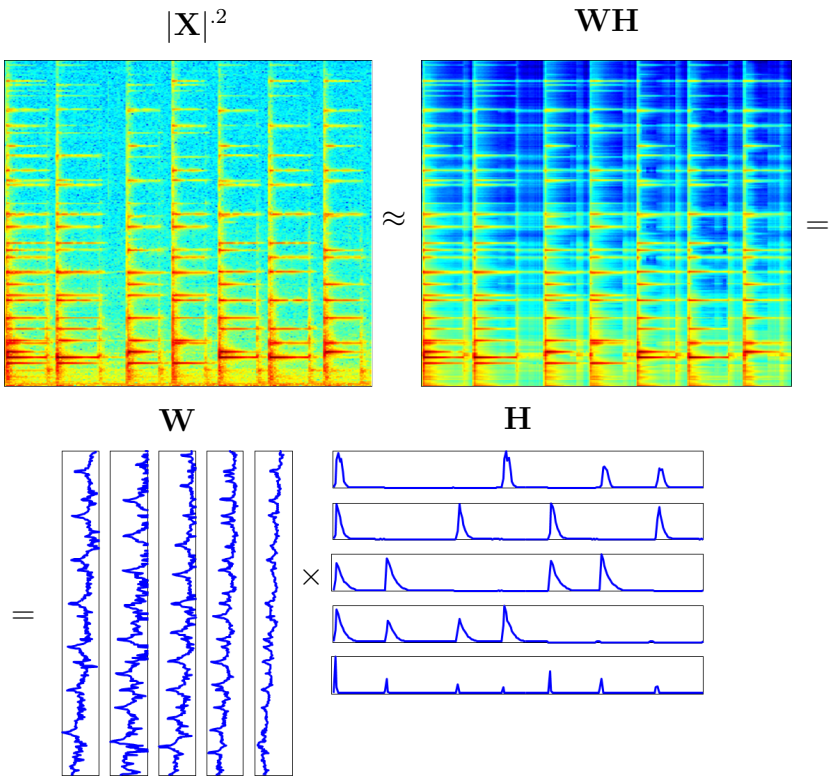


FIGURE 1.2 – An example of IS-NMF decomposition with $K = 5$ components of a piano expert available at https://www.irit.fr/~Cedric.Fevotte/extras/neco09/Piano.wav.

$\mathbf{W}$ represent characteristic spectral patterns of individual audio objects, notably here the four notes and the sound of piano hummers (5th component). Moreover, the rows of matrix $\mathbf{H}$ represent the activations of these objects in time. It is clear that such an object-based decomposition opens a door for various applications. For example, one can transcribe the music (i.e., estimate the music score from audio) [SB03, BBR07] by simply thresholding the rows of $\mathbf{H}$ to identify the individual notes. One can also perform source separation [Vir07, FBD09] by separating individual notes and then regrouping them.

To summarize, the NMF modeling is very attractive for audio thanks to the following properties :

— it is an object-based decomposition, thus allowing various applications in audio manipulation and analysis,

— it is quite general and, thus, suitable for various types of sounds such as music, speech and environmental sounds (though, it is a little bit less efficient for speech than for music, since speech is usually exhibits stronger pitch variation and thus one needs much more characteristic spectral patterns to describe it well),

— in contrast to other models such as, e.g., Gaussian mixture models (GMMs), it allows handling polyphony [8].

Let us now turn back to the NMF with IS divergence (IS-NMF). A very attractive property of IS-NMF, which we use extensively in this work, is that it allows the following probabilistic interpretation. Let us assume the complex-valued STFT coefficients x_{fn} mutually independent and each coefficient distributed as

$$x_{fn} \sim \mathcal{N}_c(0, [\mathbf{WH}]_{fn}), \tag{1.11}$$

where in a more general vector case $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the *proper* complex Gaussian distribution [NM93] with probability density function (pdf)

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\pi \boldsymbol{\Sigma}|} \exp\left[-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^H (\mathbf{x}-\boldsymbol{\mu})\right], \tag{1.12}$$

$\boldsymbol{\mu}$ being complex-valued mean vector, and $\boldsymbol{\Sigma}$ being complex-valued Hermitian covariance matrix. It was proven in this case [FBD09] that the NMF optimization criterion (1.3) with IS divergence is strictly equivalent to the maximum likelihood (ML) estimation of $\mathbf{W}$ and $\mathbf{H}$.

This probabilistic reformulation, though equivalent, is more attractive since it models directly the STFT coefficients x_{fn} (thus the signal itself), and not the power spectrogram $|x_{fn}|^2$, where the phase information is lost.

Another attractive property of IS divergence, in light of application to audio, is that it is scale-invariant [FBD09], i.e.,

$$d_{IS}(\lambda b, \lambda a) = \lambda d_{IS}(b, a) \tag{1.13}$$

for any $\lambda > 0$. This makes it equally sensitive to sounds with low and high power.

1.1.4 Application to single channel source separation

The single channel source separation problem is usually formulated as follows. It is assumed that J signals, called *sources*, are added to form a so-called *mixture*

$$\tilde{x}(t) = \sum_{j=1}^{J} \tilde{s}_j(t), \tag{1.14}$$

where t stands for sample index in time domain, $\tilde{x}(t)$ denote mixture samples, and $\tilde{s}_j(t)$ denote jth source samples. The problem is to estimate unknown sources under the mixing assumption (1.14), given the observed mixture.

Thanks to the linearity of the STFT transform the mixing equation (1.14) rewrites in the STFT domain as

$$x_{fn} = \sum_{j=1}^{J} s_{jfn}, \tag{1.15}$$

where x_{fn} and s_{jfn} denote the STFT coefficients of the mixture and the sources, respectively.

Now we assume that the power spectrogram of each source $|\mathbf{S}_j|^2$ is modeled by an IS-NMF as [FBD09]

$$|\mathbf{S}_j|^{\cdot 2} \approx \mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j, \tag{1.16}$$

with $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ and $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$. Equivalently, as mentioned in the previous section, this can be re-formulated in a probabilistic manner as

$$s_{jfn} \sim \mathcal{N}_c(0, [\mathbf{W}_j \mathbf{H}_j]_{fn}). \tag{1.17}$$

We then stuck together the IS-NMF source models as $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_J]$ and $\mathbf{H} = [\mathbf{H}_1^T, \mathbf{H}_2^T, \ldots, \mathbf{H}_J^T]^T$, and one can easily show, thanks to mixing equation (1.15) and assumption (1.17), that expression (1.11) holds, and thus the mixture power spectrogram is as well modeled with IS-NMF as

$$|\mathbf{X}|^{\cdot 2} \approx \mathbf{W}\mathbf{H}. \tag{1.18}$$

Assuming all models well estimated, the sources can then be recovered by Wiener filtering as

$$\widehat{\mathbf{S}}_j = \frac{\mathbf{W}_j\mathbf{H}_j}{\mathbf{WH}} \odot \mathbf{X}, \tag{1.19}$$

which corresponds to the minimum mean squared error (MMSE) estimator in the STFT domain under assumptions (1.17) and (1.15).

The main problem here is that the source models $(\mathbf{W}_j, \mathbf{H}_j)$ cannot be directly estimated since the sources are unknown. However, the mixture model $(\mathbf{W}, \mathbf{H})$ may be estimated from the observed $\mathbf{X}$. Even though we assume that by chance each rank-1 component $\mathbf{w}_k\mathbf{h}_k$ of decomposition $(\mathbf{W}, \mathbf{H})$ corresponds to just one source, they still need to be correctly regrouped to form $(\mathbf{W}_j, \mathbf{H}_j)_{j=1}^J$, and usually this is unfeasible without additional prior information.

Various strategies exist to estimate source models with different levels of supervision. Let us mention just some of them :

— *Non-supervised (or blind) :* NMF components may be clustered a posteriori based on some criterion [SG09].

— *Semi-supervised :* Examples of some of J sources, but not of all sources, are available (say J^* examples, $0 < J^* < J$). Then, the spectral dictionaries $\mathbf{W}_j$ $(j = 1, \ldots, J^*)$ may be pre-trained on these source examples, concatenated to form $\mathbf{W}$ with some additional columns to describe the remaining sources $(j = J^*+1, \ldots, J)$, and then $(\mathbf{W}, \mathbf{H})$ might be learned from the mixture while keeping pre-trained $\mathbf{W}_j$ $(j = 1, \ldots, J^*)$ fixed [SRS07].

— *Weakly-supervised :* There are no examples of clean sources, but there are examples of source mixtures with less than J sources. For example, to separate "piano + bass + drums" mix, one would have "piano + bass", "piano + drums" and "bass + drums" example mixtures. It is possible to learn NMF source models from such weak annotations (see [LSCJ08] or [7] for details).

— *Supervised :* There are examples of all J sources. Then, as in semi-supervised case, $\mathbf{W}_j$ $(j = 1, \ldots, J)$ may be pre-trained on these source examples, concatenated to form $\mathbf{W}$, and then the $\mathbf{H}$ may be estimated from the mixture, while keeping $\mathbf{W}$ fixed [SRS07].

There are many other ways to regularize NMF model estimation, including audio object compression, where some information to guide model estimation may be extracted from the clean sources [PGB10, LPB+12], or informed source separation, where the estimation is guided by some available complementary information (e.g., music score [EPMP14] or text [9]) or by a user [10].

1.2 Nonnegative tensor factorization

1.2.1 Problem statement

By tensors we mean L-way arrays or simply datasets indexed by L indices. For example, in case of $L = 2$ we are back to matrices and in case of $L = 3$ we have sort of "boxes". Since in this work we are dealing only with 3-way tensor, we limit our presentation here to the particular case of $L = 3$.

Let our data represented by a 3-way tensor $\boldsymbol{\mathcal{B}} = [b_{jfn}]_{j,f,n=1}^{J,F,N}$ of size $J \times F \times N$ with nonnegative entries. There are many kinds of nonnegative tensor factorization (NTF) models such as TUCKER3 [Kie00] and many others [CZPA09]. Among the

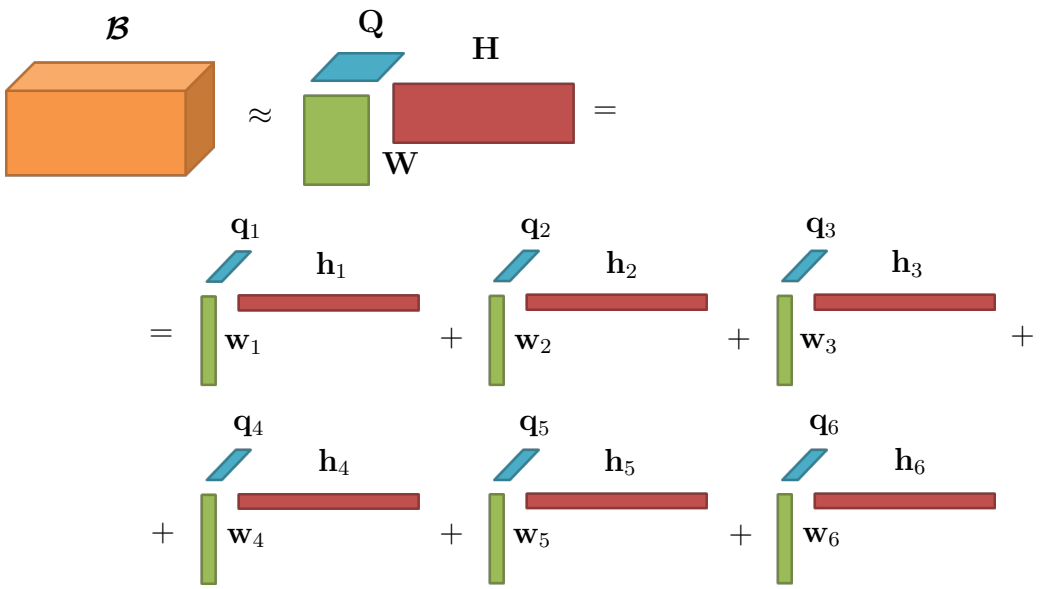


FIGURE 1.3 – NTF (CANDECOMP / PARAFAC) with $K = 6$ as a sum of rank-1 tensors.

most popular ones there is a so-called CANDECOMP or PARAFAC model [Bro97]. Since this is the only NTF model we are using in this work, we will describe only this model and will refer to it as NTF throughout this document.

The data tensor $\boldsymbol{\mathcal{B}}$ is approximated as a sum of K rank-1 3-way nonnegative tensors as

$$\boldsymbol{\mathcal{B}} \approx \boldsymbol{\mathcal{V}} = \sum_{k=1}^{K} \mathbf{q}_k \circ \mathbf{w}_k \circ \mathbf{h}_k^T, \tag{1.20}$$

with $\mathbf{Q}$ being a $J \times K$ nonnegative matrix, $\mathbf{q}_k$ being its k-th column, $\mathbf{W}$ and $\mathbf{H}$ being defined as before for NMF, and $\circ$ denoting tensor outer product. This decomposition is represented on Figure 1.3.

Similarly to NMF, NTF parameters are found by optimizing

$$(\mathbf{Q}, \mathbf{W}, \mathbf{H}) = \arg \min_{\mathbf{Q}' \geq 0, \mathbf{W}' \geq 0, \mathbf{H}' \geq 0} C(\mathbf{Q}', \mathbf{W}', \mathbf{H}') \tag{1.21}$$

$$= \arg \min_{\mathbf{Q}' \geq 0, \mathbf{W}' \geq 0, \mathbf{H}' \geq 0} D(\boldsymbol{\mathcal{B}} \| \boldsymbol{\mathcal{V}}), \tag{1.22}$$

with $D(\boldsymbol{\mathcal{B}} \| \boldsymbol{\mathcal{V}})$ being some divergence, and we consider here only the IS divergence, and $\boldsymbol{\mathcal{V}}$ being specified as in (1.20).

It is interesting to note that, in contrast to NMF, for NTF (with $L > 2$) the conditions for uniqueness of solution of (1.22) are much milder [Kru77, LC10].

1.2.2 Algorithms

Similarly to NMF, various optimization strategies are possible to optimize (1.22). We here summarize the MU rules for the case of IS divergence. In case of NTF these rules are easier to be formulated in scalar form and consist in alternating between the

following updates :

$$q_{jk} \leftarrow q_{jk} \frac{\sum_{f,n=1}^{F,N} w_{fk} h_{kn} b_{jfn} v_{jfn}^{-2}}{\sum_{f,n=1}^{F,N} w_{fk} h_{kn} v_{jfn}^{-1}}, \tag{1.23}$$

$$w_{fk} \leftarrow w_{fk} \frac{\sum_{j,n=1}^{J,N} q_{jk} h_{kn} b_{jfn} v_{jfn}^{-2}}{\sum_{j,n=1}^{J,N} q_{jk} h_{kn} v_{jfn}^{-1}}, \tag{1.24}$$

$$h_{kn} \leftarrow h_{kn} \frac{\sum_{j,f=1}^{J,F} q_{jk} w_{fk} b_{jfn} v_{jfn}^{-2}}{\sum_{j,f=1}^{J,F} q_{jk} w_{fk} v_{jfn}^{-1}}, \tag{1.25}$$

where $\boldsymbol{\mathcal{V}} = [v_{jfn}]_{j,f,n=1}^{J,F,N}$, and it is recomputed as in (1.20) after each update.

1.2.3 Application to multichannel source separation

In the multichannel scenarios it is assumed that the sources are recorded by several $I > 1$ microphones. In this case each source goes in general through a different acoustic path to attend each of microphones. This so-called spatial diversity is usually exploited by a source separation algorithm on top of source characteristics (as in the single channel case) to achieve better separation quality. In multichannel scenarios a distinction is often made between the (over-)determined case ($I \geq J$: at least as much mixtures as sources) and the under-determined case ($I < J$: less mixtures than sources), which is more difficult.

Previous works utilizing NTF for multichannel source separation [FCC05, PE06] were applying it directly to the spectrograms of channels (or mixtures) stuck into a 3-valence tensor. Indeed, this is quite intuitive as idea, since this is almost the only observed nonnegative 3-valence tensor available in this case. Without going here into the details of how the sources and the mixtures are related, let us assume that $\{\mathbf{X}_i\}_{i=1}^{I}$ are the STFTs of I mixtures. They are then stuck to form a 3-valence tensor $\boldsymbol{\mathcal{X}} = [x_{ifn}]_{i,f,n=1}^{I,F,N}$, which is modeled as

$$|\boldsymbol{\mathcal{X}}|^{.2} \approx \boldsymbol{\mathcal{V}} = \sum_{k=1}^{K} \mathbf{q}_k \circ \mathbf{w}_k \circ \mathbf{h}_k^T, \tag{1.26}$$

with $\mathbf{Q} \in \mathbb{R}_+^{I \times K}$. Each entry q_{ik} of matrix $\mathbf{Q}$ represents the contribution of the k-th component of the decomposition into the i-th channel. Similarly to NMF, optimization of this model with IS divergence is equivalent to the ML criterion optimization assuming that

$$x_{ifn} \sim \mathcal{N}_c\left(0, \left[\sum_{k=1}^{K} \mathbf{q}_k \circ \mathbf{w}_k \circ \mathbf{h}_k^T\right]_{ifn}\right), \tag{1.27}$$

and all x_{ifn} are mutually independent.

The main drawbacks of the multichannel NTF modeling (1.26) are :

1. This fully nonnegative decomposition of the mixture (power) spectrogram ignores completely the STFT phase, while the phase modeling is very important for audio source separation, especially under far-field assumptions (i.e., when the distances between sources and microphones are considerably greater than the distances between microphones) [11].
2. Since the coefficients q_{ik} are not varying over frequency, this decomposition is limited to model instantaneous mixtures (not convolutive ones), where the sources

are simply added after multiplication by some scalar gains (not filtering). This makes its applicability to real world scenarios very limited.

3. Finally, within IS-NTF probabilistic formulation (1.27) it is assumed that all STFT coefficients x_{ifn} are mutually independent, i.e., over time, frequency and channels. While independence over time and frequency might be a good approximation of the reality, the independence over channels is certainly a too coarse approximation. Indeed, since the same sources are mixed up in different channels, the channels are not independent.

These drawbacks will be addressed within multichannel multisource NMF/NTF modeling we will present in the next chapter.

1.3 Conclusion

In this chapter we have presented some basics on NMF and NTF modeling with a strong focus on IS divergence that has a very attractive Gaussian interpretation. We have discussed the applicability of this modeling to audio processing and especially to single-channel and multichannel audio source separation. We have revealed and discussed the limits of existing approaches in the multichannel case.

Chapitre 2

Proposed multichannel multisource NMF/NTF modeling

In this chapter I introduce a general formulation of multichannel multisource NMF/NTF modeling that unifies to some extend all the models we introduced in papers [1, 2, 3, 4].

2.1 Main idea

Recall that previous state-of-the-art attempts to use NTF for multichannel source source separation are relying on NTF modeling of tensor of multichannel mixture spectrograms. As already mentioned, such a modeling is only a very coarse approximation of multichannel mixing and it is often not reliable at all (e.g., for convolutive mixtures). However, as discussed in Section 1.1.3, NMF remains a very good model for source spectrograms.

Our main idea relies on a sort of semi-nonnegative modeling and consists in

— either modeling power spectraogram of each source with IS-NMF (*multichannel NMF*) as in (1.16),

— or modeling power spectraograms of all sources stuck in a 3-valence tensor $\boldsymbol{\mathcal{S}} = [s_{jfn}]_{j,f,n=1}^{J,F,N}$ with IS-NTF (*multichannel NTF*) as

$$|\boldsymbol{\mathcal{S}}|^{\cdot 2} \approx \boldsymbol{\mathcal{V}} = \sum_{k=1}^{K} \mathbf{q}_k \circ \mathbf{w}_k \circ \mathbf{h}_k^T, \tag{2.1}$$

while modeling the mixing process directly in the signal domain instead of the domain of nonnegative spectrograms. This becomes possible thanks to the Gaussian interpretation of IS-NMF, as in (1.17), or of IS-NTF as

$$s_{jfn} \sim \mathcal{N}_c\left(0, \left[\sum\nolimits_{k=1}^{K} \mathbf{q}_k \circ \mathbf{w}_k \circ \mathbf{h}_k^T\right]_{jfn}\right), \tag{2.2}$$

which specifies a Gaussian distribution of the sources $\boldsymbol{\mathcal{S}}$ (not their power spectrograms $|\boldsymbol{\mathcal{S}}|^{\cdot 2}$), thus allowing modeling mixing process in the signal domain. Moreover, as we will see, such a modeling allows handling other distortions such as missing samples or quantization, which seems to be almost impossible within a fully nonnegative framework.

To summarize in two words, instead of modeling the observed mixture spectrogram tensor with NTF as in (1.26) we propose NTF modeling of the latent source spectrogram tensor as in (2.1).

2.2 Assumptions and modeling

Let us now present our general formulation unifying models in [1, 2, 3, 4]. We assume that there are J audio sources going altogether through a combination of various linear transforms such as for example :

- filtering (appears, e.g., in convolutive audio source separation [1, 2]),
- summation (appears, e.g., in audio source separation [1, 2, 3]),
- STFT and inverse STFT (useful for switching between STFT and time domains, as needed in [3]),
- subsampling (appears, e.g., within problems with missing samples like audio declipping or more generally audio inpainting [3]).

As a final step, a quantization might be applied, which is necessary in audio compression or simply to store the result. At the end we get an M_x-length vector $\mathbf{x}$ of observations, which is a concatenation of all resulting samples. The result may be in any domain : time, time-frequency or other. For the sake of generality $\mathbf{x}$ is assumed complex-valued, i.e., $\mathbf{x} \in \mathbb{C}^{M_x}$. Let again $\boldsymbol{\mathcal{S}}$ a 3-valence tensor of source STFT coefficients, which is unknown in general. Since the sources may be computed from $\boldsymbol{\mathcal{S}}$ via the inverse STFT transform, which is linear, and all other transforms applied to sources are linear (except the quantization), source STFTs $\boldsymbol{\mathcal{S}}$ and the resulting observations $\mathbf{x}$ are related (up to the quantization step) by a linear transform. This can be written as

$$\mathbf{x} = \mathbf{A}\,\mathrm{vec}(\boldsymbol{\mathcal{S}}) + \mathbf{z}, \tag{2.3}$$

where $\mathrm{vec}(\cdot)$ is an operator vectorizing a tensor of size $J \times F \times N$ into a vector of size $M_s = J \cdot F \cdot N$, $\mathbf{A} \in \mathbb{C}^{M_s \times M_x}$ is a matrix representing the resulting linear transform, and $\mathbf{z} \in \mathbb{C}^{M_x}$ is a quantization noise or any other noise.

It is further assumed that the noise components z_m $(m = 1, \ldots, M_x)$ are mutually independent and each component follows a zero-mean Gaussian distribution

$$z_m \sim \mathcal{N}_c(0, \sigma^2_{z,m}), \tag{2.4}$$

with a fixed variance $\sigma^2_{z,m}$ assumed to be known. The variance $\sigma^2_{z,m}$ depends on the component's index m, e.g., to be able handling non-uniform quantization. Gaussian assumption is not completely true for quantization, and mutual independence assumption is only true for scalar quantization (not for vector one). However these approximations are reasonable.

It is also assumed that all the entries of $\mathrm{vec}(\boldsymbol{\mathcal{S}})$ are mutually independent and each entry follows a zero-mean Gaussian distribution as in (2.2). As such, this is a multichannel NTF modeling (2.1) which is in fact a generalization of multichannel NMF modeling (1.16). Indeed, if each column of matrix $\mathbf{Q}$ in (2.1) is supposed to be normalized such as it sums to 1, each q_{jk} represents the contribution of the k-th rank-1 component into the modeling of the j-th source. Now, assuming each column of matrix $\mathbf{Q}$ has all entries but one equal to zero, i.e. each rank-1 component contributes into the modeling of just one source, one can easily show that (2.1) reduces to (1.16) (up to some trivial permutation issues). In light of the above explanation, multichannel NTF has the following potential advantages over multichannel NMF :

- one rank-1 component may contribute into modeling of several sources, and
- in contrast to multichannel NMF, where one needs to specify the number of components K_j for each source, for multichannel NTF one only needs to define the

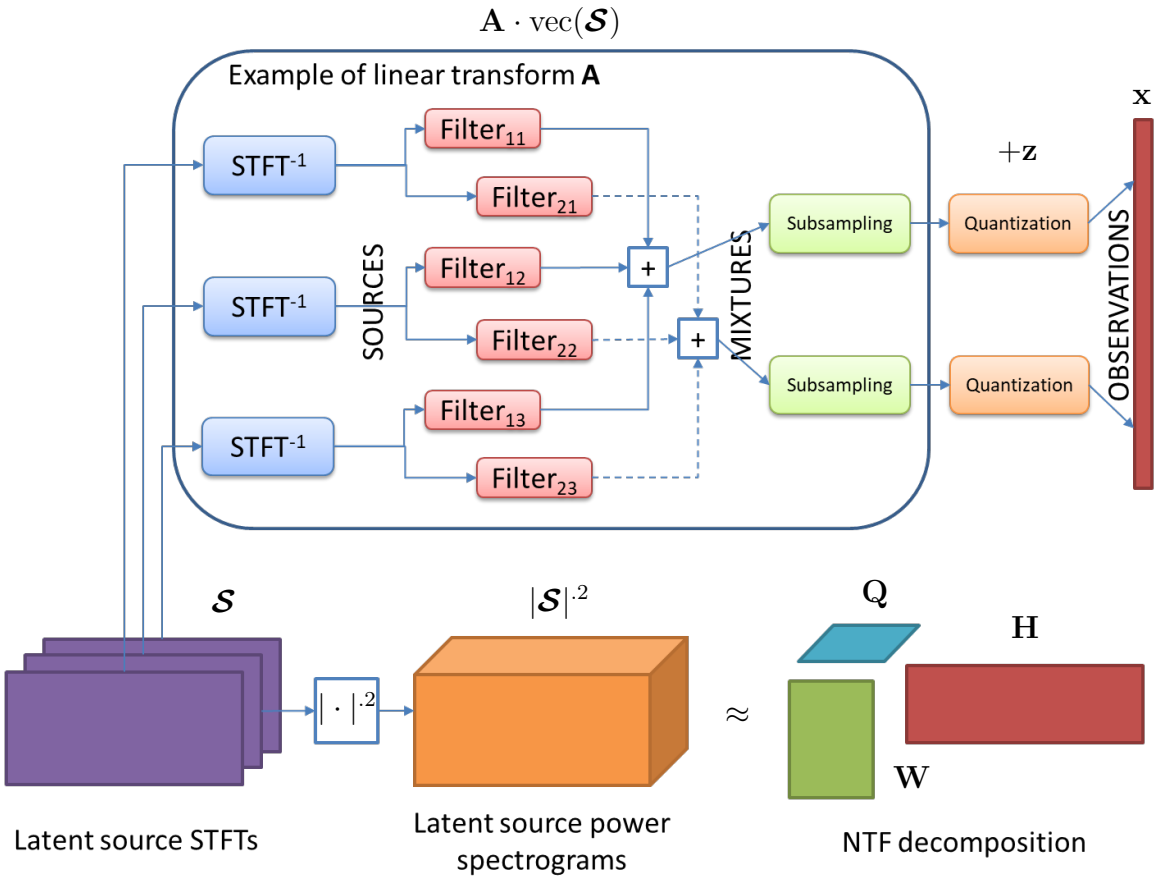


FIGURE 2.1 – General multichannel multisource NTF modeling together with an example of linear transform $\mathbf{A}$, where 3 sources are convolutively mixed into 2 channels in time domain and then subsampled.

total budget of components K, which is then automatically distributed between sources via learning of matrix $\mathbf{Q}$.

As such, we here use multichannel NTF rather than multichannel NMF.

The above-described general modeling is schematized on Figure 2.1 together with an example of linear transform $\mathbf{A}$, where 3 sources are convolutively mixed into 2 channels in time domain and then subsampled. However, this is just one example and there are many other possibilities, e.g., one may not going back to time domain with inverse STFT and do filtering directly in the STFT domain, as in [1, 2, 3]. In the latter case the observations $\mathbf{x}$ are in the STFT domain as well.

Given the model specified above and the observations $\mathbf{x}$, the goal is to estimate both the parameters of NTF decomposition $(\mathbf{Q}, \mathbf{W}, \mathbf{H})$ and the tensor of latent source STFTs $\boldsymbol{\mathcal{S}}$. The sources may be then reconstructed in time domain via the inverse STFT.

It should be noted right away that this formulation is not practical at all since, even though everything is Gaussian, thus leading to tractable computations, the dimensions $M_s = J \times F \times N$ and M_x of matrix $\mathbf{A}$ are huge in practice making the computations unfeasible. However, as we will see later on, this will be taken care about by introducing some approximations/simplifications. For the moment we keep the formulation as it is for the sake of generality.

On the other hand, this formulation only partially generalizes the models from [1, 2, 3, 4]. More precisely :

— As for convolutive multichannel source separation [1, 2], the transform $\mathbf{A}$ is not completely known, since it is based on the parameters of mixing filters (see Fig. 2.1) that are not known in general and must be estimated as well.

— Several other nonnegative structures, not only NMF or NTF, are considered in [2] to approximate the latent source power spectrograms $|\boldsymbol{\mathcal{S}}|.^2$.

— A so-called *full-rank spatial model* [DVG10a] [2] is not covered by this formulation, since it models so-called *non-point* sources, i.e., sources that are slightly decorrelated over channels.

However, since generalizing everything is difficult (though not impossible), we keep our current formulation as it is, and all those differences will be discussed case by case.

2.3 Estimation criteria and algorithms

2.3.1 Estimation criteria

Let $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$ denote the whole set of parameters of the NTF model. The estimation is usually performed in two steps :

— The model $\boldsymbol{\theta}$ is estimated in the ML sense, i.e., maximazing the likelihood of the observations given the model, as :

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}'} p(\mathbf{x}|\boldsymbol{\theta}'). \tag{2.5}$$

In case some prior distribution on parameters $\boldsymbol{\theta}$ is given, maximum a posteriori (MAP) estimation may be used instead [2].

— Given the estimated model $\boldsymbol{\theta}$, the sources $\boldsymbol{\mathcal{S}}$ are estimated with the MMSE estimator as :

$$\widehat{\boldsymbol{\mathcal{S}}} = \mathbb{E}[\boldsymbol{\mathcal{S}}|\mathbf{x}; \boldsymbol{\theta}]. \tag{2.6}$$

2.3.2 Algorithms

Note that since everything is Gaussian and the transform $\mathbf{A}$ is linear, the posterior distribution of $\mathrm{vec}(\boldsymbol{\mathcal{S}})$, given the observations and the model, is Gaussian as well, and it can be shown [1] to be expressed as

$$p(\mathrm{vec}(\boldsymbol{\mathcal{S}})|\mathbf{x}; \boldsymbol{\theta}) = N_c(\mathrm{vec}(\boldsymbol{\mathcal{S}}); \mathrm{vec}(\widehat{\boldsymbol{\mathcal{S}}}), \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}^{\mathrm{post}}), \tag{2.7}$$

with $N_c(\cdot;\cdot,\cdot)$ defined in (1.12), and posterior mean $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{S}}}) \in \mathbb{C}^{M_s}$ and posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}^{\mathrm{post}} \in \mathbb{C}^{M_s \times M_s}$ computed as

$$\mathrm{vec}(\widehat{\boldsymbol{\mathcal{S}}}) = \mathbf{G}\mathbf{x}, \tag{2.8}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}^{\mathrm{post}} = \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}} - \mathbf{G}\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}, \tag{2.9}$$

$$\mathbf{G} = \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}\mathbf{A}^H(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}\mathbf{A}^H + \boldsymbol{\Sigma}_{\mathbf{z}})^{-1}, \tag{2.10}$$

where [1]

$$\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}} = \mathrm{diag}(\mathrm{vec}(\boldsymbol{\mathcal{V}})), \tag{2.11}$$

1. Within this document the operation $\mathrm{diag}(\cdot)$ when applied to a square matrix means a column vector consisting of the elements of the diagonal of this matrix, while when applied to a vector means a diagonal matrix with this vector on the diagonal.

with $\boldsymbol{\mathcal{V}}$ as in (2.1),

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \mathrm{diag}([\sigma_{z,1}^2, \ldots, \sigma_{z,M_x}^2]), \tag{2.12}$$

and matrix $\mathbf{G}$ is a so-called Wiener filter gain. As such, this already gives us the solution for the source estimation in (2.6), which is obtained via Wiener filtering [Kay93] as in (2.8). We will now present a way for optimizing the ML criterion (2.5) to find model parameters.

The likelihood in (2.5) writes

$$p(\mathbf{x}|\boldsymbol{\theta}) = N_c(\mathbf{x}; \underline{0}, \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}\mathbf{A}^H + \boldsymbol{\Sigma}_{\mathbf{z}}), \tag{2.13}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}$ and $\boldsymbol{\Sigma}_{\mathbf{z}}$ specified in (2.11) and (2.12), respectively ; and there is no closed-form solution maximizing it over $\boldsymbol{\theta}$. As such, optimization strategies such as the expectation maximization (EM) algorithm [DLR77] or more generally the MM algorithm [HL04] are usually used. An MM approach is proposed in [SKAU13] and many variants of EM are possible depending, e.g., on the choice of the latent data [12, 13]. We here detailed just one EM algorithm variant that is suitable in most cases and has quite simple formulation and interpretation. Note that this EM algorithm does not always coincide with algorithms described in [1, 2, 3, 4] for the corresponding models, but it is applicable for all those models.

This algorithm is referred to as *GEM-MU* [2, 4], since it is rather a generalized EM (GEM) algorithm [DLR77] (i.e., the maximization step does not maximize the corresponding auxiliary function, but only insures it is non-decreasing under parameters update), and it is based on the MU rules to update NTF model parameters within the maximization step. The algorithm consists simply in iterating between computing the conditional expectation of latent source powers spectrogram tensor $\boldsymbol{\mathcal{P}} = |\boldsymbol{\mathcal{S}}|^{\cdot 2}$ (E-step), and updating the NTF model parameters $\boldsymbol{\theta}$ with MU rules (1.23), (1.24), (1.25) while approximating the estimated tensor (M-step). More precisely :

— *E-step :* Compute conditional expectation of source power spectrograms :

$$\mathrm{vec}(\widehat{\boldsymbol{\mathcal{P}}}) = \mathbb{E}[\mathrm{vec}(|\boldsymbol{\mathcal{S}}|^{\cdot 2})|\mathbf{x}; \boldsymbol{\theta}] = \mathrm{vec}(|\widehat{\boldsymbol{\mathcal{S}}}|^{\cdot 2}) + \mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}^{\mathrm{post}}), \tag{2.14}$$

with $\widehat{\boldsymbol{\mathcal{S}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{S}}}^{\mathrm{post}}$ computed as in (2.8) and (2.9), respectively.

— *M-step :* Apply one or several iterations of MU rules (1.23), (1.24), (1.25), while substituting data tensor $\boldsymbol{\mathcal{B}}$ by $\widehat{\boldsymbol{\mathcal{P}}}$.

Let us note again that this algorithm is not practical at all since requires inversion and multiplication of matrices of very high dimensions M_s and M_x (see, e.g., (2.8), (2.9) and (2.10)). However, this is avoided in each particular case by employing some approximations that make matrix $\mathbf{A}$ block-diagonal either over time frames [4] or over both time frames and frequency bins [1, 2, 3].

2.3.3 Summary of applications

We have shown the advantage of the proposed multichannel multisource NMF/NTF modeling over the state of the art for a range of applications, including :

— blind source separation [1],

— supervised and semi-supervised source separation [2], where this models may be combined with other models in a flexible and systematic manner,

— informed source separation or, saying differently, audio objects compression [3, 4],

— audio inpainting including audio declipping and compressive sampling recovery [4].

These applications will be discussed below in details.

2.4 Conclusion

We have presented a quite general formulation of a multichannel multisource NTF modeling unifying to some extend the models developed in [1, 2, 3, 4] that will be discussed in the following chapters. Though this very general formulation is computationally intractable, approximations will be introduced to overcome this.

Chapitre 3

Paper 1 : Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation

This work was done in collaboration with Cédric Févotte. In [1] we have introduced multichannel NMF model and applied it with success to blind source separation of convolutive and instantaneous mixtures, as well as to blind separation of professionally produced stereo music recordings. The model was then extended to multichannel NTF in [7]. In my opinion this is a key contribution in the domain, and it can be briefly resumed as an extension of NMF modeling applicability from single-channel source separation to multichannel source separation. This statement might be supported by the fact that the paper [1] has been extensively cited (538 citations according to Google Scholar on November 19, 2019) and has received the IEEE Signal Processing Society Best Paper Award in 2014.

3.1 Audio source separation of multichannel mixtures

Convolutive mixing is one of the most realistic mixing models for static (i.e., non-moving) sources. As such, this kind of mixing is very often considered in audio source separation. In contrast to single-channel mixing (1.14), where the sources are just added, it is assumed that I mixture signals $\tilde{x}_i(t)$ $(i = 1, \ldots, I)$ are obtained from J sources $\tilde{s}_j(t)$ $(j = 1, \ldots, J)$ through a convolutive mixing as :

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau) \tilde{s}_j(t-\tau) + \tilde{z}_i(t), \tag{3.1}$$

where $\tilde{a}_{ij}(t)$ is the finite-impulse response of some (causal) filter and $\tilde{z}_i(t)$ is some additive noise (e.g., quantization noise). The goal is again to estimate sources $\tilde{s}_j(t)$ from the known mixtures $\tilde{x}(t)$. In general, the convolutive filters coefficients $\tilde{a}_{ij}(t)$ are not known either.

Instantaneous mixing is also often considered for research purposes, and it is a simplified version of convolutive mixing, where each filter is replaced by multiplication by just one gain as

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \tilde{a}_{ij} \tilde{s}_j(t) + \tilde{z}_i(t). \tag{3.2}$$

This mixing is much less realistic and occurs in practice quite rarely. One exception are artificially-mixed professionally produced music recordings, where a so-called "pan pot" mixing corresponds to instantaneous mixing. However, this is rather valid for some old recordings, in modern recordings a lot of reverberation and other effects are added.

There is also a distinction between (over-)determined ($I \geq J$) and under-determined ($I < J$) cases. The under-determined case is obviously more challenging and we were mostly targeting this case, while developing multichannel NMF. However, it is also applicable without restriction in the (over-)determined case.

Convolutive mixing equation (3.1) may be rewritten in the STFT domain as

$$x_{ifn} = \sum_{j=1}^{J} a_{ijf} s_{jfn} + z_{ifn}, \tag{3.3}$$

with x_{ifn}, s_{jfn} and z_{ifn} being, respectively, STFT coefficients of mixtures, sources and noise; and a_{ijf} being coefficients of discrete Fourier transform (DFT) of filters $\tilde{a}_{ij}(t)$. Equation (3.3) is verified only approximately, and it is referred to as *narrowband approximation*. It holds when the filter length L is "significantly" shorter than the STFT window size [PS00]. Equation (3.3) can be also rewritten in matrix form as

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{z}_{fn}, \tag{3.4}$$

with vectors $\mathbf{x}_{fn} = [x_{1fn}, \ldots, x_{Ifn}]^T$, $\mathbf{s}_{fn} = [s_{1fn}, \ldots, s_{Jfn}]^T$ and $\mathbf{z}_{fn} = [z_{1fn}, \ldots, z_{Ifn}]^T$; and matrix $\mathbf{A}_f = [a_{ijf}]_{i,j=1}^{I,J}$.

One can note that convolutive mixing (3.1) reduces to instantaneous mixing (like (3.2)) for each frequency bin f in the STFT domain (3.3). However, even if one manages to separate instantaneous mixtures in each frequency bin, one still needs correctly grouping individual bin-wise source estimates to reconstruct sources globally. This problem is usually referred to as *permutation alignment problem* [SAM10].

Prior to our work, several methods were proposed to solve convolutive mixing source separation formulated in the STFT domain under narrowband approximation (3.3). However, non of those approaches were addressing this problem globally in a principle way. For example, one of the best approaches proposed by Sawada *et al.* [SAM10] consists in solving instantaneous mixing (3.4) by an independent component analysis (ICA)-like method [OP04] for each frequency bin, and then in solving permutation alignment by grouping sources according to their temporal correlation. As we have already mentioned at the end of Section 1.2.3, state-of-the-art NTF-based methods are as well suffering from numerous drawbacks.

Multichannel NMF allows overcoming all above-mentioned shortcomings. In particular, in contrast to [SAM10], permutation alignment and frequency-wise source estimation are addressed jointly within a global probabilistic Gaussian modeling.

3.2 Multichannel NMF model

Assuming convolutive mixing narrowband approximation (3.4) and sources distributed as in (1.17) we obtain multichannel NMF modeling as proposed in [1]. Similarly, according to our extension in [7], multichannel NTF is obtained by assuming the sources distributed as in (2.2).

A schematic representation of multichannel NMF is given on Figure 3.1. On can see that this representation is a partial case of our general formulation on Figure 2.1,

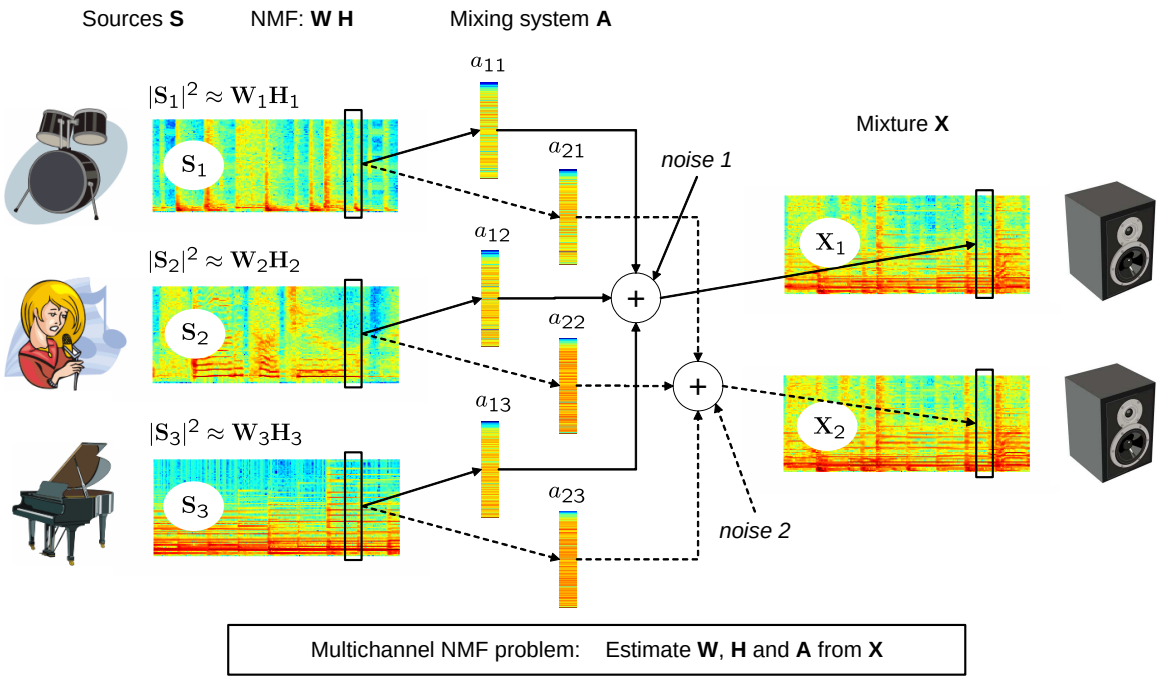


FIGURE 3.1 – Representation of convolutive mixing system and formulation of Multichannel NMF problem (figure from [1]).

except that there is no subsampling and everything (i.e., modeling and observations) is formulated in the STFT domain. Mixing equation (3.4) corresponds to equation (2.3) in the general formulation, where, thanks to narrowband approximation, the linear transform $\mathbf{A}$ becomes block-diagonal over time and frequency so as the computations in (2.8), (2.9) and (2.10) simplify to multiplication and inversion of matrices/vectors of size I or J. This allows developing very efficient parameter optimization strategies.

While the original GEM algorithm described in [1] is different from the GEM-MU algorithm described in Section 2.3.2, the latter is applicable as well with only difference that the E-step and the M-step should be completed to allow updating the mixing parameters $\mathbf{A}_f$. The corresponding detailed implementation can be found in [7], where it is also shown that the GEM-MU algorithm converges faster than the original GEM from [1] (see Fig. 1. in [7]).

3.3 Results

We have evaluated the proposed approach on both instantaneous and convolutive mixtures. Though, as we have already mentioned, multichannel NMF solves permutation alignment and frequency-wise source estimation jointly, it is still quite sensible to the initialization of model parameters. As such, we used state-of-the-art reference algorithms to obtain good initializations. For instantaneous mixtures we used the algorithm by Vincent [Vin07], and for convolutive ones the algorithm by Sawada *et al.* [SAM10]. We have shown that multichannel NMF improves the source separation results over the reference algorithms in both instantaneous and convolutive cases (see Table II in [1]).

Moreover, the proposed approach has been evaluated on the corresponding tasks of the international Signal Separation Evaluation Campaign (SiSEC 2008) [VAB09].

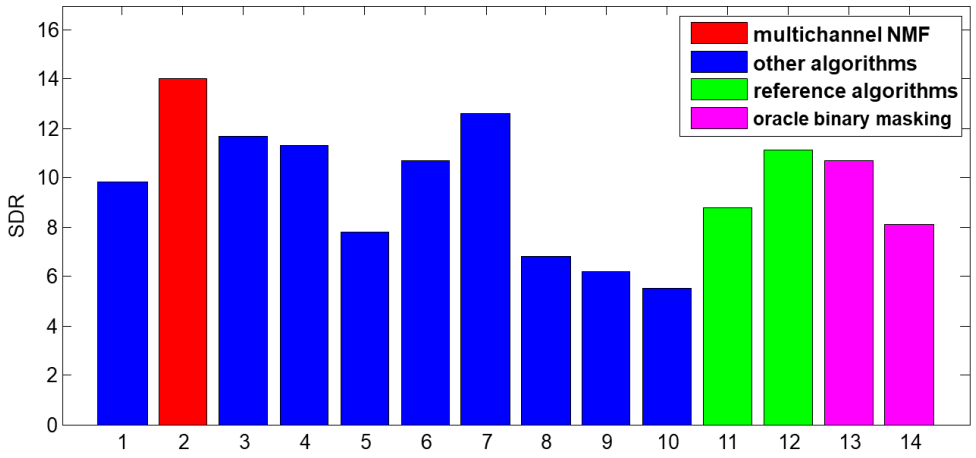


FIGURE 3.2 – Results for "Under-determined speech and music mixtures - instantaneous mixtures" SiSEC 2008 [VAB09] task. The results are plotted in terms of source to distortion ratio (SDR) [VGF06] (higher is better).

In particular, the approach has shown the best results for "Under-determined speech and music mixtures - instantaneous mixtures" SiSEC 2008 task among 10 competing algorithms, see Figure 3.2 and/or Table 2 in [VAB09]. It is also interesting to note from Figure 3.2 that the multichannel NTF outperforms considerably the so-called *oracle binary masking*, where the sources are estimated using an "ideal" binary mask derived based on the knowledge of true sources ("oracles"). This showcase (when a non-oracle estimation outperforms an oracle one) demonstrates the great potential of multichannel Wiener filtering (2.8) over a simple (but limited) binary masking.

3.4 Impact and followings

This work had a quite important impact in our scientific community, since, as it was already mentioned, it has been extensively cited and the paper [1] has received the IEEE Signal Processing Society Best Paper Award in 2014. Moreover, the multichannel NMF modeling was followed, extended or applied to other problems by myself (with other co-authors, as we will see below) or other researchers. Let us just list few examples :

— We have extended multichannel NMF to a more general audio source separation framework [2] allowing it to be combined with other models and to be applied in other source separation scenarios (e.g., supervised or semi-supervised).

— We have applied it to other problems [3, 4] such as audio objects compression or audio inpainting (e.g., declipping).

— Sawada *et al.* [SKAU13] proposed different (MM-based) algorithms for multichannel NMF, and have also extended it to a case of a generalized multichannal EUC distance.

— Kitamura *et al.* [KOS+16] have proposed more efficient algorithms for multichannel NMF in case of (over-)determined mixtures.

— Several researchers [LGH19, KLIM19] have pushed multichannel NMF into deep learning world by replacing the latent NMF models with pre-trained variational autoencoders (VAEs) [KW13].

— Several book chapters [12, 13] [KSH18] have been written on this topic.

3.5 Conclusion

We have presented multichannel NMF modeling as developed for and applied to blind audio source separation in convolutive and instantaneous mixtures. We have shown and discussed some results and improvements over the corresponding sate-of-the-art methods. This work had a quite significant impact in our scientific community, and, in particular, it was well remarked and followed.

Chapitre 4

Paper 2 : A general flexible framework for the handling of prior information in audio source separation

This work was done in collaboration with Emmanuel Vincent and Frédéric Bimbot. We have remarked that many audio source separation approaches within various scenarios (single-channel or multichannel ; blind, semi-supervised or supervised ; etc ...) fall within a so-called *local Gaussian modeling (LGM)* framework that includes and generalizes the multichannel NMF [1]. Though those approaches are not all based on NMF spectral source modeling. For example, some of them are based on Gaussian mixture models (GMMs) or hidden Markov models (HMMs) as spectral source models and on various spatial models. Motivated by this observation we have developed a flexible framework [2] allowing combining various spectral and spatial source models in a systematic manner. Moreover, the parameters of those models may be fully or partly fixed or learned from the observed mixture, while optionally given some prior distribution. Those choices should be based on the available prior knowledge about the source separation problem, including availability of training data from which some parameters might be pre-trained. We have developed a suitable GEM algorithm allowing estimating model parameters whatever the model specification. A corresponding software implementation called Flexible Audio Source Separation Toolbox (FASST) has been developed in Matlab and is available at [14]. It was later re-implemented in C++ and Python. FASST toolbox is used by researchers and engineers from the community, and the corresponding paper [2] has been well remarked (298 citations according to Google Scholar on November 19, 2019). In particular this work shows the ability of multichannel NMF, due to its probabilistic Gaussian nature, to be combined with other Gaussian models and to be enforced by probabilistic priors.

4.1 Local Gaussian modeling

Let us first explain local Gaussian modeling (LGM) framework that is a more general concept than multichannel NMF/NTF. Convolutive mixing equation (3.4) may be rewritten as

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{jfn} + \mathbf{z}_{fn}, \qquad \mathbf{y}_{jfn} = \mathbf{a}_{jf} s_{jfn}, \tag{4.1}$$

where $\mathbf{a}_{jf}$ is the j-th column of matrix $\mathbf{A}_f$, and vector $\mathbf{y}_{jfn} \in \mathbb{C}^I$ is a so-called *spatial image* of j-th source consisting in contributions of this source to each mixture. Very often in audio source separation we are interested in estimating those spatial images rather than the sources s_{jfn} themselves, since anyway there is a great scale ambiguity between the sources and the mixing coefficients a_{ijf}. Now we assume that the source coefficients s_{jfn} are mutually independent and each coefficient follows a zero-mean Gaussian distribution

$$s_{jfn} \sim \mathcal{N}_c(0, v_{jfn}), \tag{4.2}$$

where variances v_{jfn} may be structured as before via NMF as in (1.16) or NTF as in (2.1), but more generally can be structured differently.

Using (4.1) and (4.2) one can easily show that the spatial source image vector $\mathbf{y}_{jfn}$ is distributed as

$$\mathbf{y}_{jfn} \sim \mathcal{N}_c(\underline{0}, \mathbf{R}_{jf} v_{jfn}), \tag{4.3}$$

where $\mathbf{R}_{jf} = \mathbf{a}_{jf}\mathbf{a}_{jf}^H$ is a matrix of rank 1. After we have published multichannel NMF paper [1], the Duong *et al.* [DVG10a] have baptized this model (i.e., $\mathbf{R}_{jf} = \mathbf{a}_{jf}\mathbf{a}_{jf}^H$) *rank-1 spatial model* and they have proposed to consider a so-called *full-rank spatial model*, where matrix $\mathbf{R}_{jf}$ is not restricted to be expressed as $\mathbf{a}_{jf}\mathbf{a}_{jf}^H$, can be of any rank, and can be left free during estimation or constrained differently (see [DVG10a] for details). The full-rank modeling allowed to model *non-point sources* (as opposed to *point sources* that a located in single points), e.g., like a piano playing in a room. Moreover, even in case of point sources, it was shown to improve source separation performance in case of long reverberations (greater than STFT window size) when the narrowband approximation (3.3) becomes less exact, and in case of slightly moving (non-static) sources [DVG10a].

To summarize, the modeling (4.3), where parameters consisting of

— *spatial covariances* $\mathbf{R}_{jf}$, and

— *spectral variances* v_{jfn}

may be structures somehow or given some prior distribution, is called *local Gaussian modeling (LGM)* [VAG09]. Many existing approaches fall within this quite general framework.

Note that, as it was already mentioned at the end of Section 2.2, our general multichannel NTF formulation in Chapter 2 does not generalize the LGM for the following reasons :

— First, in case of LGM with full-rank spatial covariances $\mathbf{R}_{jf}$ there is no more notion of point source coefficients s_{jfn} and they cannot be even estimated properly. Indeed, since $\mathbf{R}_{jf}$ is not of rank 1, the entries of spatial image vector $\mathbf{y}_{jfn}$ in (4.3) are decorrelated over channels.

— Second, within LGM modeling spectral variances v_{jfn} are not necessarily structured by NMF of NTF.

4.2 Motivation

First, let us give an intuition why in 2010, when we have started working on this project, it was a right time to propose such a general framework. The point is that before (say before 2009) different audio source separation problems were treated by qualitatively different methods. For example, single-channel separation was often treated by methods based on NMF, while multichannel separation ((over-)determined or

under-determined) by methods based on ICA [Com94]. It was not clear at all how to use ICA in the single-channel case or how to use NMF in the multichannel case. As such, a potential unification of approaches was difficult. Though it is still not clear how to use ICA in the single-channel case, since introduction of multichannel NMF [1] it became clear that all those scenarios may be tackled at least by NMF modeling, which opened us a door and gave us an inspiration for a possible generalization.

Moreover, we have remarked that many state-of-the-art approaches (at least 16 approaches, see Table I of [2]), including multichannel NMF [1], fall into the LGM framework (4.3) and may be classified according to possible combinations of the following characteristics :

— *Problem dimensionality :* single-channel, under-determined, or (over-)determined.

— *Level of supervision :* non-supervised (or blind), semi-supervised, or supervised.

— *Mixing type :* instantaneous or convolutive.

— *Spatial covariance model :* rank-1 [1] or full-rank [DVG10a].

— *Spectral variance model :* NMF [1], harmonic NMF [VBB09], GMM [15], HMM [Att03] or source-filter model [DRDF10].

— *Signal representation :* linear (e.g., STFT) or quadratic [HBB92] (e.g., equivalent rectangular bandwidth (ERB) [DVG10b]).

We see that all possible combinations of these characteristics generates already a lot of possibilities, and the number of possibilities is yet greater since each source can be modeled with a particular model (e.g., in terms of spatial covariance model and spectral variance model).

Our main motivation was to propose a general framework allowing incorporating easily any combination of those characteristics in a principle way within a generic and flexible implementation. In other words, while current approach was often consisting in specifying a necessary combination of characteristics from the above list (model design), designing a suitable algorithm and then implementing it (see Fig. 4.1, top), our intent was to replace all these steps performed by a user/researcher by a specification of constraints from a library of constraints (see Fig. 4.1, bottom). The rest of the job should be done by the generic algorithm we developed. Note that we do not claim generalizing any possible solution for audio source separation, but many of those that follow the LGM framework (4.3).

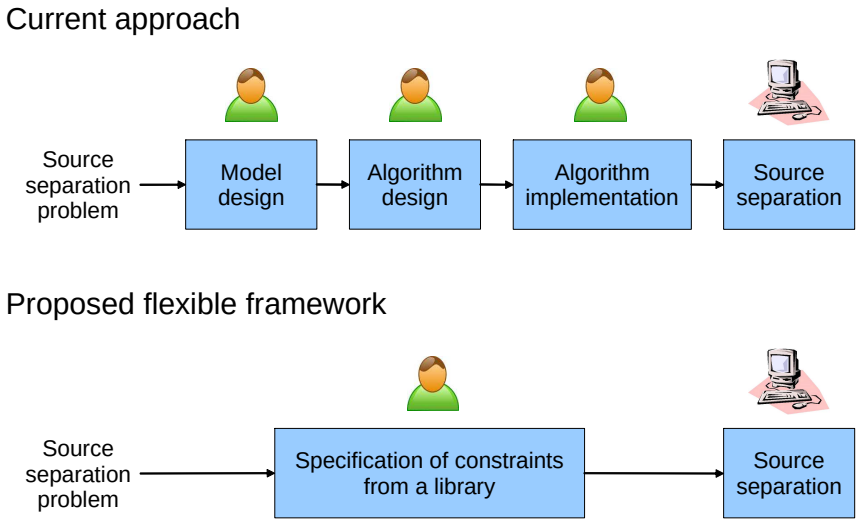


FIGURE 4.1 – Current way of addressing a new source separation problem (top) and the way of addressing it using the proposed flexible framework (bottom) (figure from [2]).

4.3 Formulation

In few words, our general framework is implemented as follows. The characteristics listed in the previous section can be specified for a source separation problem at whole, and individual characteristics may be specified for each source. The latter mostly include spatial covariance models and spectral variance models. In particular, the most of different possibilities are for spectral variance models, since they include the models mentioned in the previous section (NMF, harmonic NMF, GMM, etc ...), but also their combinations in an hierarchical fashion. One example of such an hierarchical decomposition, as applied to a recording of several xylophone notes, is represented on Figure 4.2. Finally, each parameter subset may be

— either left free,

— or fixed (e.g., if pre-trained or specified based on some prior knowledge),

— or given some probabilistic prior

during the estimation process.

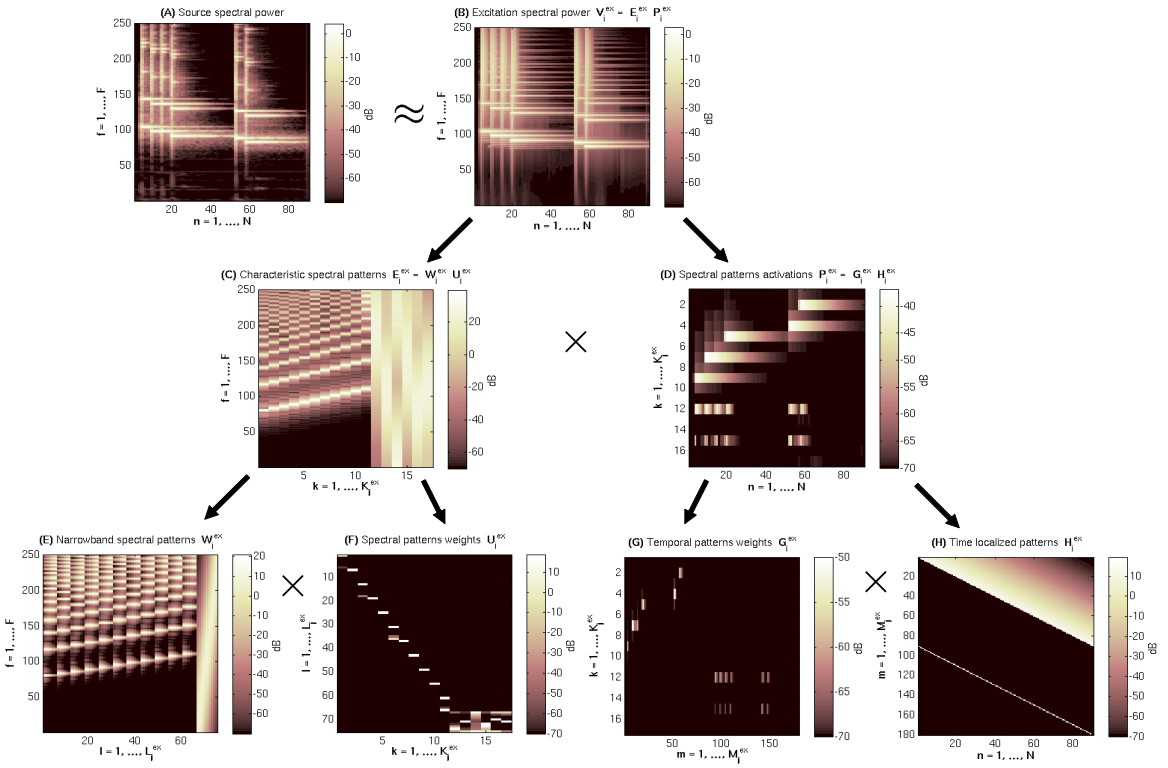


FIGURE 4.2 – A particular hierarchical spectral decomposition as applied to the spectral power of several xylophone notes (figure from [2]).

4.4 Algorithm

The generic algorithm we designed is also a variant of MU-GEM approach described in Section 2.3.2, though implemented for MAP estimation, since some parameters may be given probabilistic priors. It is schematized on Figure 4.3, and during the M-step it applies a specific (pre-defined) constraint to each parameter subset $\theta_{j,k}$. This allows achieving desired generality of the optimization strategy.

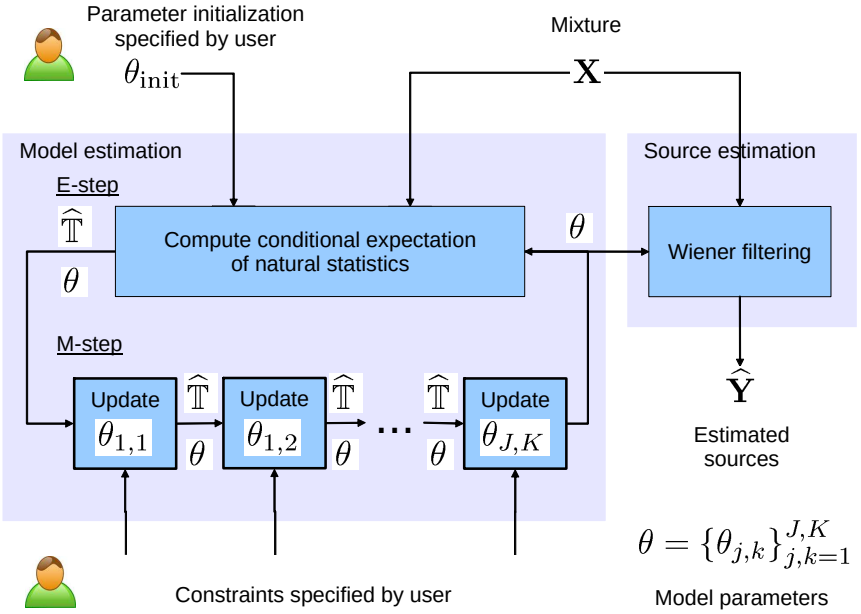


FIGURE 4.3 – Overview of the proposed general MU-GEM algorithm for parameter estimation and source separation (figure from [2]).

4.5 Implementation

The framework was implemented in Matlab, baptized *Flexible Audio Source Separation Toolbox (FASST)* and released for public use [14] under a general public license (GPL). It was later re-implemented in Python by Jean-Louis Durrieu. Finally, a C++ implementation [SVB+14] was developed and released by members of PANAMA team from INRIA - Rennes, where this work was done.

4.6 Conclusion

We have presented a general flexible framework for audio source separation [2]. Based on prior knowledge on a particular audio source separation problem, the framework allows specifying various constraints that are then taken into account for model estimation within the corresponding generic algorithm. This work was well remarked and the corresponding software implementation (FASST toolbox [14]) we released is used by researchers and engineers.

Chapitre 5

Paper 3 : Coding-based informed source separation : Nonnegative tensor factorization approach

This work was done in collaboration with Antoine Liutkus, Roland Badeau and Gaël Richard. Note that, though a great research effort in audio source separation was done, in general none of approaches allows attending any desired quality of the estimated sources. Indeed, there are even works reporting theoretical quality bounds (obtained by so-called *oracle estimators*) [VGP07] that cannot be overcome by wide classes of conventional source separation approaches. As such, researchers considered a different new setting lying sort of in between audio source separation and audio compression. It is assumed that the clean sources and the mixture are available at a so-called *encoding* stage, where any kind of information may be extracted in order to guide source separation at a so-called *decoding* stage, where clean sources are not available any more. The extracted information should be compact enough to be efficiently stored or transmitted. Interestingly, this problem has appeared in more or less the same time in both audio coding research community, where it was called *spatial audio object coding (SAOC)* [ERF+08], and in audio source separation research community, where it was called *informed source separation (ISS)* [PGB10, LPB+12].

Note that at the time we started working on this topic I was well-placed to work on such a problem related to both audio source separation and audio compression. Indeed, as we have already seen, I did a lot of work on audio source separation. Moreover, in 2007 I have done a one year postdoctoral stay in Sweden in Royal Institute of Technology (KTH), where I was working with Prof. Bastiaan Kleijn on audio compression [16, 17, 18]. More specifically, I was designing practical audio compression schemes based on probabilistic model-based quantization and encoding under high-rate theory assumptions [ZSN08]. As we will see, such kind of probabilistic compression can be elegantly married with probabilistic model-based (e.g., multichannel NTF) source separation.

5.1 Motivation

As it was already mentioned, the ISS problem has started to be studied independently and more or less at the same time by researchers from audio sources separation community, where it was called ISS [PGB10, LPB+12], and by researchers from audio compression community, where it was called SAOC [ERF+08]. However, no links were

established between those studies, and I am even not sure that one community was really aware about the research carried by another community. Moreover :

— As for ISS, the methods proposed by researchers from audio sources separation community [PGB10, LPB+12] were too source separation-inspired. For example, Liutkus *et al* [LPB+12] proposed to estimate NTF model from clean sources at the encoder, quantize and transmit its parameters, and then to use this quantized model at the decoder to estimate sources by Wiener filtering. A major drawback of this approach is that, even assuming quite high transmission rate, it cannot achieve any desired quality, since the achievable quality is bounded by that of oracle estimators [VGP07]. There was even a "believe" that ISS methods are in principle unable overcoming quality bounds of oracle estimators [PGB10, LPB+12]. It is however quite obvious from compression perspective that, in case when the rate is high enough, one can simply compresses the sources so as to achieve any desired quality.

— As for SAOC, most of proposed methods [ERF+08] were relying on estimating and transmitting cues such as channel correlation, spatial coherence, source localization parameter, etc ... This is however a very poor modeling, as compared to NTF for example. In SAOC [ERF+08] it is also proposed to transmit perceptually-encoded residual signals to achieve high quality result.

Our motivation for this work was to mix-up the two worlds (source separation and compression), while keeping the best of each of them.

5.2 Coding-based ISS at glance

We have introduced a so called *Coding-based ISS (CISS)*, first in a preliminary conference publication [19], as a concept, and then in the journal publication [3], where the multisource NTF model was used. To introduce it briefly, the CISS consists in :

— *CISS encoder* :
  — Estimate a Gaussian model (whatever Gaussian model) from clean sources.
  — Quantize, encode and transmit the model parameters.
  — Quantize, encode and transmit the sources based on the posterior distribution of the sources, given the mixture and quantized model.

— *CISS decoder* :
  — Decode quantized model parameters.
  — Decode the sources based on the posterior distribution of the sources, given the mixture and quantized model.

Figure 5.1 provides a very high-level illustration to understand global advantages of CISS over conventional ISS and source coding. Let us comment through the subplots :

— (A) : In conventional ISS methods (e.g., [LPB+12]) the sources themselves are not transmitted, and thus the performance is bounded by that of oracle estimators [VGP07].

— (B) : If we apply conventional source coding to the sources without using the mixture, we may achieve any quality, given a sufficient bit-rate. However, the rate that is inversely proportional to the data log-likelihood would be quite high. Indeed, the distribution is quite broad (see blue Gaussian on on Fig. 5.1 (B)).

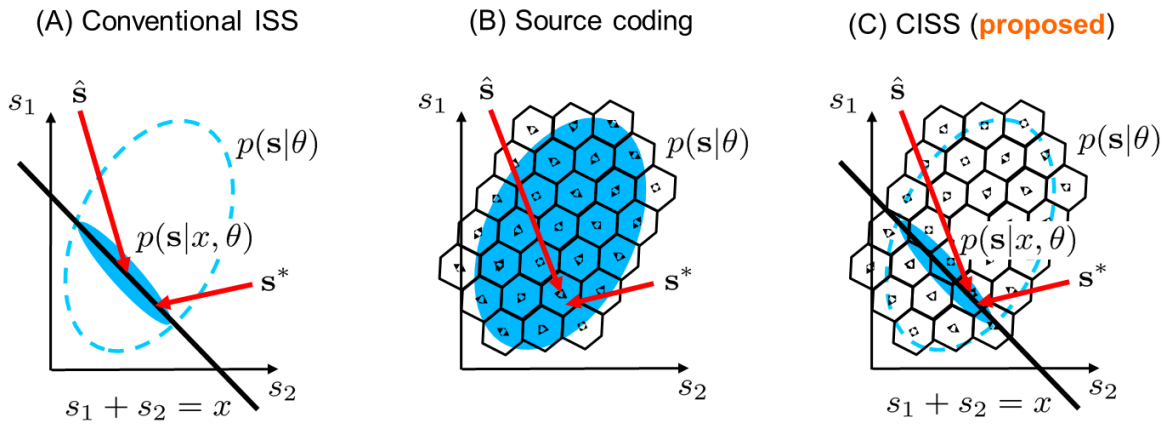


FIGURE 5.1 – Simplified visualization of the following probabilistic model-based methods applied in one TF point : (A) conventional ISS [LPB+12], (B) source coding and (C) the proposed coding-based ISS (CISS). Notations : x : mixture, $\mathbf{s} = [s_1, s_2]^T$: sources, $p(\mathbf{s}|\theta)$: *a priori* source distribution, $p(\mathbf{s}|x,\theta)$: *a posteriori* source distribution, $\mathbf{s}^*$: true sources, $\hat{\mathbf{s}}$: estimated sources.

— (C) : Within CISS

  — on one hand, as in source coding, we quantize and encode the sources, thus we may achieve any quality,

  — on the other hand, as in source separation or in conventional ISS, we quantize and encode using the posterior source distribution, given the mixture. Since the posterior distribution is much narrower than the prior distribution (see Fig. 5.1 (C)), we need much smaller rate, as compared to source coding without using the mixture.

5.3 CISS based on multisource NTF

We now present the CISS scheme based on multisource NTF (NTF-CISS) as published in [3]. Note that for simplicity we considered in [3] the single channel ISS scenario. However, extension to multichannel scenario is straightforward, and we have done it later and published in a conference paper [20].

We assume single-channel mixing (1.15) directly in the STFT domain. Sources, represented as before by tensor $\boldsymbol{\mathcal{S}}$ in the STFT domain, are assumed to follow multisource NTF model (2.1), (2.2). This formulation falls again within our general modeling represented on Figure 2.1.

Note that the posterior distribution-based CISS, as sketched in the previous section, may be equivalently seen as : first computing a rough estimation of sources $\widehat{\boldsymbol{\mathcal{S}}}$ by standard Wiener filtering and then encoding the residual $\boldsymbol{\mathcal{S}} - \widehat{\boldsymbol{\mathcal{S}}}$ based on the posterior distribution covariances. We see again a relation with SAOC [ERF+08], where residual signals may be optionally encoded.

We now present in broad lines the NTF-CISS [3] encoding/decoding :

— *NTF-CISS encoder* :

  — Estimate NTF $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$ from clean sources $\boldsymbol{\mathcal{S}}$ by applying MU rules (1.23), (1.24), (1.25) with $\boldsymbol{\mathcal{B}} = |\boldsymbol{\mathcal{S}}|^{\cdot 2}$.

  - Quantize model parameters (leading to $\bar{\boldsymbol{\theta}}$), entropy-encode them (in [3] we used uniform scalar quantization of parameters in log-domain (i.e., $\log(\mathbf{Q})$, $\log(\mathbf{W})$, $\log(\mathbf{H})$) and GMM-based arithmetic coding [ZSN08]) and transmit.
  - Compute rough source estimates $\widehat{\boldsymbol{\mathcal{S}}} = \mathbb{E}[\boldsymbol{\mathcal{S}}|\mathbf{X}; \bar{\boldsymbol{\theta}}]$ by standard Wiener filtering (2.8).
  - Residuals $\boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{S}} - \widehat{\boldsymbol{\mathcal{S}}}$ being a posteriori zero-mean,
    - decorrelate them with Karhunen-Loeve Transform (KLT) based on posterior covariances (2.9),
    - apply uniform scalar quantization (leading to $\bar{\boldsymbol{\mathcal{R}}}$) in KLT domain [ZSN08] and entropy-encode based on statistics of posterior covariances (2.9).
- *NTF-CISS dencoder* :
  - Decode NTF model parameters $\bar{\boldsymbol{\theta}}$.
  - Compute posterior means $\widehat{\boldsymbol{\mathcal{S}}}$ as in (2.8) and posterior covariances as in (2.9).
  - Decode residuals $\bar{\boldsymbol{\mathcal{R}}}$ in the KLT domain and transform them back to the STFT domain by applying the inverse KLT computed from posterior covariances.
  - Reconstruct sources as $\bar{\boldsymbol{\mathcal{S}}} = \widehat{\boldsymbol{\mathcal{S}}} + \bar{\boldsymbol{\mathcal{R}}}$.

Note that even though the above NTF-CISS scheme overview rely on general equations (2.8) and (2.9) for posterior mean and covariance computation, since everything is formulated in the STFT domain, these expressions can be factorized over both time and frequency. As such, similarly to [1] (see also Chap. 3) and [2] (see also Chap. 4), there is no computational burden related to (2.8) and (2.9).

Let us mention few other attractive features of the proposed NTF-CISS approach :

- As it was already mentioned, it is easily extendable to multichannel case within, e.g., general formulation presented in Chapter 2. We have indeed proposed and published such extension in [20].
- In addition, perceptual modeling is possible within NTF-CISS framework. We have later published a study on perceptual modeling in [21].
- In contrast to SAOC [ERF+08], where two different models are used to encode the rough source estimates and the residuals (thus both models need to be transmitted), in our case NTF is used for both rough source estimates (posterior means) and for arithmetic coding of residuals (posterior covariances), and NTF parameters are the only parameters transmitted.
- In contrast to some other NMF/NTF-based audio compression methods [NV10, NVV11] (though not addressing directly the ISS problem), where the STFT magnitudes and phases are encoded independently, in NTF-CISS framework they are encoded jointly (by directly coding STFT coefficients), which is possible thanks to the probabilistic Gaussian NTF formulation.

Another important findings we did in [3] are as follows :

- Based on a previous study [17], we concluded that at least in the high-rate regime the optimal rate for model transmission is constant and independent on the total rate. In other words, in the high-rate regime any extra rate should be spent for signal (here sources) transmission.
- We have shown theoretically, under some approximations, that NTF parameters should be quantized in log-domain.
- We have also derived optimal rate distribution between different NTF parameter sets (i.e., matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$).

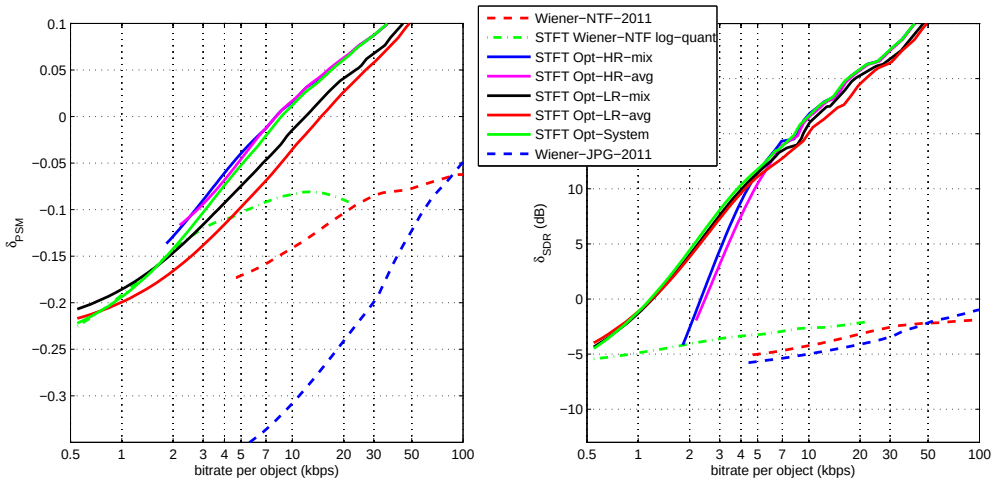


FIGURE 5.2 – CISS-NTF with different ways of optimizing parameters (solid lines), compared to state of the art [LPB+12] (dotted lines). δPSM and δSDR denote the improvements over the corresponding measures computed for the oracle Wiener filtering [VGP07] source estimates in the STFT domain (figure from [3]).

5.4 Results

Various ISS approaches were evaluated and compared in terms of rate vs. source separation performance curves. As for source separation performance metrics, we used both

— perceptual similarity measure (PSM) of PEMO-Q [HK06], and

— source to distortion ratio (SDR) [VGF06].

Instead of plotting absolute values of these metrics, we used δPSM and δSDR that are improvements of the corresponding measures computed for the oracle Wiener filtering [VGP07] results.

The results are plotted on Figure 5.2. One can see that all evaluated state-of-the-art ISS methods cannot go beyond zero, since by construction they cannot outperform oracle Wiener filtering. On the contrary, thanks to the source residuals encoding, NTF-CISS outperforms oracle Wiener filtering for high rates, and outperforms classical ISS methods for all rates.

5.5 Conclusion

In this chapter we have first described the ISS problem. We have then introduced the CISS and NTF-CISS frameworks that rely on both source separation and compression principals to solve ISS problem efficiently. We have shown experimental results demonstrating considerable superiority of NTF-CISS over prior art ISS methods.

Chapitre 6

Paper 4 : Solving time-domain audio inverse problems using nonnegative tensor factorization

This work was done in collaboration with Çağdaş Bilen and Patrick Pérez in Technicolor, within a collaborative ANR JCJC project MAD (Missing Audio Data). At the beginning of this project we were wandering whether NMF/NTF modeling, being a representation of nonnegative audio (power-)spectrograms, is suitable for reconstructing missing audio sample in time domain, e.g., problems like audio declipping [AEJ+12, KJM+13, KBG15, SKD14] or more generally audio inpainting [AEJ+12]. In fact, it became possible with IS-NMF/NTF thanks to its Gaussian interpretation ; and the fact that subsampling in time (loosing samples in time domain) and STFT (inverse STFT) are linear transforms. Moreover, this setting and the modeling we proposed fit into the general scheme described in Chapter 2 (see also Fig. 2.1).

Historically this work was conducted as follows. We first were thinking how to apply NMF just for audio declipping, but then we have realized that the approach we had in mind may be generalized making possible several other potential applications, existing or new. As such, to promote the use of NMF/NTF modeling for those applications, we have first decided publishing a series of conference papers [22, 23, 24, 25, 26], each paper focusing on just one application. We have then formulated the framework in a general way and published it in the journal paper [4] together with the applications including

— audio declipping [22, 26, 4],

— joint audio inpainting and source separation [23, 4],

— compressive sampling recovery [4],

— compressive sampling-based ISS (CS-ISS) [24, 4].

6.1 General framework formulation

We here give a high level presentation of the proposed framework without going too deep into technical details (see [4] for more details). It is assumed that J sources are mixed forming a single-channel mixture (1.14). Moreover, it is assumed that all the signals, including sources and mixtures, are possibly quantized and subsampled. The observations are the remaining quantized samples of the sources and the mixture, and the goal is to reconstruct the original (non-quantized) sources and the mixture

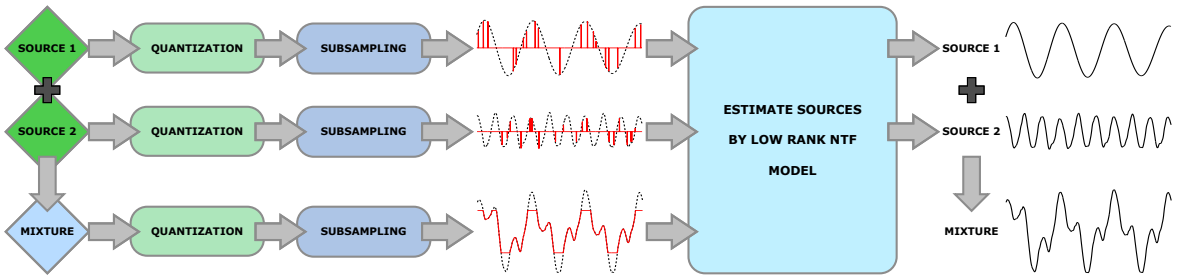


FIGURE 6.1 – General formulation of time-domain inverse problems.

(the latter can be obtained by simply summing up the reconstructed sources). The above problem formulation is schematized on Figure 6.1, and one can note that this formulation is a partial case of our more general formulation presented in Chapter 2 (see also Fig. 2.1). Indeed, our formulation here is limited to the case of single-channel mixtures for the sake of simplicity, though extending it to multichannel case is quite straightforward. In fact, we have done such an extension for a punctual study of multichannel audio declipping [26]. Moreover, though the mixture is single-channel, our overall formulation might be considered as multichannel, since both the sources and the mixture are partially observed, and the source observations may be considered as additional channels.

6.2 Modeling and algorithms

The modeling we used is exactly the multichannel NTF modeling as described in Chapter 2. As for the algorithms for model parameters estimation and signals reconstruction, the GEM-MU algorithm described in Section 2.3.2 and Wiener filtering (2.8) are applied with some approximations detailed just below.

As mentioned in Section 2.3.2, the GEM-MU algorithm in its general formulation requires multiplication and inversion of matrices of very high dimensions. Recall that in Chapters 3, 4 and 5 (papers [1], [2] and [3]) this computational burden was avoided thanks to the reformulation of the corresponding problems in the STFT domain and thanks to the narrowband approximation (3.3), which allowed factorizing all those matrix operations both in time and in frequency. Unfortunately, in contrast to [1, 2, 3], in this study we have not managed factorizing in both time and frequency, but only in time. Indeed, missing observations in time domain introduces very strong and important posterior dependencies in frequency domain. The factorization in time was achieved using the following relaxation.

We distinguish between the following three domains :

— the *time domain*,

— the *framed time domain*, which is the time domain signal chunked into overlapping frames and windowed (just up to DFT computation to obtain the STFT), and

— the *STFT domain* itself.

The framed time domain and the STFT domain are both redundant and related by a unitary transform, which is the DFT. We relax the problem by assuming that the samples are missing in the framed time domain and that the frames are independent. This is indeed an approximation, since we drop all the dependencies between overlapping frames in time. This relaxation allows drastically reducing the computational

load to the order of F (F being the number of frequency bins). Also, since the observations are in framed time domain and the NTF model is in the STFT domain, we need constantly switching within the algorithm between those two domains, and, for better efficiency, this is implemented using the fast Fourier transform (FFT). This allowed us implementing and testing the proposed framework for various applications. However, it should be noted that the computational load remains quite high, since F may be of order of 1000. As such, alternative solutions allowing further reducing the computational load are welcome.

6.3 Applications and results

In this section we present how the above-described general framework may be applied to different existing and new problems. We also show some experimental evaluation results and comparisons with the relevant state of the art methods. For the sake of conciseness we do not give much details on the experimental data and setup (an interested reader may find this information in the paper [4]).

6.3.1 Audio declipping

There are many audio processing problems, where the goal is to estimate audio samples that are for some reason missing in time or time-frequency domain. Those problems include audio declipping and declicking, compressive sampling recovery, packet loss concealment, bandwidth extension, etc ... Recently, inspired by image inpainting [BSCB00] where the goal is to reconstruct missing parts in images, all these problems were baptized as *audio inpainting* problems by Alder *et al.* [AEJ+12]. However, in my opinion the main contribution of Alder *et al.* in this work [AEJ+12] consists in changing the insight on these problems by proposing considering them as inverse problems, e.g., like source separation. We were working on those problems within the MAD (Missing Audio Data) ANR project [1] coordinated by Valentin Emiya, one of the co-authors of [AEJ+12].

The goal of audio declipping (a particular audio inpainting problem) consists in recovering time samples missed due to clipping (saturation). Since the publication of audio inpainting paper [AEJ+12], several machine learning-based methods were proposed and investigated to solve audio declipping problem. Those include sparsity-based methods [AEJ+12, KJM+13], cosparsity-based methods [KBG14, KBG15] and structural (so-called *social*) sparsity-based methods [SKD14].

At the time we started working on this project, to our best knowledge, there were no work trying to apply NMF or NTF for audio declipping. Indeed, it seems tricky at first glance, since the missing data and the modeling are in two different domains : time and STFT, respectively. Very likely this is a reason why most of the existing NMF-based audio inpainting methods were designed for reconstructing missing data in the STFT domain [LRKO+11, SRS11, ŞYC12].

Within our general formulation the audio declipping problem is addressed by assuming that there is just one non-observed source ($J = 1$) and the clipped mixture samples are missing (in fact, the source and the mixture are the same signal here). It is easy to understand that in this case the NTF modeling degenerates to the NMF modeling, since one of the dimensions of the latent source tensor $\boldsymbol{\mathcal{S}}$ is one ($J = 1$).

1. http://mad.lif.univ-mrs.fr/

Note that in case of clipping the clipped samples are indeed missing, but some additional information about them is available. Notably, it is known that the original value of a clipped sample is above (below) the clipping threshold. Taking into account those so-called *clipping constraints* is very important for better signal estimation. However, the clipping constraints are linear inequalities that make the posterior distribution of the unknown samples no longer Gaussian. This makes it difficult to manage those constraints within the proposed framework. However, we have proposed few rather ad hoc tricks allowing managing them (see [22, 4] for details).

We compared the proposed approach with the following state of the art methods :

— orthogonal matching pursuit (OMP) [AEJ+12],

— iterative hard-thresholding (HT) [KJM+13],

— cosparsity (Cosp) [KBG15],

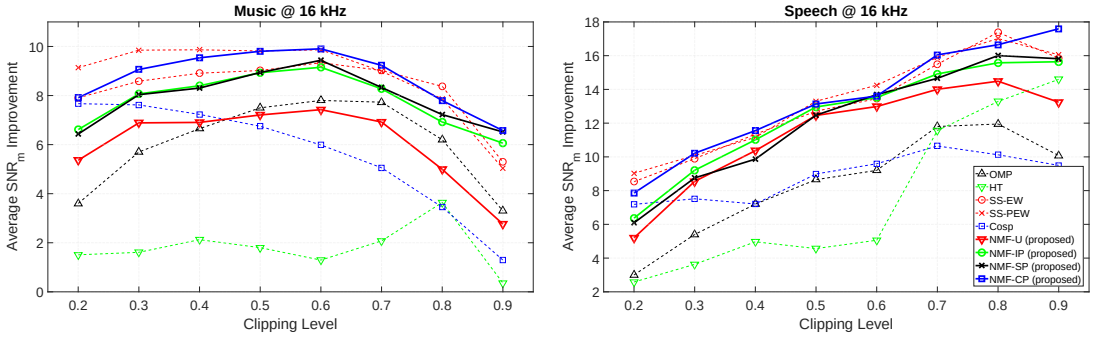
— social sparsity with empirical Wiener operator (SS-EW) [SKD14], and

— social sparsity with posterior empirical Wiener operator (SS-PEW) [SKD14].

The comparison was done on 10 music and 10 speech signals clipped at 8 different levels. The performance was measured in terms of signal to distortion ratio ($\mathrm{SNR}_{\mathrm{m}}$) improvement over the clipped signal, where the "m" subscript means that the SNR is computed only over the time support of clipped samples (see [4] for details). The results are plotted on Figure 6.2, where NMF-U denotes the proposed method without clipping constraints, and NMF-IP, NMF-SP and NMF-CP the proposed method with various strategies of managing clipping constraints. One can see that the proposed NMF-CP method gives results that are comparable to social sparsity-based method SS-PEW [SKD14]. These results show that using structural signal models (here social sparsity and NMF) leads to better declipping performance than the models based on local sparsity only [AEJ+12, KJM+13, KBG15]. One can also note that managing clipping constraint is indeed very important, especially for music signals (see NMF-U vs. NMF-CP on Fig. 6.2).

It shall be noted that to achieve these results it was very important to chose a suitable NMF model order K depending on the signal type. We have empirically found that $K = 20$ and $K = 28$ are the most suitable values for music and speech signals, respectively. Choosing smaller or higher K was leading to the performance drop. As for performance drop with smaller K, it is quite easy to understand. Indeed, with smaller K the model is not able approximating anymore the signal spectrogram with enough details. For example, NMF with $K = 2$ would not be precise enough to well approximate the spectrogram on Figure 1.2. As for performance drop with higher K, it is a bit more difficult to understand why over-estimating model order would lead to the performance degradation. Let us explain it on one example. Assume we have a periodic signal that is clipped, then the missing data support will be periodic as well. Overestimating model order would lead to appearance of sort of "phantom" (non-existing) periodic (so with low-rank spectrograms) signals that would freely oscillate in the missing support, thus leading to performance degradation. We have indeed observed such a behavior experimentally, while declipping a snar drum waveform. We will come back to this point in Section 6.3.3 below, where (surprisingly) we will see an opposite behaviour of over-estimated model, when the missing support is random.

We have also extended the proposed declipping algorithm to the case of multichannel mixtures and published as a separate conference publication [26]. The proposed approach is not anymore an instance of the general framework from [4], since in [4] we have formulated it for single-channel mixtures only, but it is an instance of the general



(a) Average SNR_m improvement computed over 10 music signals.

(b) Average SNR_m improvement computed over 10 speech signals.

FIGURE 6.2 – The average performance of all the audio declipping algorithms as a function of the clipping threshold. Lower threshold corresponds to more severe clipping (figure from [4]).

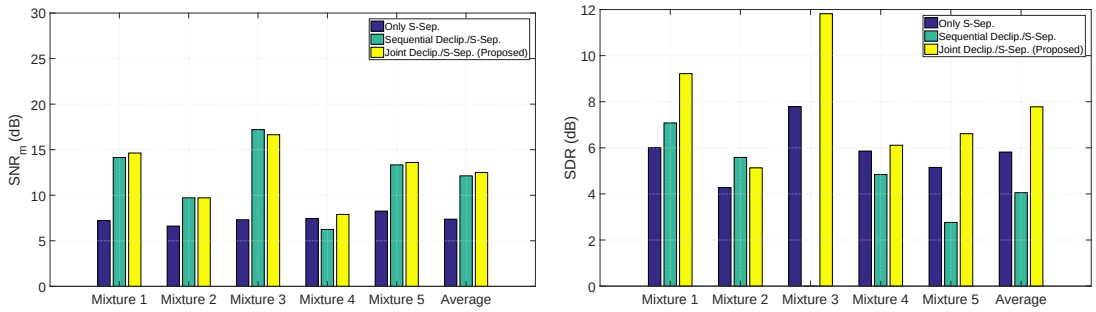
formulation presented in Chapter 2 (see also Fig. 2.1). Moreover, in contrast to single-channel NMF-based declipping algorithm [22, 4] we have discussed just above, where the multi-source nature is not important (recall that $J = 1$ in [22, 4]), in the multi-channel case it is very important to model latent sources with multi-source ($J > 1$) NTF model. Indeed, this allows the resulting approach exploiting the fact that different audio sources contribute differently in different channels, and thus remain highly correlated. To our best knowledge there were no previous works exploiting such correlations to declip multichannel signals. A naive approach would be obviously to declip each channel independently using any single-channel declipping algorithm. We have proven experimentally that the proposed approach outperforms the naive approach relying on single-channel NMF-based algorithm [22] (see [26] for details).

6.3.2 Joint audio inpainting and source separation

Real world audio mixtures are often degraded, for example they may be clipped, as we have seen in the previous section. However, in most of research on audio source separation it is assumed that the mixture is not degraded, which might be limiting for some real-world scenarios. A dummy solution in this case is a so-called *sequential approach* consisting in first declipping the mixture by any declipping algorithm and then applying a source separation algorithm to the decipped mixture. However, such an approach is sort of suboptimal, since it may suffer from errors propagation : estimation errors produced at the declipping stage cannot be corrected at the source separation stage. As such, we have proposed addressing this problem in a systematic and joint manner, where the sources are estimated directly from the mixture with missing samples without any prior missing samples imputation (declipping). We have introduced the concept of "joint audio inpainting (here declipping) and source separation" [23, 4]. To our best knowledge this was the first time that the problem of audio source separation from clipped mixtures was addressed "properly", i.e., in a systematic fashion, as opposed to the above-described sequential approach.

Within our general formulation this problem is addressed by assuming that there are $J > 1$ non-observed sources and the clipped mixture samples are missing. In other words, it is exactly as on Figure 6.1, except that none of the source samples are observed.

We have tested the proposed approach on 5 mixtures of 3 music sources and compared it to the source separation only (i.e., without any declipping) and to the sequential



(a) Declipping performance for clipping level 0.2.

(b) Source separation performance for clipping level 0.2.

FIGURE 6.3 – The declipping and source separation performance of joint optimization compared to sequential (figure from [4]).

(dummy) approach we have mentioned above. In order to have descent source separation performance, we have injected some segmental information into all approaches under comparison (for more details see [4] and Fig. 3 therein). The approaches were accessed both in terms of declipping performance (SNR$_m$) and source separation performance (SDR). The results are shown on Figure 6.3 for a quite severe clipping (mixtures rescaled between -1 and 1 are clipped in between -0.2 and 0.2) (see Fig. 4 in [4] for more results). One can remark that while joint approach, as compared to sequantial approach, does not bring any improvement in terms of declipping (see Fig. 6.3a), it outperforms in average both separation only and sequantial approaches in terms of source separation quality (see Fig. 6.3b).

6.3.3 Compressive sampling recovery

Compressive sampling [CW08] consists in randomly sampling a signal in some domain for the sake of compression. Compressive sampling recovery is an inverse problem consisting in reconstruction of the original signal from those random samples. For the sake of more efficient compression it is better to sample in a domain that is *incoherent* to another domain, where the signal has some structure, e.g., sparsity or low-rankness. As such, we are here considering random sampling in time domain which is incoherent to the STFT domain, where audio signals are sparse and their spectrograms are usually of low rank.

The implementation of compressive sampling recovery within our general framework (Fig. 6.1) is exactly the same as that of audio declipping (Sec. 6.3.1), except that there is no clipping constraint to be taken into account, i.e., a missing sample may be of any value.

The results of recovering a 4s long music signal from different percentages of random samples kept and for varying model order K are shown on Figure 6.4 in terms of SNR$_m$, together with results of a shape preserving piecewise cubic interpolation. [2] One can note that the proposed method greatly outperforms the linear interpolation.

Another interesting observation is that the results do not degrade with increasing model order, which is totally opposite to what we have observed for audio declipping problem (Sec. 6.3.1). In other words, a random missing support prevents the model from overfitting the data even if its order K is high. This should be related somehow to the

2. For the interpolation, *theinterp1()* function of Matlab 2016a is used with *phcip* method, which gave the best results among the available interpolation methods.

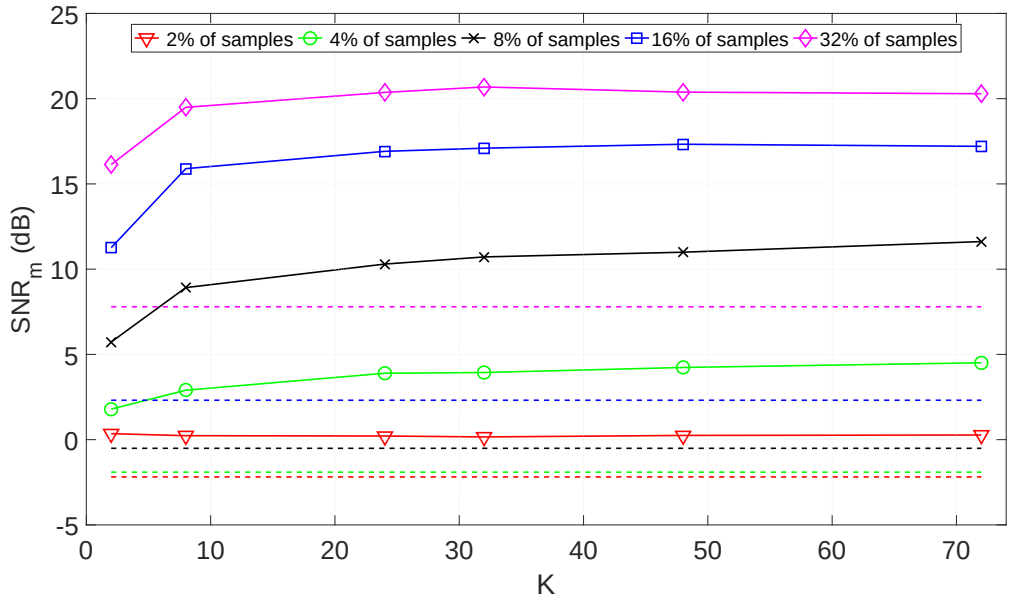


FIGURE 6.4 – The reconstruction performance measured in terms of SNR_m of a 4s long music signal from its random samples. The reconstruction results with our proposed algorithm (solid lines) are shown for different percentage of samples and different number of components, K, used in our approach. The results with shape preserving piecewise cubic interpolation are also shown for comparison (dashed lines), with the colors indicating corresponding percentage of samples (figure from [4]).

missing data theory [Gra09], where different natures of data missingness are considered : data missing completely at random (MCAR) and data missing not at random (MNAR). In compressive sampling recovery the data are MCAR, since the missed samples support is random and independent on the signal itself. In audio declipping the data are MNAR, since the missed samples support depends on the values of the signal itself.

6.3.4 Compressive sampling-based ISS

Here we go back to the ISS problem (see Chapter 5). Note that all prior ISS schemes, including [LPB+12] and [3], have encoders that are computationally demanding. Moreover, they have a higher computational load at the encoding stage than at the decoding stage. For example, in [LPB+12] and [3] at the encoder the STFT is computed and the NTF model is estimated with an iterative algorithm, while at the decoder the inverse STFT is computed, while there is no need in NTF model estimation (its parameters are transmitted). However, for some applications there might be a need of a very fast encoder, possibly at the expense of a more computationally demanding decoder. For example, for archiving purposes one needs compressing and storing everything, while de-compression may be necessary only from time to time, on demand. In this case having a very fast encoder would lead to overall time and energy consumption savings.

As such, our motivation in this work was to build an ISS scheme with a very fast encoder, while moving the computational load from the encoder to the decoder. Our idea is based on both the concepts of

— the compressive sampling [CW08] (see also Sec. 6.3.3), and

— the distributed source/video coding [XLC04, GARRM05], where the encoder (several encoders) is (are) very simple, and most of redundency of the encoded data is exploited at the decoder.

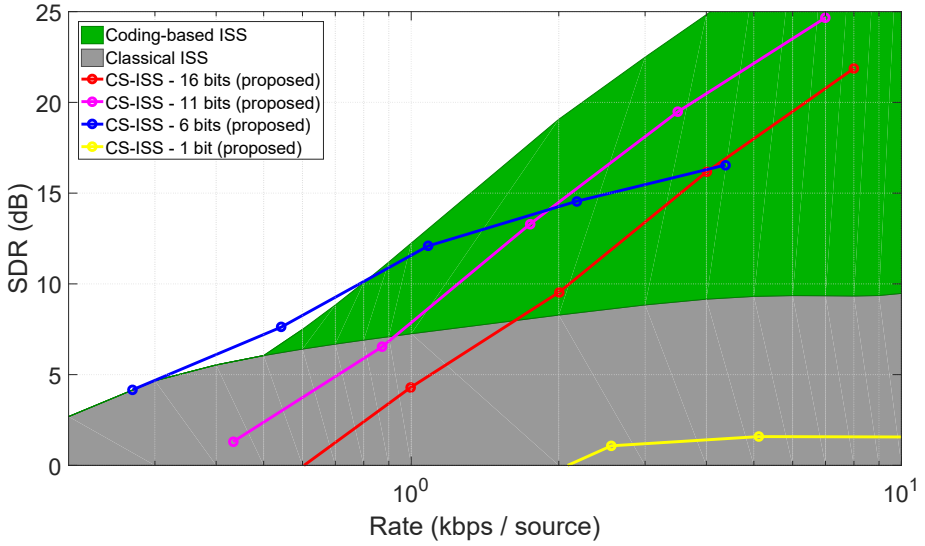


FIGURE 6.5 – The rate-distortion performance of CS-ISS using different quantization levels of the encoded samples. The performance of the ISS algorithm from [LPB+12] and the coding-based ISS algorithm from [3] are also shown for comparison (figure from [4]).

More precisely we propose the following ISS scheme called *compressive sampling-based ISS (CS-ISS)*. At the encoder the sources are simply randomly samples, the samples are uniformly quantized, (optionally) entropy-encoded and transmitted to the decoder. It is indeed quite difficult to imagine an encoder with yet lighter computations, there is even no need to compute any transform. Note that only the quantized source sample values need to be transmitted to the decoder, since exactly the same random samples support can be regenerated at the decoder by the same random generator initialized with the same seed. At the decoder the sources are reconstructed from the transmitted source samples and the mixture (recall, in ISS the mixture is assumed to be known at both the encoder and the decoder, see Chapter 5).

The last (decoding) step may be implemented within our general formulation as follows. It is assumed that there are $J > 1$ sources and their quantized samples are partly observed (on the random sampling support). It is also assumed that the mixture is fully observed. In other words, the setup is exactly as on Figure 6.1, except that the mixture is not clipped.

We have compared the proposed CS-ISS with a classical ISS approach [LPB+12] and with the CISS [3] presented in Chapter 5. The results in terms of SDR vs. rate curves are shown on Figure 6.5. Plain curves correspond to CS-ISS results with different levels of quantization (bits per source sample). We see that the optimal quantization level of CS-ISS varies depending on the overall rate : for low rates 6 bits leads to better performance, while for high rates it is 11 bits. Overall, we see that, as compared to the state of the art ISS schemes, CS-ISS gives the same performance for low rates and slightly worth performance for high rates. This loss in performance is a price to pay for an encoder that is extremely fast.

6.4 Conclusion

We have described a general framework based on NTF modeling in the STFT domain for solving audio inverse problems with data missing in time domain. We

have applied this framework for a wide range of new and existing applications. NTF modeling, being a powerful model of audio spectrograms, has demonstrated for most of applications the performances that are on par with or superior to the state-of-the-art.

Chapitre 7

Other work

Here I briefly mention some other projects/topics I was working on since I have defended my PhD in 2006. Those projects/topics are listed in separate sections below without any particular structure, though more or less in a chronological order.

Also, I would like to highlight that while working in Technicolor (about 8 years) I learned a lot on image processing and computer vision, and was involved in some projects in these areas. Though I have not published a lot in these domains [27, 5], this was a great and very interesting experience for me. Moreover, this allowed me proposing and developing some new approaches in audio processing [28, 6] inspired by paradigms proposed in image processing and computer vision. This will be mentioned below.

7.1 Flexible speech and audio coding

In 2007 I have done a one year postdoctoral stay in Sweden in Royal Institute of Technology (KTH). I was working with Prof. Bastiaan Kleijn on a European union funded project FlexCode. [1] We were designing and developing new flexible speech and audio compression schemes ; flexible in the sense that they can be instantly recast to operate on any available/desired bit rate from a continuum of bit rates. This flexibility was achieved thanks to probabilistic model-based quantization and encoding under high-rate theory assumptions [ZSN08]. We have published several papers on the topic [16, 17, 18].

7.2 Learning from uncertain data

Another attractive feature of the probabilistic multichannel NTF modeling (Chap. 2) and LGM in general (Sec. 4.1) is that it allows not only estimating the sources via Wiener filtering (2.8), but also their posterior covariances (2.9). The latter may be used as a measure of goodness of the estimated values or uncertainty about those estimates. This uncertainty may be efficiently taken into account while learning from the estimated sources.

In collaboration with Mathieu Lagrange and Emmanuel Vincent we have developed such GMM learning schemes as applied to speaker recognition task [29, 30] and singer identification task [31]. In both cases some LGM-based source separation algorithm was used to enhance either speech or singing voice, and then the uncertainty was propagated

1. http://www.flexcode.rwth-aachen.de/

through feature computation (we used Mel-frequency cepstral coefficients (MFCCs)) up to learning GMMs.

In collaboration with Simon Arberet, Rémi Gribonval and Frédéric Bimbot we investigated another use case of exploiting such uncertainty. More specifically, we were first using one LGM-based source separation algorithm to separate sources, and then we were using another source separation algorithm (e.g., based on different models) with models learned from the uncertain output of the first algorithm [32, 33]. This approach allowed for a sort of "sequential" fusion of the two source separation algorithms.

7.3 Source localization

I was also working with Charles Blandin (an intern in INRIA) and Emmanuel Vincent on source localization, where the goal is to estimate the directions of arrival (DoAs) of different sources in stereo ($I = 2$) or multichannel ($I > 2$) mixtures. In particular, source localization may be very useful for sources separation, since estimated DoAs can be used to initialize source separation algorithm or inject some prior information in it. We have proposed some new source localization approaches and provided a thorough experimental evaluation of existing and proposed approaches [34, 35]. All the approaches were released for public use within a so-called *BSS Locate toolbox*[2] in case of stereo mixture. A multichannel version of the toolbox was developped and released later by researchers from INRIA. Our journal paper [35] was well remarked (187 citations according to Google Scholar on November 19, 2019).

7.4 Source separation evaluation

I was participating in the organization of the second community-based Signal Separation Evaluation Campaign (SiSEC 2010) [36, 37]. I have also co-authored a journal paper [38] resuming the results, findings and conclusions that can be drawn from several signal separation evaluation campaigns over 4 years.

7.5 Informed source separation

I was also working on other variants of informed source separation that are not really related to compression. More precisely, in those approaches the source separation process is informed by either some auxiliary information (e.g., music score [EPMP14] in case of music audio separation) or by some information provided by a user via a dedicated interface [BMW14]. Note that both the latter approaches and the compression-related approaches (presented in Chapter 5 and Section 6.3.4) are often referred to as "informed source separation", which might be misleading sometimes.

Inspired by score-informed music source separation [EPMP14], we have proposed text-informed speech separation [39, 9]. We have also extended the same kind of idea to propose a solution for text-informed speech inpainting [40] in the case when long portions of speech signal are lost. We have also proposed several user-guided source separation approaches [7, 10, 25] including one interactive method [10], where user may continuously refining the guidance in order to improve the separation result. We have also introduced a so-called *on-the-fly audio source separation* paradigm [41, 42, 28]

2. http://bass-db.gforge.inria.fr/bss_locate/

that greatly facilitates the user interaction, thus allowing source separation guidance by non-professional users. The principle is very simple : a user types in a dedicated interface some keywords describing the sources (e.g., "dog barking" and "wind"), and then the system retrieves from the internet corresponding source examples that are immediately used to guide source separation process (a demo video can be found at [3]). This was inspired by on-the-fly object category retrieval approach [CZ12] proposed in computer vision community. Finally, within the PhD of Sanjeel Parekh co-advised between Technicolor and Télécom ParisTech, we have proposed source separation frameworks guided by motion [43] and by video information [44].

7.6 Audio-visual scenes understanding

Another part of Sanjeel Parekh's PhD work consisted in mining audio-visual objects in large video collections with "weak" labels. The labels are weak in the sense that all videos are annotated with some global labels, but there is no spatial or temporal information about the location of the corresponding audio-visual objects, e.g., a video may be labeled as "train", but there is no annotation neither about train appearance in video stream nor about where one can hear a train in audio stream. Based on a specific deep neural network (DNN) architecture and a non-supervised NMF decomposition of audio, we have developed an approach allowing at the same time (*i*) identifying a video by a label, (*ii*) locating the corresponding object in the video stream, and (*iii*) separating the corresponding sound from the audio stream [45, 46].

7.7 A bit of image/video processing : Faces

I was also publishing a little on topics in computer vision related to facial analysis. Within one project we have proposed a new facial landmarks localization estimation algorithm [27]. Another project consisted in analysis/identification of people in complex videos like movies. First, we have created a so-called *Hannah* dataset. The dataset consists in complete annotations of "Hannah and Her Sisters" movie by Woody Allen in terms of people/characters in both video and audio streams. More precisely, faces in all frames were annotated with bounding boxes and labeled with character names/ids (see Fig. 7.1), and audio speech segments were annotated in time and labeled with the same character names/ids. The dataset was released for public use, [4] and it is described in [5], where we have also proposed a new approach for face tracking in videos that we have evaluated on the dataset.

3. http://youtu.be/mBmJW7cy710/
4. https://www.interdigital.com/data_sets/hannah-dataset

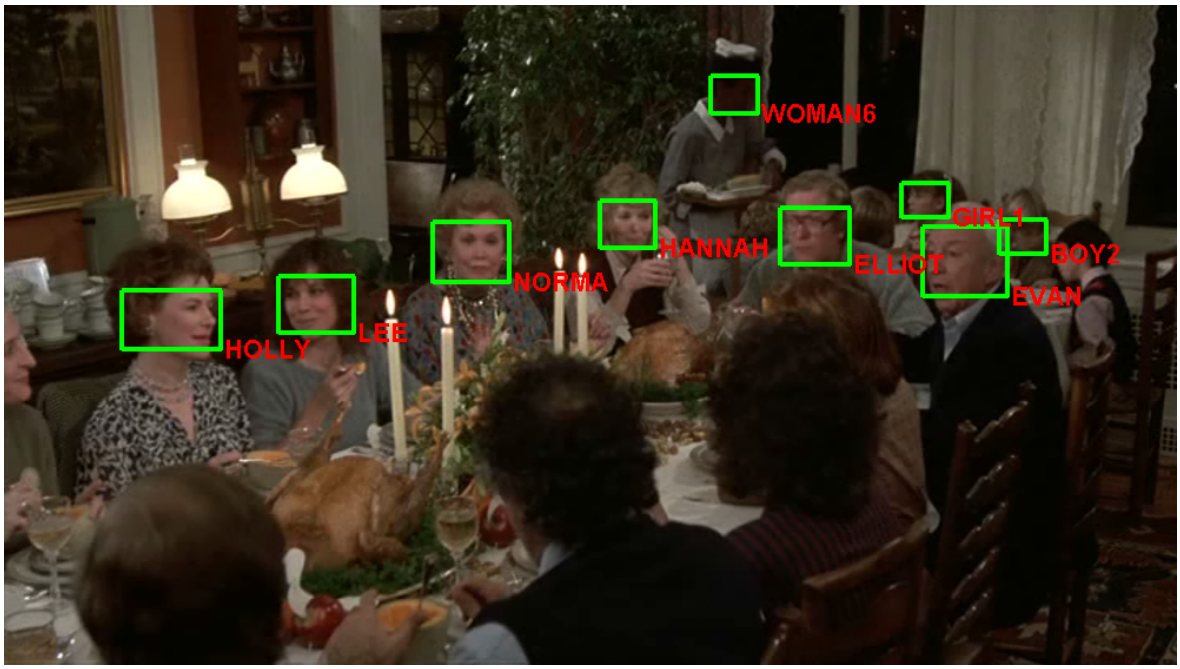


FIGURE 7.1 – Example of spatio-temporal face annotation in Hannah dataset [5].

7.8 Audio style transfer

In this project we tried again adopting an approach proposed in the area of image processing to audio processing.

A so-called *image style transfer* consists in transferring a style from one image (style image) to another image (content image), and it usually works quite well when the style image is a painting and the content image is a photo. As can be seen from an example on Figure 7.2, this kind of methods allow keeping the global structure of the content image while transferring the color palette and the local textures (painter's brush style) from the style image. The work by Gatys *et al* [GEB16] published in 2016 became very popular and gave rise to many new scientific publications, software and online applications allowing image style transfer.

Inspired by work of Gatys *et al* [GEB16], we have proposed an audio style transfer approach [6] which allows transferring sound texture and characteristic spectral patterns from one audio signal to another one (a demo can be found at [5]). However, in contrast to image style transfer, most of audio style transfer methods (including ours [6]) do not provide such spectacular and satisfactory results as in image processing. This is possibly because, on one hand, in audio style transfer it is not very clear what should be transferred exactly and, on the other hand, manipulating audio seems to be more delicate than manipulating images.

5. https://egrinstein.github.io/2017/10/25/ast.html

FIGURE 7.2 – Example of image style transfer (from [6]).

7.9 Tutorials, review paper and book chapters

In 2014, together with other collaborators, I have given two 3-hour long tutorials on informed source separation and on nonnegative matrix factorization at IEEE ICASSP 2014 and IEEE ICME 2014 conferences, respectively. In 2017 I have co-authored a review paper [11] on consolidated perspective on multimicrophone speech enhancement and source separation, published in IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) journal. In 2018 I have co-authored 4 book chapters : 3 on audio source separation [12, 13, 47], and one on audio event detection and scene analysis [48].

Chapitre 8

Conclusion

I have presented my work on solving various inverse problems in audio using probabilistic multichannel NMF/NTF modeling of latent source spectrograms. A unified view of multichannel multisource NMF/NTF modeling was first presented in Chapter 2, and then specialized through Chapters 3 to 6, while following corresponding papers [1, 2, 3, 4] and covering various applications. These approaches have proven their effectiveness for various source separation scenarios, for informed source separation (audio objects compression), and for audio inpainting.

The main short take home message that should be retained from all that is :

"When trying to use NTF model on multichannel and/or degraded audio (e.g., clipped or subsampled), do not try to find an observed tensor to apply NTF to, but rather apply it to the latent tensor of source power spectrograms."

We are now at the beginning of deep learning era and many deep learning-based solutions for the same or similar problems have been proposed. However, NMF/NTF-based methods have not lost their popularity so far. For example in the EDICS of IEEE/ACM Transactions on Audio, Speech, and Language Processing journal [1] there are the following topics for AUD-SEP (Audio and Speech Source Separation) : *Single-channel and multichannel source separation; computational acoustic scene analysis; NMF-based source separation; deep learning methods for source separation.* Indeed, while often very efficient, deep learning is not a universal solution in any situation. Provided that there is a sufficient amount of training data, deep learning-based solutions are usually more powerful than the NMF-based ones, since DNNs allow approaching a greater variety of non-linear complex functions. However, in contrast to NMF, deep learning-based solutions are more difficult to set up (many parameters to tune), are not applicable when there is few or no training data available, are usually not so flexible (i.e., once some conditions have been changed, a deep network usually needs to be retrained), are difficult to be interpreted, and, as a consequence, it is usually difficult to inject some available prior information into the system. As such, instead of choosing between deep learning and NMF, many recent works, e.g. [NLV16], are rather trying combining these two approaches.

In my future research I would like exploring deep learning-based approaches. However, inline with what was just said in the previous paragraph, I am not going completely abandoning NMF or NMF-related ideas, but I will rather try developing hybrid

1. https://signalprocessingsociety.org/publications-resources/ieeeacm-transactions-audio-speech-and-language-processing/edics

approaches. As for applications, I was inspired by impressive results the deep learning-based approaches allowed obtaining in the area of image processing. Notably, they allowed transferring a style from one image to another [GEB16], changing face attributes (e.g., age or mustached) in an image [UGP+17], generating high-resolution images of faces of non-existing celebrities [KALL17], etc ... As such, I would like turning towards similar audio manipulation and audio generation applications. As it was already mentioned in Section 7.8, I have already started working on audio style transfer [6]. However, direct application of methods proposed in image processing (e.g., Gatys *et al* [GEB16] method for style transfer) to audio often does not lead to a desirable result. This is possibly because manipulating audio is more delicate than manipulating images and in audio the features to be manipulated are different. Nevertheless, quite impressive results were obtained for speech synthesis [VDODZ+16] and music translation (across musical instruments, genres, and styles) [MWPT18]. Finally, in line with the PhD thesis of Sanjeel Parekh and in line with my work on informed source separation (Sec. 7.5), I would like developing deep learning-based multimodal approaches to couple audio with other modalities such as images, video and symbolic information (e.g., text or music scores). I hope this will facilitate manipulating, generating and inpainting one modality from another.

References I co-authored

[1] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[2] 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, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[3] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation : Nonnegative tensor factorization approach," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1699–1712, 2013. 1, 41, 42

[4] Ç. Bilen, A. Ozerov, and P. Pérez, "Solving time-domain audio inverse problems using nonnegative tensor factorization," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5604–5617, 2018.

[5] A. Ozerov, J.-R. Vigouroux, L. Chevallier, and P. Pérez, "On evaluating face tracks in movies," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 3003–3007.

[6] E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 586–590.

[7] 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 (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[8] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 121–124.

[9] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, July 2015.

[10] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*, Florence, Italy, May 2014.

[11] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on*

Audio, Speech and Language Processing (TASLP), vol. 25, no. 4, pp. 692–730, 2017.

[12] A. Ozerov and H. Kameoka, "Gaussian model based multichannel separation," in *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[13] A. Ozerov, C. Févotte, and E. Vincent, "An introduction to multichannel NMF for audio source separation," in *Audio Source Separation*. Springer, 2018, pp. 73–94.

[14] "FASST - Flexible Audio Source Separation Toolbox," http://bass-db.gforge.inria.fr/fasst/, accessed : 2019-07-28.

[15] 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, 2007.

[16] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*. IEEE, 2007, pp. 535–539.

[17] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 243–246.

[18] A. Ozerov and W. B. Kleijn, "Asymptotically optimal model estimation for quantization," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 1031–1042, 2011.

[19] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation : source coding meets source separation," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New Paltz, New York, USA, Oct. 2011, pp. 257–260.

[20] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *EUSIPCO, 20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[21] S. Kirbiz, A. Ozerov, A. Liutkus, and L. Girin, "Perceptual coding-based informed source separation," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 959–963.

[22] Ç. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[23] ——, "Joint audio inpainting and source separation," in *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, August 2015.

[24] ——, "Compressive sampling-based informed source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[25] ——, "Automatic allocation of NTF components for user guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[26] A. Ozerov, Ç. Bilen, and P. Pérez, "Multichannel audio declipping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[27] L. Chevallier, J.-R. Vigouroux, A. Goguey, and A. Ozerov, "Facial landmarks localization estimation by cascaded boosted regression," in *International Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2013, pp. 103–115.

[28] D. El Badawy, N. Q. Duong, and A. Ozerov, "On-the-fly audio source separation - A novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[29] A. Ozerov, M. Lagrange, and E. Vincent, "GMM-based classification from noisy features," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011.

[30] ——, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech & Language*, vol. 27, no. 3, pp. 874–894, 2013.

[31] M. Lagrange, A. Ozerov, and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," in *13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[32] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-gmm estimation for underdetermined instantaneous audio source separation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 751–758.

[33] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval, "A tractable framework for estimating and combining spectral source models for audio source separation," *Signal Processing*, vol. 92, no. 8, pp. 1886–1901, 2012.

[34] C. Blandin, E. Vincent, and A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2616–2619.

[35] 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.

[36] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. 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 (LVA/ICA'10)*, Saint-Malo, France, Sep. 2010, pp. 114 – 122.

[37] S. Araki, F. Theis, G. Nolte, D. Lutter, A. Ozerov, V. Gowreesunker, H. Sawada, and N. Q. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010) : Biomedical source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 123–130.

[38] E. Vincent, S. Araki, F. 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, 2012.

[39] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[40] P. Prablanc, A. Ozerov, N. Q. Duong, and P. Pérez, "Text-informed speech inpainting via voice conversion," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 878–882.

[41] D. El Badawy, N. Q. Duong, and A. Ozerov, "On-the-fly audio source separation," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2014, pp. 1–6.

[42] D. El Badawy, A. Ozerov, and N. Q. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 256–260.

[43] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6–10.

[44] ——, "Guiding audio source separation by video object information," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 61–65.

[45] ——, "Weakly supervised representation learning for unsynchronized audio-visual events." in *CVPR Workshops*, 2018, pp. 2518–2519.

[46] S. Parekh, A. Ozerov, S. Essid, N. Q. Duong, P. Pérez, and G. Richard, "Identify, locate and separate : Audio-visual object extraction in large video collections using weak supervision," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019 IEEE Workshop on*. IEEE, 2019.

[47] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF : divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 73–94.

[48] S. Essid, S. Parekh, N. Q. Duong, R. Serizel, A. Ozerov, F. Antonacci, and A. Sarti, "Multiview approaches to event detection and scene analysis," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 243–276.

Other references

[AEJ+12] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio inpainting. *IEEE Transactions on Audio, Speech and Language Processing*, 20(3) :922 – 932, 2012.

[Att03] Hagai Attias. New EM algorithms for source separation and deconvolution with a microphone array. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–297. IEEE, 2003.

[BBR07] Nancy Bertin, Roland Badeau, and Gaël Richard. Blind signal decompositions for automatic transcription of polyphonic music : NMF and K-SVD on the benchmark. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pages I–65. IEEE, 2007.

[BMW14] Nicholas Bryan, Gautham J. Mysore, and Ge Wang. Isse : An interactive source separation editor. In *CHI Conference on Human Factors in Computing Systems*, Toronto, Canada, 04/2014 2014. ACM, ACM. .

[Bro97] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2) :149–171, 1997.

[BSCB00] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *SIGGRAPH'00*, pages 417–424, 2000.

[Com94] Pierre Comon. Independent component analysis, a new concept ? *Signal processing*, 36(3) :287–314, 1994.

[CW08] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25 :21–30, 2008.

[CZ12] Ken Chatfield and Andrew Zisserman. Visor : Towards on-the-fly large-scale object category retrieval. In *Asian Conference on Computer Vision*, pages 432–446. Springer, 2012.

[CZPA09] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations : applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22, 1977.

[DRDF10] Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :564–575, 2010.

[DVG10a] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1830–1840, 2010.

[DVG10b] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 73–80. Springer, 2010.

[EPMP14] S. Ewert, B. Pardo, M. Müller, and M.D. Plumbley. Score-informed source separation for musical audio recordings : An overview. *IEEE Signal Processing Magazine*, 31(3) :116–124, May 2014.

[ERF+08] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen. Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding. In *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008.

[FBD09] 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.

[FCC05] D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *Proc. of the Irish Signals and Systems Conference*, Dublin, Sep. 2005.

[FI11] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9) :2421–2456, 2011.

[GARRM05] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero. Distributed video coding. *Proceedings of the IEEE*, 93(1) :71 – 83, January 2005.

[GEB16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[Gra09] John W Graham. Missing data analysis : Making it work in the real world. *Annual review of psychology*, 60 :549–576, 2009.

[HBB92] Franz Hlawatsch and G Faye Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE signal processing magazine*, 9(2) :21–67, 1992.

[HK06] Rainer Huber and Birger 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, 2006.

[HL04] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1) :30–37, 2004.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv :1710.10196*, 2017.

[Kay93] Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[KBG14] Srđan Kitić, Nancy Bertin, and Rémi Gribonval. Audio declipping by cosparse hard thresholding. In *iTwist - 2nd international - Traveling Workshop onInteractions betweenSparse models and Technology*, Namur, Belgium, August 2014.

[KBG15] Srđan Kitić, Nancy Bertin, and Rémi Gribonval. Sparsity and cosparsity for audio declipping : a flexible non-convex approach. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 243–250. Springer, 2015.

[Kie00] Henk AL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics : A Journal of the Chemometrics Society*, 14(3) :105–122, 2000.

[KJM+13] Srdan Kitic, Laurent Jacques, Nilesh Madhu, Michael Peter Hopwood, Ann Spriet, and Christophe De Vleeschouwer. Consistent iterative hard thresholding for signal declipping. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5939–5943. IEEE, 2013.

[KLIM19] Hirokazu Kameoka, Li Li, Shota Inoue, and Shoji Makino. Supervised determined source separation with multichannel variational autoencoder. *Neural Computation*, pages 1–24, 2019.

[KOS+16] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(9) :1622–1637, 2016.

[Kru77] Joseph B Kruskal. Three-way arrays : rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2) :95–138, 1977.

[KSH18] Hirokazu Kameoka, Hiroshi Sawada, and Takuya Higuchi. General formulation of multichannel extensions of nmf variants. In *Audio Source Separation*, pages 95–124. Springer, 2018.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2013.

[LC10] Lek-Heng Lim and Pierre Comon. Multiarray signal processing : Tensor decomposition meets compressed sensing. *Comptes Rendus Mecanique*, 338(6) :311–320, 2010.

[LCP+08] Hans Laurberg, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data : On the uniqueness of nmf. *Computational intelligence and neuroscience*, 2008, 2008.

[LGH19] S. Leglaive, L. Girin, and R. Horaud. Semi-supervised multichannel speech enhancement with variational autoencodersand non-negative matrix factorization. In *IEEE Int. Conf. onAcoustics, Speech and Signal Process. (ICASSP)*, Brighton, UK, 2019.

[LPB+12] 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. 1, 41, 42

[LRKO+11] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain De Cheveigne, and Shigeki Sagayama. Computational auditory induction as a missing-data model-fitting problem with Bregman divergence. *Speech Communication*, 53(5) :658–676, 2011.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788, 1999.

[LS01] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing 13 (NIPS'2000)*, 2001.

[LSCJ08] Hans Laurberg, Mikkel N Schmidt, Mads Graesboll Christensen, and Soren Holdt Jensen. Structured non-negative matrix factorization with sparsity patterns. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1693–1697. IEEE, 2008.

[MWPT18] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network. *arXiv preprint arXiv :1805.07848*, 2018.

[NLV16] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(9) :1652–1664, 2016.

[NM93] Fredy D Neeser and James L Massey. Proper complex random processes with applications to information theory. *IEEE transactions on information theory*, 39(4) :1293–1302, 1993.

[NV10] Joonas Nikunen and Tuomas Virtanen. Object-based audio coding using non-negative matrix factorization for the spectrogram representation. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[NVV11] Joonas Nikunen, Tuomas Virtanen, and Miikka Vilermo. Multichannel audio upmixing based on non-negative tensor factorization representation. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 33–36. IEEE, 2011.

[OP04] Paul D O'Grady and Barak A Pearlmutter. Soft-LOST : EM on a mixture of oriented lines. In *International Conference on Independent Component Analysis and Signal Separation*, pages 430–436. Springer, 2004.

[PE06] R. M. Parry and I. A. Essa. Estimating the spatial position of spectral components in audio. In *ICA*, pages 666–673, 2006.

[PGB10] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Trans. Audio, Speech, Language Process.*, 18(6) :1464–1475, 2010.

[PS00] Lucas Parra and Clay Spence. Convolutive blind separation of non-stationary sources. *IEEE transactions on Speech and Audio Processing*, 8(3) :320–327, 2000.

[SAM10] Hiroshi Sawada, Shoko Araki, and Shoji Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3) :516–527, 2010.

[SB03] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pages 177–180. IEEE, 2003.

[SG09] M. Spiertz and V. Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of International Conference on Digital Audio Effects (DAFx'09)*, Como, Italy, Sept. 2009.

[SKAU13] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5) :971–982, 2013.

[SKD14] Kai Siedenburg, Matthieu Kowalski, and Monika Dörfler. Audio declipping with social sparsity. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1577–1581. IEEE, 2014.

[SRS07] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007.

[SRS11] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Missing data imputation for time-frequency representations of audio signals. *Journal of signal processing systems*, 65(3) :361–370, 2011.

[SVB+14] Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraa-Labastie, Xabier Jaureguiberry, Dung T Tran, and Frédéric Bimbot. The flexible audio source separation toolbox version 2.0. in Show & Tell, IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[ŞYC12] Umut Şimşekli, Y Kenan Yılmaz, and A Taylan Cemgil. Score guided audio restoration via generalised coupled tensor factorisation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5369–5372. IEEE, 2012.

[UGP+17] Paul Upchurch, Jacob R Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Q Weinberger. Deep feature interpolation for image content changes. In *CVPR*, volume 1, page 3, 2017.

[VAB09] Emmanuel Vincent, Shoko Araki, and Pau Bofill. The 2008 signal separation evaluation campaign : A community-based approach to large-scale evaluation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 734–741. Springer, 2009.

[VAG09] Emmanuel Vincent, Simon Arberet, and Rémi Gribonval. Underdetermined instantaneous audio source separation via local gaussian modeling. In *International Conference on Independent Component Analysis and Signal Separation*, pages 775–782. Springer, 2009.

[VBB09] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :528–537, 2009.

[VDODZ+16] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet : A generative model for raw audio. In *SSW*, page 125, 2016.

[VGF06] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4) :1462–1469, 2006.

[VGP07] E. Vincent, R. Gribonval, and M. Pumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8) :1933 – 1950, August 2007.

[Vin07] Emmanuel Vincent. Complex nonconvex lp norm minimization for underdetermined source separation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 430–437. Springer, 2007.

[Vir07] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3) :1066–1074, 2007.

[XLC04] Z. Xiong, A.D. Liveris, and S. Cheng. Distributed source coding for sensor networks. *IEEE Signal Processing Magazine*, 21(5) :80–94, September 2004.

[ZSN08] David Y Zhao, Jonas Samuelsson, and Mattias Nilsson. On entropy-constrained vector quantization using gaussian mixture models. *IEEE Transactions on Communications*, 56(12) :2094–2104, 2008.

Appendix

Paper 1 (Ozerov & Févotte, *IEEE TASLP*, 2010)

Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation

Alexey Ozerov, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***Abstract*—We consider inference in a general data-driven object-based model of multichannel audio data, assumed generated as a possibly underdetermined convolutive mixture of source signals. We work in the short-time Fourier transform (STFT) domain, where convolution is routinely approximated as linear instantaneous mixing in each frequency band. Each source STFT is given a model inspired from nonnegative matrix factorization (NMF) with the Itakura–Saito divergence, which underlies a statistical model of superimposed Gaussian components. We address estimation of the mixing and source parameters using two methods. The first one consists of maximizing the exact joint likelihood of the multichannel data using an expectation-maximization (EM) algorithm. The second method consists of maximizing the sum of individual likelihoods of all channels using a multiplicative update algorithm inspired from NMF methodology. Our decomposition algorithms are applied to stereo audio source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings. Our EM method produces competitive results with respect to state-of-the-art as illustrated on two tasks from the international Signal Separation Evaluation Campaign (SiSEC 2008).**

***Index Terms*—Expectation-maximization (EM) algorithm, multichannel audio, nonnegative matrix factorization (NMF), nonnegative tensor factorization (NTF), underdetermined convolutive blind source separation (BSS).**

I. Introduction

NONNEGATIVE matrix factorization (NMF) is an unsupervised data decomposition technique with effervescent popularity in the fields of machine learning and signal/image processing [1]. Much research about this topic has been driven by applications in audio, where the data matrix is taken as the magnitude or power spectrogram of a sound signal. NMF was for example applied with success to automatic music transcription [2], [3] and audio source separation [4], [5]. The factorization amounts to decomposing the spectrogram data into a sum of rank-1 spectrograms, each of which being the expression of an elementary spectral pattern amplitude-modulated in time. However, while most music recordings are available in multichannel format (typically, stereo), NMF in its standard setting is only suited to single-channel data. Extensions to multichannel data have been considered, either by stacking up the spectrograms of each channel into a single matrix [6] or by considering nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [7]. These approaches inherently assume that the original sources have been mixed instantaneously, which in modern music mixing is not realistic, and they require a posterior binding step so as to group the elementary components into instrumental sources. Furthermore they do not exploit the redundancy between the channels in an optimal way, as will be shown later.

The aim of this work is to remedy these drawbacks. We formulate a multichannel NMF model that accounts for convolutive mixing. The source spectrograms are modeled through NMF and the mixing filters serve to identify the elementary components pertaining to each source. We consider more precisely I sampled signals $\tilde{x}_i(t)$ $(i = 1, \ldots, I, t = 1, \ldots, T)$ generated as convolutive noisy mixtures of J point source signals $\tilde{s}_j(t)$ $(i = 1, \ldots, J)$ such that

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau) \tilde{s}_j(t-\tau) + \tilde{b}_i(t) \tag{1}$$

where $\tilde{a}_{ij}(\tau)$ is the finite-impulse response of some (causal) filter and $\tilde{b}_i(t)$ is some additive noise. The time-domain mixing given by (1) can be approximated in the short-time Fourier transform (STFT) domain as

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \tag{2}$$

where $x_{i,fn}$, $s_{j,fn}$ and $b_{i,fn}$ are the complex-valued STFTs of the corresponding time signals, $a_{ij,f}$ is the complex-valued discrete Fourier transform of filter $\tilde{a}_{ij}(\tau)$, $f = 1, \ldots, F$ is a frequency bin index, and $n = 1, \ldots, N$ is a time frame index. Equation (2) holds when the filter length L is assumed "significantly" shorter than the STFT window size $(2F - 2)$ [8]. Equation (2) can be rewritten in matrix form, such that

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn} \tag{3}$$

where $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$, $\mathbf{s}_{fn} = [s_{1,fn}, \ldots, s_{J,fn}]^T$, $\mathbf{b}_{fn} = [b_{1,fn}, \ldots, b_{I,fn}]^T$, and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$.

Manuscript received December 24, 2008; revised August 17, 2009. Current version published February 10, 2010. This work was supported in part by the French ANR project SARAH (StAndardisation du Remastering Audio Haute-Définition). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

A. Ozerov was with the Institut Telecom, Telecom ParisTech, CNRS LTCI, 75014 Paris, France. He is now with the METISS Team of IRISA/INRIA, 35042 Rennes Cedex, France (e-mail: alexey.ozerov@irisa.fr).

C. Févotte is with CNRS LTCI, Telecom ParisTech, 75014 Paris, France (e-mail: cedric.fevotte@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2031510

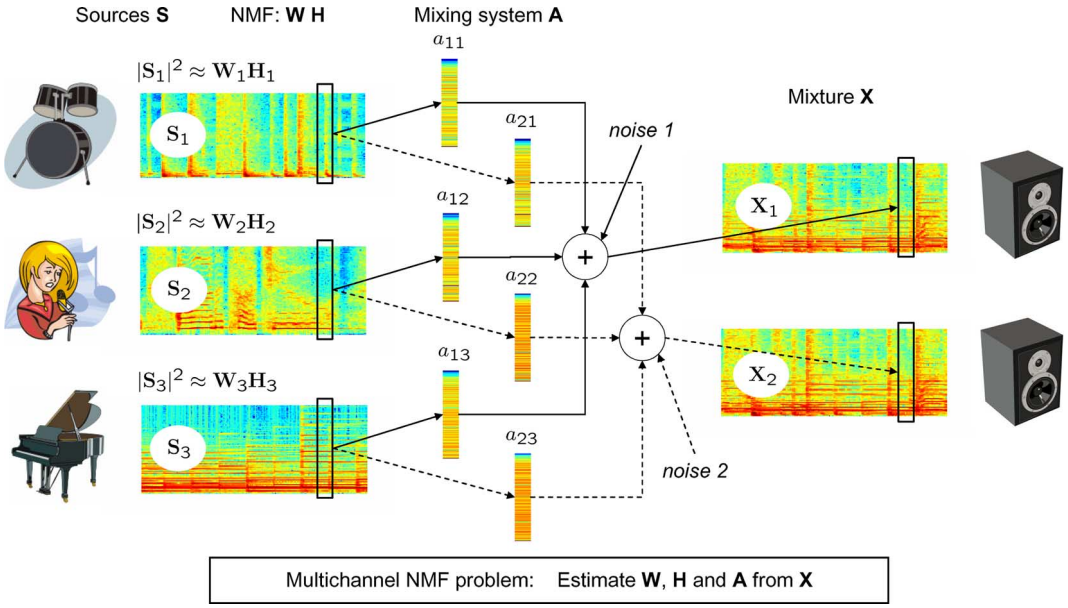


Fig. 1. Representation of convolutive mixing system and formulation of Multichannel NMF problem.

A key ingredient of this work is to model the $F \times N$ power spectrogram $|\mathbf{S}_j|^2 = [|s_{j,fn}|^2]_{fn}$ of source j as a product of two nonnegative matrices $\mathbf{W}_j$ and $\mathbf{H}_j$, such that

$$|\mathbf{S}_j|^2 \approx \mathbf{W}_j \mathbf{H}_j. \tag{4}$$

Given the observed mixture STFTs $\mathbf{X} = \{x_{i,fn}\}_{ifn}$, we are interested in joint estimating the source spectrogram factors $\{\mathbf{W}_j, \mathbf{H}_j\}_j$ and the mixing system $\{\mathbf{A}_f\}_f$, as illustrated in Fig. 1. Our problem splits into two subtasks: 1) defining suitable estimation criteria, and 2) designing algorithms optimizing these criteria.

We adopt a statistical setting in which each source STFT is modeled as a sum of latent Gaussian components, a model introduced by Benaroya *et al.* [9] in a supervised single-channel audio source separation context. A connection between full maximum-likelihood (ML) estimation of the variance parameters in this model and NMF using the Itakura–Saito (IS) divergence was pointed out in [10]. Given this source model, hereafter referred to as *NMF model*, we introduce two estimation criteria together with corresponding inference methods.

- The first method consists of maximizing the exact joint log-likelihood of the multichannel data using an expectation-maximization (EM) algorithm [11]. This method fully exploits the redundancy between the channels, in a statistically optimal way. It draws parallels with several model-based multichannel source separation methods [12]–[18], as described throughout the paper.
- The second method consists of maximizing the sum of individual log-likelihoods of all channels using a multiplicative update (MU) algorithm inspired from NMF methodology. This approach relates to the above-mentioned NTF techniques [6], [7]. However, in contrast to standard NTF which inherently assumes instantaneous mixing, our approach addresses a more general convolutive structure and does not require the posterior binding of the elementary components into J sources.

The general multichannel NMF framework we describe yields a data-driven object-based representation of multichannel data that may benefit many tasks in audio, such as transcription or object-based coding. In this article we will more specifically focus on the convolutive blind source separation (BSS) problem, and as such we also address means of reconstructing source signal estimates from the set of estimated parameters. Our decompositions are conservative in the sense that the spatial source estimates sum up to the original mix. The mixing parameters may also be changed without degrading audio quality, so that music remastering is one potential application of our work. Remixes of well-known songs retrieved from commercial CD recordings are proposed in the results section.

Many convolutive BSS methods have been designed under model (3). Typically, an instantaneous independent component analysis (ICA) algorithm is applied to data $\{\mathbf{x}_{fn}\}_{n=1,\ldots,N}$ in each frequency subband f, yielding a set of J source subband estimates per frequency bin. This approach is usually referred to as frequency-domain ICA (FD-ICA) [19]. The source labels remain however unknown because of the ICA standard permutation indeterminacy, leading to the well-known FD-ICA permutation alignment problem, which cannot be solved without using additional *a priori* knowledge about the sources and/or about the mixing filters. For example in [20] the sources in different frequency bins are grouped *a posteriori* relying on their temporal correlation, thus using prior knowledge about the sources, and in [21], [22] the sources and the filters are estimated assuming a particular structure of convolutive filters, i.e., using prior knowledge about the filters. The permutation ambiguity arises from the individual processing of each subband, which implicitly assumes mutual independence of one source's subbands. This is not the case in our work where our source model implies a coupling of the frequency bands, and joint estimation of the source

parameters and mixing coefficients frees us from the permutation alignment problem.

Our EM-based method is related to some multichannel source separation techniques employing Gaussian mixture models (GMMs) as source models. Univariate independent and identically distributed (i.i.d.) GMMs have been used to model source samples in the time domain for separation of instantaneous [12], [13] and convolutive [12] mixtures. However, such time-domain GMMs are not of the most relevance for audio as they do not model temporal correlations in the signal. In [14], Attias proposes to model the sources in the STFT domain using multivariate GMMs, hence taking into account temporal correlations in the audio signal, assumed stationary in each window frame. The author develops a source separation method for convolutive mixtures, supervised in the sense that the source models are pre-trained in advance. A similar approach with log-spectral domain GMMs is developed by Weiss *et al.* in [15]. Arberet *et al.* [16] propose a multivariate GMM-based separation method for instantaneous mixing that involves a computationally efficient strategy for learning the source GMMs separately, using intermediate source estimates obtained by some BSS method. As compared to these works, we use a different source model (the NMF model), which might be considered more suitable than the GMM for musical signals. Indeed, the NMF is well suited to polyphony as it basically takes the source to be a sum of elementary components with characteristic spectral signatures. In contrast, the GMM takes the source as a single component with many states, each representative of a characteristic spectral signature, but not mixed *per se*. To put it in an other way, in the NMF model a summation occurs in the STFT domain (or equivalently, in the time domain), while in the GMM the summation occurs on the distribution of the frames. Moreover, as discussed later, the computational complexity of inference in our model grows linearly with the number of components while the complexity of standard inference in GMMs grows combinatorially.

The remaining of this paper is organized as follows. NMF source model and noise model are introduced in Section II. Section III is devoted to the definition of our two estimation criteria, with corresponding optimization algorithms. Section IV presents results of our methods to stereo source separation in various settings, including blind and supervised separation of music and speech sources in synthetic instantaneous and convolutive mixtures, as well as in professionally produced music recordings. Conclusions are drawn in Section V. Preliminary aspects of this work are presented in [23]. We here considerably extend on the simulations part as well as on the theoretical developments related to our algorithms.

II. Models

A. Sources

Let $K \geq J$ and $\{\mathcal{K}_j\}_{j=1}^{J}$ be a nontrivial partition of $\mathcal{K} = \{1, \ldots, K\}$. Following [9], [10], we assume the complex random variable $s_{j,fn}$ to be a sum of $\#\mathcal{K}_j$ latent components, such that

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \tag{5}$$

where $w_{fk}, h_{kn} \in \mathbb{R}^+$ and $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the *proper* complex Gaussian distribution [24] with probability density function (pdf)

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\pi \boldsymbol{\Sigma}|} \exp\left[-(\mathbf{x} - \boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]. \tag{6}$$

In the rest of the paper, the quantities $s_{j,fn}$ and $c_{k,fn}$ are, respectively, referred to as "source" and "component". The components are assumed *mutually* independent and *individually* independent across frequency f and frame n. It follows that

$$s_{j,fn} \sim \mathcal{N}_c\left(0, \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn}\right). \tag{7}$$

Denoting $\mathbf{S}_j$ the $F \times N$ STFT matrix $[s_{j,fn}]_{fn}$ of source j and introducing the matrices $\mathbf{W}_j = [w_{fk}]_{f,k \in \mathcal{K}_j}$ and $\mathbf{H}_j = [h_{kn}]_{k \in \mathcal{K}_j, n}$, respectively, of dimensions $F \times \#\mathcal{K}_j$ and $\#\mathcal{K}_j \times N$, it is easily shown [10] that the minus log-likelihood of the parameters describing source j writes

$$-\log p(\mathbf{S}_j | \mathbf{W}_j, \mathbf{H}_j) \stackrel{c}{=} \sum_{fn} d_{IS}\left(|s_{j,fn}|^2 | [\mathbf{W}_j \mathbf{H}_j]_{fn}\right)$$

where "$\stackrel{c}{=}$" denotes equality up to a constant and

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{8}$$

is the IS divergence. In other words, ML estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ given source STFT $\mathbf{S}_j$ is equivalent to NMF of the power spectrogram $|\mathbf{S}_j|^2$ into $\mathbf{W}_j \mathbf{H}_j$, where the IS divergence is used. MU and EM algorithms for IS-NMF are, respectively, described in [25], [26] and in [10]; in essence, this paper describes a generalization of these algorithms to a multichannel multisource scenario. In the following, we will use the notation $\mathbf{P}_j = \mathbf{W}_j \mathbf{H}_j$, i.e., $p_{j,fn} = \mathbb{E}\{|s_{j,fn}|^2\}$.

Our source model is related to the GMM used for example in [14], [16] in the same source separation context, with the difference that one source frame is here modeled as a sum of $\#\mathcal{K}_j$ elementary components while in the GMM one source frame is modeled as a process which can take one of many states, each characterized by a covariance matrix. The computational complexity of inference in our model with our algorithms described next grows linearly with the total number of components while the derivation of the equivalent EM algorithm for GMM leads to an algorithm that has combinatorial complexity with the number of states [12], [13], [15]. It is possible to achieve linear complexity in the GMM case also, but at the price of approximate inference [14], [16]. Note that all considered algorithms, either for the NMF model or GMM, only ensure convergence to a stationary point of the objective function, and, as a consequence, the final result depends strongly on the parameters initialization. We wish to emphasize that we here take a fully data-driven approach in the sense that no parameter is pre-trained.

B. Noise

In the most general case, we may assume noisy data and the following algorithms can easily accommodate estimation of noise statistics under Gaussian independent assumptions and given covariance structures such as $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}$ or $\boldsymbol{\Sigma}_{\mathbf{b},n}$. In

this paper, we consider for simplicity stationary and spatially uncorrelated noise such that

$$b_{i,fn} \sim \mathcal{N}_c\left(0, \sigma^2_{i,f}\right) \tag{9}$$

and $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma^2_{i,f}]_i)$. The musical data we consider in Section IV-A is not noisy in the usual sense, but the noise component can account for model discrepancy and/or quantization noise. Moreover, this noise component is required in the EM algorithm to prevent from potential numerical instabilities (see Section III-A1 below) and slow convergence (see Section III-A6 below). In Section IV-D, we will consider several scenarios: when the variances are equal and fixed to a small value $\tilde{\sigma}^2$, when the variances are estimated from data, and most importantly when annealing is performed via the noise variance, so as to speed up convergence as well as favor global solutions.

C. Convolutive Mixing Model Revisited

With (5), the mixing model (3) can be recast as

$$\mathbf{x}_{fn} = \mathring{\mathbf{A}}_f \mathbf{c}_{fn} + \mathbf{b}_{fn} \tag{10}$$

where $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T \in \mathbb{C}^{K\times 1}$ and $\mathring{\mathbf{A}}_f$ is the so called "augmented mixing matrix" of dimension $I \times K$, with elements defined by $\mathring{a}_{ik,f} = a_{ij,f}$ if and only if $k \in \mathcal{K}_j$. Thus, for every frequency bin f, our model is basically a linear mixing model with I channels and K elementary Gaussian sources $c_{k,fn}$, with structured mixing coefficients (i.e., subsets of elementary sources are mixed identically). Subsequently, we will note $\mathbf{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}([w_{fk}h_{kn}]_k)$ the covariance of $\mathbf{c}_{fn}$.

III. Methods

A. Maximization of Exact Likelihood With EM

1) Criterion: Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}, \mathbf{\Sigma}_\mathbf{b}\}$ be the set of all parameters, where $\mathbf{A}$ is the $I \times J \times F$ tensor with entries $a_{ij,f}$, $\mathbf{W}$ is the $F \times K$ matrix with entries w_{fk}, $\mathbf{H}$ is the $K \times N$ matrix with entries h_{kn}, and $\mathbf{\Sigma}_\mathbf{b}$ are the noise covariance parameters. Under previous assumptions, data vector $\mathbf{x}_{fn}$ has a zero-mean proper Gaussian distribution with covariance

$$\mathbf{\Sigma}_{\mathbf{x},fn}(\boldsymbol{\theta}) = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f} \tag{11}$$

where $\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}([p_{j,fn}]_j)$ is the covariance of $\mathbf{s}_{fn}$. ML estimation is consequently shown to amount to minimization of

$$C_1(\boldsymbol{\theta}) = \sum_{fn} \mathrm{trace}\left(\mathbf{x}_{fn}\mathbf{x}_{fn}^H \mathbf{\Sigma}^{-1}_{\mathbf{x},fn}\right) + \log\det \mathbf{\Sigma}_{\mathbf{x},fn}. \tag{12}$$

The noise covariance term $\mathbf{\Sigma}_{\mathbf{b},f}$ appears necessary so as to prevent from ill-conditioned inverses that occur if 1) $\mathrm{rank}(\mathbf{A}_f) < I$, and in particular if $I > J$, i.e., in the overdetermined case, or if 2) $\mathbf{\Sigma}_{\mathbf{s},fn}$ has more than $(J - I)$ null diagonal coefficients in the underdetermined case $(I < J)$. Case 2) might happen in regions of the time–frequency plane where sources are inactive.

For fixed f and n, the BSS problem described by (3) and (12), and the following EM algorithm, is reminiscent of works by Cardoso *et al.*, see, e.g., [27] for the square noise-free case, [17] for other cases and [18] for use in an audio setting. In these papers, a grid of the representation domain is chosen, in each cell of which the source statistics are assumed constant. This is not required in our case where we instead solve F parallel linear instantaneous mixtures tied across frequency by the source model.[1]

2) Indeterminacies: Criterion (12) suffers from obvious scale, phase and permutation indeterminacies.[2] Regarding scale and phase, let $\hat{\boldsymbol{\theta}} = \{\{\mathbf{A}_f\}_f, \{\mathbf{W}_j, \mathbf{H}_j\}_j\}$ be a minimizer of (12) and let $\{\mathbf{D}_f\}_f$ and $\{\mathbf{\Lambda}_j\}_j$ be sets of respectively *complex* and *nonnegative* diagonal matrices. Then, the set

$$\tilde{\boldsymbol{\theta}} = \left\{\left\{\mathbf{A}_f\mathbf{D}_f^{-1}\right\}_f, \left\{\mathrm{diag}\left(\left[|d_{jj,f}|^2\right]_f\right)\mathbf{W}_j\mathbf{\Lambda}_j^{-1}\right\}_j, \{\mathbf{\Lambda}_j\mathbf{H}_j\}_j\right\}$$

leads to $\mathbf{\Sigma}_{\mathbf{x},fn}(\hat{\boldsymbol{\theta}}) = \mathbf{\Sigma}_{\mathbf{x},fn}(\tilde{\boldsymbol{\theta}})$, hence same likelihood value. Similarly, permuted diagonal matrices would also leave the criterion unchanged. In practice, we remove the scale and phase ambiguity by imposing $\sum_i |a_{ij,f}|^2 = 1$ and $a_{1j,f} \in \mathbb{R}^+$ (and scaling the rows of $\mathbf{W}_j$ accordingly) and then by imposing $\sum_f w_{fk} = 1$ (and scaling the rows of $\mathbf{H}_j$ accordingly). With these conventions, the columns of $\mathbf{A}_f$ convey normalized mixing proportions between the channels, the columns of $\mathbf{W}$ convey normalized frequency shapes and all time-dependent amplitude information is relegated into $\mathbf{H}$.

3) Algorithm: We derive an EM algorithm based on *complete data* $\{\mathbf{X}, \mathbf{C}\}$, where $\mathbf{C}$ is the $K \times F \times N$ STFT tensor with coefficients $c_{k,fn}$. The complete data pdfs $\{p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11] or [29, Appendix]) and the set $\{\mathbf{R}_{\mathbf{xx},f}, \mathbf{R}_{\mathbf{xs},f}, \mathbf{R}_{\mathbf{ss},f}, \{u_{k,fn}\}_{kn}\}_f$ defined by

$$\mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \quad \mathbf{R}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H \tag{13}$$

$$\mathbf{R}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H, \quad u_{k,fn} = |c_{k,fn}|^2 \tag{14}$$

is shown to be a *natural (sufficient) statistics* [29] for this family. Thus, one iteration of EM consists of computing the expectation of the natural statistics conditionally on the current parameter estimates (E step) and of reestimating the parameters using the updated natural statistics, which amounts to maximizing the conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int[\log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})]p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}')d\mathbf{C}$ (M step). The resulting updates are given in Algorithm 1, with more details given in Appendix A.

Algorithm 1 EM algorithm (one iteration)

- **E step**. Conditional expectations of natural statistics:

$$\widehat{\mathbf{R}}_{\mathbf{xx},f} = \mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \tag{15}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H, \tag{16}$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{s},fn} - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn} \tag{17}$$

$$\hat{u}_{k,fn} = \left[\hat{\mathbf{c}}_{fn}\hat{\mathbf{c}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{c},fn} - \mathbf{G}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f\mathbf{\Sigma}_{\mathbf{c},fn}\right]_{k,k} \tag{18}$$

[1]In [17] and [27], the ML criterion can be recast as a measure of fit between observed and parameterized covariances, where the measure of deviation writes $D(\mathbf{\Sigma}_1|\mathbf{\Sigma}_2) = \mathrm{trace}(\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}) - \log\det\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1} - I$ and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are positive definite matrices of size $I \times I$ (note that the IS divergence is obtained in the special case $I = 1$). The measure is simply the KL divergence between the pdfs of two zero-mean Gaussians with covariances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Such a formulation cannot be used in our case because $\mathbf{\Sigma}_1 = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$ is not invertible for $I > 1$.

[2]There might also be other less obvious indeterminacies, such as those inherent to NMF (see, e.g., [28]), but this study is here left aside.

where
$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1} \tag{19}$$
$$\hat{\mathbf{c}}_{fn} = \mathbf{G}_{\mathbf{c},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{c},fn} = \boldsymbol{\Sigma}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1} \tag{20}$$
$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f} \tag{21}$$
$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left(\left[\sum_{k\in\mathcal{K}_j} w_{fk}h_{kn}\right]_j\right) \tag{22}$$
$$\boldsymbol{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}\left([w_{fk}h_{kn}]_k\right) \tag{23}$$

and $\mathring{\mathbf{A}}_f$ is defined in Section II-C.

- **M step**. Update the parameters:

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f}\widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \tag{24}$$
$$\boldsymbol{\Sigma}_{\mathbf{b},f} = \mathrm{diag}\left(\widehat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{xs},f}^H - \widehat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\right) \tag{25}$$
$$w_{fk} = \frac{1}{N}\sum_n \frac{\hat{u}_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f \frac{\hat{u}_{k,fn}}{w_{fk}}. \tag{26}$$

- Normalize $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-A2.

4) Implementation Issues: The computation of the source Wiener gain $\mathbf{G}_{\mathbf{s},fn}$ given by (19) requires the inversion of the $I \times I$ matrix $\boldsymbol{\Sigma}_{\mathbf{x},fn}$ at every time–frequency (TF) point. When $I > J$ (overdetermined case) it may be preferable for sake of computational efficiency to use the following alternative formulation of $\mathbf{G}_{\mathbf{s},fn}$, obtained using Woodbury matrix identity [30]

$$\mathbf{G}_{\mathbf{s},fn} = \boldsymbol{\Xi}_{\mathbf{s},fn}^{-1}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \tag{27}$$

with

$$\boldsymbol{\Xi}_{\mathbf{s},fn} = \mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f + \boldsymbol{\Sigma}_{\mathbf{s},fn}^{-1}. \tag{28}$$

This second formulation requires the inversion of the $J \times J$ matrix $\boldsymbol{\Xi}_{\mathbf{s},fn}$ instead of the inversion of the $I \times I$ matrix $\boldsymbol{\Sigma}_{\mathbf{x},fn}$. The same idea applies to the computation of $\mathbf{G}_{\mathbf{c},fn}$, (20), if $I > K$. Thus, this second formulation may become interesting in practice only if $I > J$ and $I > K$, i.e., if $I > K$ (recall that $K \geq J$). As we only consider undetermined mixtures in the experimental part of this article ($I < J$), we turn to the original formulation given by (19). As we more precisely consider stereo mixtures, we only need inverting 2×2 matrices per TF point and our MATLAB code was efficiently vectorized so as to manipulate time–frequency matrices directly, thanks to Cramer's explicit matrix inversion formula. Note also that we only need to compute the diagonal elements of the $K \times K$ matrix in (18). Hence, the computational complexity of one EM algorithm iteration grows linearly (and not quadratically) with the number of components.

5) Linear Instantaneous Case: Linear instantaneous mixing is a special case of interest, that concerns for example "pan pot" mixing. Here, the mixing matrix is real-valued and shared between all the frequency subbands, i.e., $\mathbf{A}_f = \mathbf{A}_{\mathrm{inst}} \in \mathbb{R}^{I\times J}$. In that case, (24) needs only be replaced by

$$\mathbf{A}_{\mathrm{inst}} = \Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{xs},f}\right\}\left[\Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{ss},f}\right\}\right]^{-1}. \tag{29}$$

6) Simulated Annealing: If one computes $\mathbf{A}_f$ through (24), (16), (17), (19), and (21), assuming $\boldsymbol{\Sigma}_{\mathbf{b},f} = 0$, one has $\mathbf{A}_f = \mathring{\mathbf{A}}_f$ as result. Thus, by continuity, when the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{b},f}$ tends to zero, the resulting update rule for $\mathbf{A}_f$ tends to $\mathbf{A}_f \leftarrow \mathring{\mathbf{A}}_f$. Hence, the convergence of $\mathbf{A}_f$ becomes very slow for small values of $\sigma_{i,f}^2$. To overcome this difficulty and also favor global convergence, we have tested in the experimental section several simulated annealing strategies. In our framework, simulated annealing consists in setting the noise variances $\sigma_{i,f}^2$ to a common iteration-dependent value $\sigma_{i,f}^2(iter)$, initialized with an arbitrary large value $\hat{\sigma}_{i,f}^2$ and gradually decreased through iterations to a small value $\tilde{\sigma}_{i,f}^2$. Besides improving convergence speed, this scheme should also favor convergence to global solutions, as typical of annealing algorithms: the cost function is rendered flatter in the first iterations due to the (assumed) presence of high noise, smoothing out local minima, and is gradually brought back to its exact shape in the subsequent iterations.

7) Reconstruction of the Sources: Minimum mean square error (MMSE) estimates $\hat{\mathbf{s}}_{fn} = \mathbb{E}[\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}]$ of the source STFTs are directly retrieved using Wiener filter of (19). Time-domain sources may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window (see e.g., [31]).

By conservativity of Wiener reconstruction the spatial images of the estimated sources and of the estimated noise sum up to the original mix in STFT domain, i.e., $\hat{\mathbf{A}}_f$, $\hat{\mathbf{s}}_{fn}$, and $\hat{\mathbf{b}}_{fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}$ satisfy (3). Thanks to linearity of the inverse-STFT, the reconstruction is conservative in the time domain as well.

B. Maximization of Individual Likelihoods With MU Rules

1) Criterion: We now consider a different approach consisting of maximizing the sum of individual channel log-likelihoods $\sum_i \log p(\mathbf{X}_i|\boldsymbol{\theta})$, hence discarding mutual information between the channels. This is equivalent to setting the off-diagonal terms of $\mathbf{x}_{fn}\mathbf{x}_{fn}^H$ and $\boldsymbol{\Sigma}_{\mathbf{x},fn}$ to zero in criterion (12), leading to minimization of cost

$$C_2(\boldsymbol{\theta}) = \sum_{ifn} d_{IS}\left(|x_{i,fn}|^2|\hat{v}_{i,fn}\right) \tag{30}$$

where $\hat{v}_{i,fn}$ is the structure defined by

$$\hat{v}_{i,fn} = \sum_j q_{ij,f}\underbrace{\sum\nolimits_{k\in\mathcal{K}_j} w_{fk}h_{kn}}_{p_{j,fn}} \quad \left(+\sigma_{i,f}^2\right) \tag{31}$$

and $q_{ij,f} = |a_{ij,f}|^2$. For a fixed channel i, $\hat{v}_{i,fn}$ is basically the sum of the source variances modulated by the mixing weights. A noise variance term $\sigma_{i,f}^2$ might be considered, either fixed or to be estimated, but we will simply set it to zero as we will not here encounter the issues described in Section III-A6 about convergence of EM in noise-free observations.

Criterion (30) may also be read as the ML criterion corresponding to the model where the contributions of each component (and thus, of each source) to each channel would be different and independent realizations of the same Gaussian process, as opposed to the same realization. In other words, this

assumption amounts to changing our observation and source models given by (2) and (5) to

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn}^{(i)} + b_{i,fn} \quad (32)$$

$$s_{j,fn}^{(i)} = \sum_{k \in \mathcal{K}_j} c_{k,fn}^{(i)} \quad \text{with} \quad c_{k,fn}^{(i)} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \quad (33)$$

and thus changing (7) to

$$s_{j,fn}^{(i)} \sim \mathcal{N}_c\left(0, \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn}\right) \quad (34)$$

where $c_{k,fn}^{(i)}$ (resp. $s_{j,fn}^{(i)}$) denotes the contribution of component k (resp. source j) to channel i, and these contributions are assumed independent over channels (i.e., over i).

Our approach differs from the NTF approach of [6], [7] where the following PARAFAC structure [32] is considered

$$\hat{v}_{i,fn}^{NTF} = \sum_k q_{ik}^{NTF} w_{fk} h_{kn}. \quad (35)$$

It is only a sum of $I \times F \times N$ rank-1 tensors and amounts to assuming that $\hat{\mathbf{V}}_i^{NTF} = [\hat{v}_{i,fn}^{NTF}]_{fn}$ is a linear combination of $F \times N$ time–frequency patterns $\mathbf{w}_k h_k$, where $\mathbf{w}_k$ is column k of $\mathbf{W}$ and h_k is row k of $\mathbf{H}$. It intrinsically implies a linear instantaneous mixture and requires a postprocessing binding step in order to group the K elementary patterns into J sources, based on clustering of the ratios $\{q_{1k}^{NTF}/q_{2k}^{NTF}\}_k$ (in the stereo case). To ease comparison, our model can be rewritten as

$$\hat{v}_{i,fn} = \sum_k \overset{\circ}{q}_{ik,f} w_{fk} h_{kn} \quad (36)$$

subject to the constraint $\overset{\circ}{q}_{ik,f} = q_{ij,f}$ if and only if $k \in \mathcal{K}_j$ (with the notation introduced in Section II-C, we have also $\overset{\circ}{q}_{ik,f} = |\overset{\circ}{a}_{ik,f}|^2$). Hence, our model has the following merits with respect to (w.r.t.) the PARAFAC-NTF model: 1) it accounts for convolutive mixing by considering frequency-dependent mixing proportions ($\overset{\circ}{q}_{ik,f}$ instead of q_{ik}^{NTF}) and 2) the constraint that the K mixing proportions $\{\overset{\circ}{q}_{ik,f}\}_k$ can only take J possible values implies that the clustering of the components is taken care of within the decomposition as opposed to after the decomposition.

We have here chosen to use the IS divergence as a measure of fit in (30) because it connects with the optimal inference setting of Section III-A and because it was shown a relevant cost for factorization of audio power spectrograms [10], but other costs could be considered, such as the standard Euclidean distance and the generalized Kullback–Leibler (KL) divergence, which are the costs considered in [6] and [7].

2) Indeterminacies: Criterion (30) suffers from same scale, phase and permutations ambiguities as criterion (12), with the exception that ambiguity on the phase of $a_{ij,f}$ is now total as this parameter only appears through its squared-modulus. In the following, the scales are fixed as in Section III-A2.

3) Algorithm: We describe for the minimization of $C_2(\boldsymbol{\theta})$ an iterative MU algorithm inspired from NMF methodology [1], [33], [34]. Continual descent of the criterion under this algorithm was observed in practice. The algorithm simply consists of updating each scalar parameter θ_l by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t. this parameter, namely

$$\theta_l \leftarrow \theta_l \frac{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-}{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+} \quad (37)$$

where $\nabla_{\theta_l} C_2(\boldsymbol{\theta}) = [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+ - [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-$ and the summands are both nonnegative [10]. Not any cost function gradient may be separated in two such summands, but this is the case for the Euclidean, KL and IS costs, and more generally the β-divergence of which they are specific cases [10], [26]. This scheme automatically ensures the non-negativity of the parameter updates, provided initialization with a nonnegative value.

The resulting parameter updates are described in Algorithm 2, where "." indicates element-wise matrix operations, $\mathbf{1}_{N\times 1}$ is a N-vector of ones, $\mathbf{q}_{ij}$ is the $F \times 1$ vector $[q_{ij,f}]_f$ and $\mathbf{V}_i$ (resp. $\hat{\mathbf{V}}_i$) is the $F \times N$ matrix $[|x_{i,fn}|^2]_{fn}$ (resp. $[\hat{v}_{i,fn}]_{fn}$). Some details about the derivation of the algorithm are given in Appendix B.

Algorithm 2 MU rules (one iteration)

- Update $\mathbf{Q}$

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij} . \frac{\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N\times 1}}{\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N\times 1}}. \quad (38)$$

- Update $\mathbf{W}$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j . \frac{\sum_{i=1}^{I} \mathrm{diag}(\mathbf{q}_{ij}) \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right) \mathbf{H}_j^T}{\sum_{i=1}^{I} \mathrm{diag}(\mathbf{q}_{ij}) \hat{\mathbf{V}}_i^{-1} \mathbf{H}_j^T}. \quad (39)$$

- Update $\mathbf{H}$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j . \frac{\sum_{i=1}^{I} (\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j)^T \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right)}{\sum_{i=1}^{I} (\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j)^T \hat{\mathbf{V}}_i^{-1}}. \quad (40)$$

- Normalize $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-B2.

4) Linear Instantaneous Case: In the linear instantaneous case, when $q_{ij,f} = q_{ij}$, we obtain the following update rule for the mixing matrix coefficients:

$$q_{ij} \leftarrow q_{ij} . \frac{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right]}{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right]} \quad (41)$$

where $\mathrm{sum}[\mathbf{M}]$ is the sum of all coefficients in $\mathbf{M}$. Then, $\mathrm{diag}(\mathbf{q}_{ij})$ needs only be replaced by q_{ij} in (39) and (40). The overall algorithm yields a specific case of PARAFAC-NTF which directly assigns the elementary components to J directions of arrival (DOA). This scheme however requires to fix in advance the partition $\{\mathcal{K}_j\}_{j=1}^J$ of $\mathcal{K} = \{1, \ldots, K\}$, i.e., assign a given number of components per DOA. In the specific linear instantaneous case, multiplicative updates for the whole matrices $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$ can be exhibited (instead of individual updates for q_{ij}, $\mathbf{W}_j$, $\mathbf{H}_j$), but are not given here for conciseness. They are similar in form to [33], [34] and lead to a faster MATLAB implementation.

5) Reconstruction of the Source Images: Criterion (30) being equivalent to the ML criterion under the model defined by (32) and (33), the MMSE estimate $\hat{s}_{j,fn}^{(i)\mathrm{im}} = \mathbb{E}[s_{j,fn}^{(i)\mathrm{im}} | \mathbf{x}_{fn}; \boldsymbol{\theta}]$ of the

image $s_{j,fn}^{(i)\mathrm{im}} \stackrel{\mathrm{def}}{=} a_{ij,f} s_{j,fn}^{(i)}$ of source j in channel i is computed through

$$\hat{s}_{j,fn}^{(i)\mathrm{im}} = \frac{q_{ij,f} p_{i,fn}}{\hat{v}_{i,fn}} x_{i,fn} \tag{42}$$

i.e., by Wiener filtering of each channel. A noise component (if any) can similarly be reconstructed as $\hat{b}_{i,fn} = (\sigma_{i,f}^2 / \hat{v}_{i,fn}) x_{i,fn}$. Overall the decomposition is conservative, i.e., $\sum_j \hat{s}_{j,fn}^{(i)\mathrm{im}} + \hat{b}_{i,fn} = x_{i,fn}$.

IV. Experiments

In this section, we first describe the test data and evaluation criteria, and then proceed with experiments. All the audio datasets and separation results are available from our demo web page [35]. MATLAB implementations of the proposed algorithms are also available from the authors' web pages.

A. Datasets

Four audio datasets have been considered and are described below.

- **Dataset A** consists of two synthetic stereo mixtures, one instantaneous the other convolutive, of $J = 3$ musical sources (drums, lead vocals and piano) created using 17-s excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Commons License at [36] and downsampled to 16 kHz. The mixing parameters (instantaneous mixing matrix and the convolutive filters) were taken from the 2008 Signal Separation Evaluation Campaign (SiSEC'08) "under-determined speech and music mixtures" task development datasets [37], and are described below.
- **Dataset B** consists of synthetic (instantaneous and convolutive) and live-recorded (convolutive) stereo mixtures of speech and music sources, corresponding to the test data for the 2007 Stereo Audio Source Separation Evaluation Campaign (SASSEC'07) [38]. It also coincides with development dataset `dev2` of SiSEC'08 "under-determined speech and music mixtures" task. All the mixtures are 10 s long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. The synthetic convolutive filters were generated with the Roomsim toolbox [39]. They simulate a pair of omnidirectional microphones placed 1 m apart in a room of dimensions $4.45 \times 3.55 \times 2.5$ m with reverberation time 130 ms, which correspond to the setting employed for the live-recorded mixtures. The distances between the sources and the center of the microphone pair vary between 80 cm and 1.20 m. For all mixtures the source directions of arrival vary between -60° and $+60^\circ$ with a minimal spacing of 15° (for more details see [37]).
- **Dataset C** consists of SiSEC'08 test and development datasets for task "professionally produced music recordings". The test dataset consists of two excerpts (of about 22 s long) from two different professionally produced stereo songs, namely "Que pena tanto faz" by Tamy and "Roads" by Bearlin. The development dataset consists of two other excerpts (of about 12 s long) from the same songs, with all original stereo tracks provided separately. All recordings are sampled at 44 kHz (CD quality).
- **Dataset D** consists of three excerpts of length between 25 and 50 s taken from three professionally produced stereo recordings of well-known pop and reggae songs, and downsampled to 22 kHz.

TABLE I
STFT Window Lengths Used in Different Experiments

experiment section	dataset	window length		sampling freq. (Hz)
		samples	milliseconds	
IV-D, IV-E	A	1024	64	16000
IV-F	B - inst.	1024	64	16000
	B - conv.	2048	128	16000
IV-G	C	2048	46	44100
IV-H	D	2048	93	22050

B. Source Separation Evaluation Criteria

In order to evaluate our multichannel NMF algorithms in terms of audio source separation we use the signal-to-distortion ratio (SDR) numerical criterion defined in [38], which essentially compares the reconstructed source images with the original ones. The quality of the mixing system estimates was assessed with the mixing error ratio (MER) described at [37], which is an SNR-like criterion expressed in decibels. MATLAB routines for computing these criteria were obtained from the SiSEC'08 web page [37]. These evaluation criteria can only be computed when the original source spatial images (and mixing systems) are available. When not (i.e., for datasets C and D), separation performance is assessed perceptually and informally by listening to the separated source images, available online at [35].

C. Algorithm Parameters

1) STFT Parameters: In all the experiments below we used STFTs with half-overlapping sine windows, using the STFT computation tools for MATLAB available from [37]. The choice of the STFT window size is rather important, and is a matter of compromise between 1) good frequency resolution and validity of the convolutive mixing approximation of (2) and 2) validity of the assumption of source local stationarity. We have tried various window sizes (powers of 2) for every experiment, and the most satisfactory window sizes are reported in Table I.

2) Model Order: In our case the model order parameters consist of the total number of components K and the allocation of the components among the J sources, i.e., the partition $\{\mathcal{K}_1, \ldots, \mathcal{K}_J\}$. The value of J may be set by hand to the number of instrumental sources in the recording, although, as we shall discussed later, the existence of non-point sources or the existence of sources mixed similarly might render the choice of J trickier. The choice of the number of components per source may raise more questions. As a first guess one may choose a high value, so that the model can account for all of the diversity of the source; basically, one may think of one component per note or elementary sound object. This leads to increased flexibility in the model, but, at the same time, can lead to data overfitting (in case of few data), and favors the existence of local minima,

thus rendering optimization more difficult, as well as more intensive. Interestingly, it has been noted in [10] that, given a limited number of components, IS-NMF is also able to learn higher level structures in the musical signal. One or a few components can capture a large part of one source or a subset of sources, so that a coherent sound decomposition can be achieved to some extent. A similar behavior was logically observed in our multichannel scenario, with even more success as the spatial information helps to discriminate between the sources. Hence, satisfying source separation results could be obtained with small values of K.

In the experiments of Sections IV-D and IV-E we set $\#\mathcal{K}_j = 4$; however, this has minor importance there as the aim of these experiments is merely to investigate the algorithms behavior, and not to obtain optimal source separation performance. In the experiments of Sections IV-F and IV-G, $\#\mathcal{K}_j$ is chosen by hand through trials so as to obtain most satisfying results. In the experiment of Section IV-H the total number of components is arbitrary set to either $K = 15$ or 20, depending on the recording, and the numbers of components per source $\#\mathcal{K}_j$ are chosen automatically by the initialization procedure, see below.

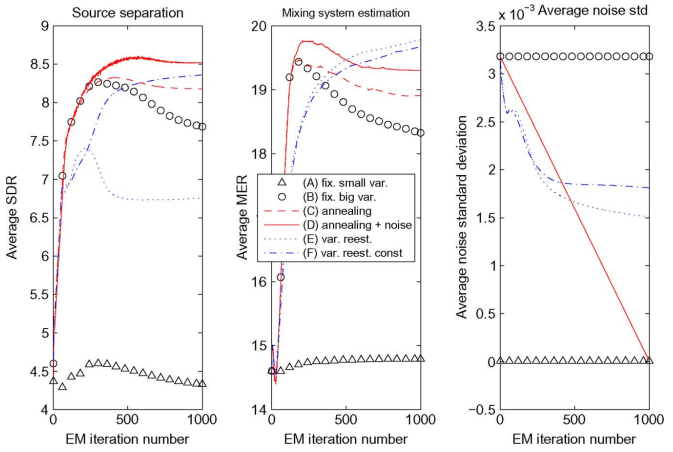


Fig. 2. EM algorithm results on convolutive mixture of dataset A, using various noise variance update schemes. (Left) Average source separation SDR. (Middle) average mixing system identification MER. (Right) average noise standard deviation. (A) Triangles: small fixed noise variance. (B) Circles: larger fixed noise variance. (C) Dashed line: annealing. (D)Solid line: annealing with noise injection. (E) Dotted line: diagonal noise covariance reestimation. (F) Dash-dotted line: isotropic noise variance reestimation.

D. Dealing With the Noise Part in the EM Algorithm

In this section, we experiment strategies for updating the noise parameters in the EM algorithm. We here arbitrarily use the convolutive mixture of dataset A and set the total number of components to $K = 12$, equally distributed between $J = 3$ sources. Our EM algorithm being sensitive to parameters initialization, we used the following *perturbed oracle* initializations so as to ensure "good" initialization: factors $\mathbf{W}$ and $\mathbf{H}$ as computed from the original sources using IS-NMF [10] and original mixing system $\mathbf{A}$, all perturbed with high level additive noise. We have tested the following noise update schemes.

- (**A**): $\mathbf{\Sigma}_{\mathbf{b},f} = \tilde{\sigma}^2\mathbf{I}_I$, with fixed $\tilde{\sigma}^2$ set to 16-bit PCM quantization noise variance.
- (**B**): $\mathbf{\Sigma}_{\mathbf{b},f} = \hat{\sigma}_f^2\mathbf{I}_I$, with fixed $\hat{\sigma}_f^2$ set to the average channel empirical variance in every frequency band divided by 100, i.e., $100\hat{\sigma}_f^2 = \sum_{in} |x_{i,fn}|^2/IN$.
- (**C**): $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ with standard deviation σ_f decreasing linearly through iterations from $\hat{\sigma}_f$ to $\tilde{\sigma}$. This is what we refer to as simulated annealing.
- (**D**): Same strategy as (C), but with adding a random noise with covariance $\mathbf{\Sigma}_{\mathbf{b},f}$ to $\mathbf{X}$ at every EM iteration. We refer to this as annealing with noise injection.
- (**E**): $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma_{i,f}^2]_i)$ is reestimated with update (25).
- (**F**): Noise covariance is reestimated like in scheme E, but under the more constrained structure $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ (isotropic noise in each subband). In that case, operator $\mathrm{diag}(\cdot)$ in (25) needs to be replaced with $\mathrm{trace}(\cdot)\mathbf{I}_I/I$.

The algorithm was run for 1000 iterations in each case and the results are presented in Fig. 2, which displays the average SDR and MER along iterations, as well as the noise standard deviations $\sigma_{i,f}$, averaged over all channels i and frequencies f. As explained in Section III-A6, we observe that with a small fixed noise variance (scheme A), the mixing parameters stagnate. With a fixed larger noise variance (scheme B) convergence starts well but then performance drops due to artificially high noise variance. Simulated annealing (scheme C) overcomes this problem, and artificial noise injection (scheme D) even improves the results (both in terms of source separation and mixing system estimation). Noise variance reestimation allows to obtain performances almost similar to annealing, but only in the case when the variance is constrained to be the same in both channels (scheme F). However, we observed that faster convergence is obtained in general using annealing with noise injection (scheme D) for similar results.

Finally, it should be noted that for the schemes with annealing (C and D) both the average SDR and MER start decreasing from about 400 iterations (for SDR) and 200 iterations (for MER). We believe this is because the final noise variance $\tilde{\sigma}^2$ (set to 16-bit PCM quantization noise variance) might be too small to account for discrepancy in the convolutive mixing equation STFT-approximation (2). Indeed, with scheme F (constrained reestimated variance) the average noise standard deviation seem to be converging to a value in the range of 0.002 (see right plot of Fig. 2), which is much larger than $\tilde{\sigma}$. Thus, if computation time is not an issue, scheme F can be considered the most advantageous because this is the only scheme to systematically increase both the average SDR and MER at every iteration and it allows to adjust a suitable noise level adaptively. However, as we want to keep the number of iterations low (e.g., 300–500) for sake of short computation time, we will resort to scheme D in the following experiments.

E. Convergence and Separation Performance

In this experiment we wish to check consistency of optimization of the proposed criteria with respect to source separation performance improvement, in the least as measured by the SDR. We used both mixtures of dataset A (instantaneous and convolutive) and ran 1000 iterations of both algorithms (EM and MU) from ten different perturbed oracle initializations, obtained as in previous section. Again we used $K = 12$ components, equally split into $J = 3$ sources. Figs. 3 and 4 report results for the instantaneous and convolutive mixtures, respectively. Plots on

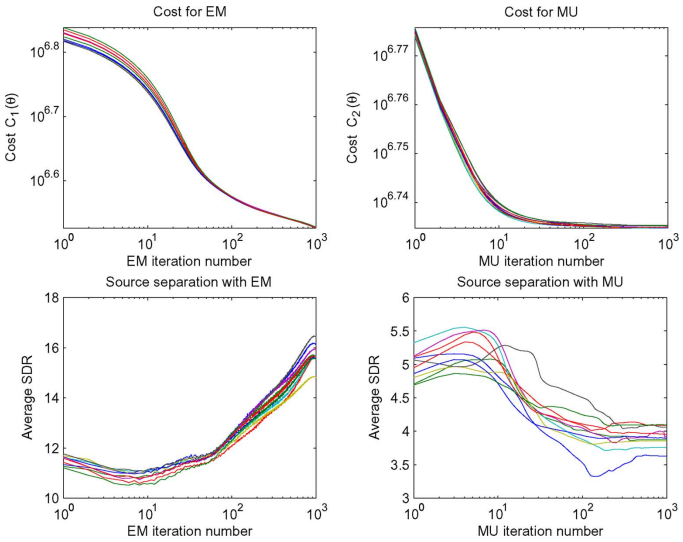


Fig. 3. Ten runs of EM and MU from ten perturbed oracle initializations using instantaneous mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

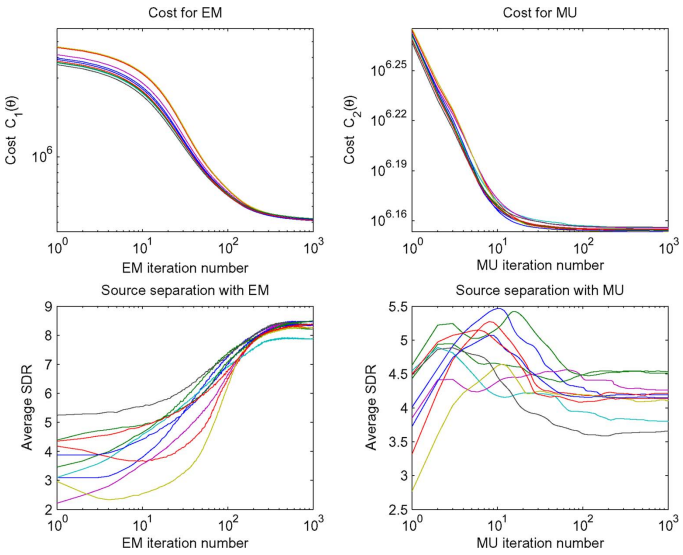


Fig. 4. Ten runs of EM and MU from ten perturbed oracle initializations using convolutive mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

top row display in log-scale the cost functions $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$ w.r.t. iterations for all ten runs. Note that cost $C_1(\boldsymbol{\theta})$ is not positive in general, see (12), so that we have added a common large constant value to all curves so as to ensure positivity, and to be able plotting cost value in the logarithmic scale. Plots on bottom row display the average SDRs.

The results show that maximization of the joint likelihood with the EM algorithm leads to consistent improvement of source separation performance in term of SDR, in the sense that final average SDR values are higher than values at initialization. This is not the case with MU, which results in nearly every case in worsening the SDR values obtained from oracle initialization. This is undoubtedly a consequence of discarding mutual information between the channels.

As for computational loads, our MATLAB implementation of EM (resp. MU) algorithm takes about 80 min (resp. 20 min) per 1000 iterations, for this particular experiment with 17-s stereo mixture (sampled at 16 kHz), $J = 3$ sources, and $K = 12$ components.

F. Blind Separation of Under-Determined Speech and Music Mixtures

In this section, we compare our algorithms with the methods that achieved competitive results at the SASSEC'07 evaluation campaign for the tasks of underdetermined mixtures of respectively speech and music signals, in both instantaneous and convolutive cases. We used the same data and evaluation criteria as in the campaign. More precisely, our algorithms are compared in the instantaneous case to the method of Vincent [40], based on source STFT reconstruction using a minimum l_0 norm constraint given a mixing matrix estimate obtained with the method of Arberet *et al.* [41]. In the convolutive case, our algorithms are compared to the method of Sawada, based on frequency-dependent complex-valued mixing matrices estimation [42], and *a posteriori* grouping relying on temporal correlations between sources in different frequency bins [20]. We used the outputs of these methods to initialize our own algorithms. In the linear instantaneous case, we were given MATLAB implementations of [40] and [41]. In the convolutive case, we simply downloaded the source image estimates from the SASSEC'07 web page [43]. In both cases we built initializations of $\mathbf{W}$ and $\mathbf{H}$ based on NMF of the source spectrogram estimates.[3]

We have found satisfactory separation results through trials using $\#\mathcal{K}_j = 4$ components for musical sources and $\#\mathcal{K}_j = 10$ components for speech sources. More components seem to be needed for speech so as to account for its higher variability (e.g., vibrato). The EM and MU algorithms were run for 500 iterations, final source separation SDR results together with reference methods results are displayed in Table II.[4] The EM method yields a significant separation improvement for all linear instantaneous mixtures. Improvement is also obtained in the convolutive case for most source estimates, but is less significant in terms of SDRs. However, and maybe most importantly, we believe our source estimates to be generally more pleasant to listen to. Indeed, one drawback of sparsity-based, nonlinear source reconstruction is musical noise, originating from unnatural, isolated time-frequency atoms scattered over the time–frequency plane. In contrast, our Wiener source estimates, obtained as a linear combination of data in each TF cell, appear to be less prone to such artifacts as can be listened to at demo web page [35]. We have entered our EM algorithm to the "under-determined speech and music mixtures" task of SiSEC'08 for instantaneous mixtures, and our results can be compared to other

[3]However, in that case we used KL-NMF instead of IS-NMF, not to fit the lower-energy residual artifacts and interferences, to which IS-NMF might be overly sensitive as a consequence of its scale-invariance. This seemed to lead to better initializations indeed.

[4]The reference algorithms performances in Table II do not always coincide with those given on the SASSEC'07 web page [43]. In the instantaneous case, this is because we have not used the exact same implementation of the l_0 minimization algorithm [40] that was used for SASSEC. In the convolutive case, this is because we have removed the dc component from all speech signals (including reference, source image estimates, and mixtures) using high-pass filtering, in order to avoid numerical instabilities.

TABLE II
SOURCE SEPARATION RESULTS FOR SASSEC DATA IN TERMS OF SDR (dB)

	female4				male4				nodrums			wdrums			average
	s1	s2	s3	s4	s1	s2	s3	s4	s1	s2	s3	s1	s2	s3	
Linear instantaneous mixtures															
l_0 min.	12.6	6.1	4.7	7.3	15.6	2.7	5.3	6.9	21.2	1.7	15.8	-0.5	3.1	28.4	9.6
EM	14.2	7.8	5.9	8.6	16.8	3.5	8.2	9.6	27.1	7.6	21.4	0.9	4.6	29.8	12.3
MU	3.9	0.9	0.1	2.2	8.6	-0.7	2.8	2.9	8.8	-6.4	3.3	10.0	2.9	19.3	4.4
Synthetic convolutive mixtures (1m)															
Sawada	5.2	5.3	3.2	2.6	4.5	0.6	4.9	2.3	3.0	1.0	-1.6	4.4	-12.7	0.6	1.3
EM	7.7	6.4	4.1	3.2	6.2	0.4	5.5	2.7	4.1	1.0	-1.8	3.9	-12.4	1.3	1.9
MU	5.2	3.3	2.7	1.4	3.4	-0.9	3.0	1.7	2.8	1.0	-2.0	5.9	-10.9	1.9	1.1
Live-recorded convolutive mixtures (1m)															
Sawada	4.1	3.8	6.0	3.3	3.0	1.6	4.8	2.4	4.1	5.1	-3.8	4.1	4.5	6.0	3.5
EM	5.3	3.6	7.2	4.3	3.5	2.1	5.6	3.1	4.5	7.3	-4.5	4.9	5.5	8.0	4.3
MU	1.6	-0.2	4.3	1.8	1.1	0.0	2.8	2.1	3.9	3.6	-4.9	4.1	4.5	7.5	2.4

methods in [44], and online at [45]. Note that among the ten algorithms participating in this task our algorithm outperformed all the other competiting methods by at least 1 dB for all separation measures (SDR, ISR, SIR, and SAR), see [44, Table 2].

G. Supervised Separation of Professionally Produced Music Recordings

We here apply our algorithms to the separation of the professionally produced music recordings of dataset B. This is a supervised setting in the sense that training data is available to learn the source spectral patterns $\mathbf{W}$ and filters. The following procedure is used.

- Learn mixing parameters $\{a^{tr}_{ij,f}\}_{if}$, spectral patterns $\mathbf{W}^{tr}_j$, and activation coefficients $\mathbf{H}^{tr}_j$ from available training signal images of source j (using 200 iterations of EM/MU); discard $\mathbf{H}^{tr}_j$.
- Clamp $\mathbf{A}$ and $\mathbf{W}$ to their trained values $\mathbf{A}^{tr}$ and $\mathbf{W}^{tr}$ and reestimate activation coefficients $\mathbf{H}$ from test data $\mathbf{X}$ (using 200 iterations of EM/MU).
- Reconstruct source image estimates from $\mathbf{A}^{tr}$, $\mathbf{W}^{tr}$ and $\mathbf{H}$.

Except for the training of mixing coefficient, the procedure is similar in spirit to supervised single-channel separation schemes proposed, e.g., in [9] and [46].

One important issue with professionally produced modern music mixtures is that they do not always comply with the mixing assumptions of (3). This might be due to nonlinear sound effects (e.g., dynamic range compression), to reverberation times longer than the analysis window length, and maybe most importantly to when the *point source* assumption does not hold anymore, i.e., when the channels of a stereo instrumental track cannot be represented as a convolution of the *same* source signal. The latter situation might happen when a sufficiently voluminous musical instrument (e.g., piano, drums, acoustic guitar) is recorded with several microphones placed close to the instrument. As such, the guitar track of the "Que pena tanto faz" song from dataset C is a non-point source image. Such tracks may be modeled as a sum of several point sources, with different mixing filters.

For the "Que pena tanto faz" song, the vocal part is modeled as an instantaneously mixed point source image with $\#\mathcal{K}_1 = 8$ components while the guitar part is modeled as a sum of three convolutively mixed point source images, each modeled with $\#\mathcal{K}_2 = \#\mathcal{K}_3 = \#\mathcal{K}_4 = 3$ components. For the "Roads" song, the bass and vocals parts are each modeled as instantaneously mixed point source images with six components, the piano part is modeled as a convolutive point source image with six components and finally, the residual background music (sum of remaining tracks) is modeled as a sum of three convolutive point source images with four components. The audio results, available at [35], tend to show better performance of the EM approach, especially on the "Roads" song. Our results can be compared to those of the other methods that entered the "professionally produced music recordings" task of SiSEC'08 in [44], and online at [47].

H. Blind Separation of Professionally Produced Music Recordings

In the last experiment, we have tested the EM and MU algorithms for the separation of professionally produced music recordings (commercial CD excerpts) in a fully unsupervised (blind) setting. We used the following parameter initialization procedure, inspired from [48], which yielded satisfactory results.

- Stack left and right mixture STFTs so as to create a $2F \times N$ complex-valued matrix $\mathbf{X}_{2\mathrm{ch}} = [\mathbf{X}_L^T\ \mathbf{X}_R^T]^T$.
- Produce a K-components IS-NMF decomposition of $|\mathbf{X}_{2\mathrm{ch}}|^2 \approx \mathbf{W}_{2\mathrm{ch}}\mathbf{H}_{2\mathrm{ch}}$.
- Initialize $\mathbf{W}$ as the average of $\mathbf{W}_L$ and $\mathbf{W}_R$, where $\mathbf{W}_{2\mathrm{ch}} = [\mathbf{W}_L^T\ \mathbf{W}_R^T]^T$. Initialize $\mathbf{H} = \mathbf{H}_{2\mathrm{ch}}$.
- Reconstruct K components $\widehat{\mathbf{C}}_{2\mathrm{ch},k} = [\widehat{\mathbf{C}}_{L,k}^T\ \widehat{\mathbf{C}}_{R,k}^T]^T$ from $\mathbf{X}_{2\mathrm{ch}}$, $\mathbf{W}_{2\mathrm{ch}}$, and $\mathbf{H}_{2\mathrm{ch}}$, using single-channel Wiener filtering (see, e.g., [10]). Produce K ad-hoc left and right component-dependent mixing filters estimates by averaging $\widehat{\mathbf{C}}_{L,k}/\Phi$ and $\widehat{\mathbf{C}}_{R,k}/\Phi$ over frames, with $\Phi = \arg(\widehat{\mathbf{C}}_{L,k})$, and normalizing according to Section III-A2. Cluster the resulting filter estimates with the K-means algorithm, whose output can be used to

define the partition $\{\mathcal{K}_j\}_{j=1}^{J}$ (using cluster indices) and a mixing system estimate $\hat{\mathbf{A}}$ (using cluster centroids).

Depending on the recording we set the number of sources J to 3 or 4 and used a total of $K = 15$ to 20 components. The EM and MU algorithms were run for 300 iterations in every case. On these specific examples the superiority of the EM method w.r.t. the MU method is not as clear as with previous datasets. A likely reason is the existence of nonpoint sources breaking the validity of mixing assumptions (2). In such precise cases, choosing not to exploit inter-channel dependencies might be better, because our model of these dependencies is now wrong. Looking for suitable probabilistic models of nonpoint sources is a new and interesting research direction.

In some cases the source image estimates contain several musical instruments and some musical instruments are spread over several source images. Besides poor initialization, this can be explained by 1) sources mixed similarly (e.g, same directions of arrival), and thus impossible to separate in our fully blind setting, 2) nonpoint sources, not well represented by our model and thus split into different source image estimates.

One way to possibly refine separation results is to reconstruct individual stereo component images (i.e., obtained via Wiener filtering (20) in case of EM method, or via (42) by replacing $p_{i,fn}$ with $w_{fk}h_{kn}$ in case of MU method), and manually group them through listening, either to separate sources mixed similarly, or to reconstruct multidirectional sound sources that better match our understanding/perception of a single source.

Finally, to show the potential of our source separation approach for music remixing, we have created some remixes using the blindly separated source images and/or the manually regrouped ones. The remixes were created in Audacity [49] by simply re-panning the source image estimates between left and right channels and by changing their gains. The audio results can be listened to at [35].

V. Conclusion

We have presented a general probabilistic framework for the representation of multichannel audio, under possibly underdetermined and noisy convolutive mixing assumptions. We have introduced two inference methods: an EM algorithm for the maximization of the channels joint log-likelihood and a MU algorithm for the maximization of the sum of individual channel log-likelihoods. The complexity of these algorithms grows linearly with the number of model components, and make them thus suitable to real-world audio mixtures with any number of sources. The corresponding CPU computational loads are in the order of a few hours for a song, which may be considered reasonable for applications such as remixing, where real-time is not an issue.

We have applied our decomposition algorithms to stereo source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings.

The EM algorithm was shown to outperform state-of-the-art methods, given appropriate initializations. Both our methods have indeed been found sensitive to parameter initialization, but we have come up with two satisfying initialization schemes. The first one, described in Section IV-F, consists in using the output of a different separation algorithm. We show that our EM algorithm improves the separation results in almost all cases. The second scheme, described in Section IV-H, consists in a single-channel NMF decomposition followed by K-means filters clustering. Our experiments tend to show that the NMF model is more suitable to music than speech: music sources can be represented by a small number of components to attain good separation performance, and informal listening indicates better separation of music signals.

Given that the mixed signals follow the mixing and point source assumptions inherent to (2), the EM method gives better separation results than the MU method, because between-channel dependencies are optimally exploited. However, the performance of the EM method may significantly drop when these assumptions are not verified. In contrast, we have observed that the MU method, which relies on a weaker model of between-channel dependencies, yields more even results overall and higher robustness to model discrepancies (that may for example occur in professionally produced recordings).

Let us now mention some further research directions. Algorithms faster than EM (both in terms of convergence rate and CPU time per iteration) would be desirable for optimization of the joint likelihood (12). As such, we envisage turning to Newton gradient optimization, as inspired from [50]. Mixed strategies could also be considered, consisting of employing EM in the first few iterations to get a sharp decrease of the likelihood before switching to faster gradient search once in the neighborhood of a solution.

Bayesian extensions of our algorithm are readily available, using for example priors favoring sparse activation coefficients h_k, or even sparse filters $q_{ij,f}$ like in [51]. Minor changes are required in the MU rules so as to yield algorithms for maximum *a posteriori* (MAP) estimation. More complex priors structure can also be envisaged within the EM method, such as Markov chains favoring smoothness of the activation coefficients $\mathbf{H}$ [10].

An important perspective is automatic order selection. In our case, that concerns the total number of components K, the number of sources J and the partition $\{\mathcal{K}_j\}_j$. Regarding the total number of components K, ideas from *automatic relevance determination* can be explored, see [52] in a NMF setting. Then the problem of partitioning can be viewed as a clustering problem with unknown number of clusters J, which is a typical machine learning problem.

While we have assessed the validity of our model in terms of source separation, our decompositions more generally provide a data-driven object-based representation of multichannel audio that could be relevant to other problems such as audio transcription, indexing and object-based coding. As such, it will be interesting to investigate the semantics revealed by the learnt spectral patterns $\mathbf{W}$ and activation coefficients $\mathbf{H}$.

Finally, as discussed in Section IV-H, new models should be considered for professionally produced music recordings, dealing with nonpoint sources, nonlinear sound effects, such as dynamic range compression, and long reverberation times.

APPENDIX A

APPENDIX A
EM ALGORITHM DERIVATION OUTLINE

The complete data minus log-likelihood can be written as

$$\begin{aligned}
&-\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) \\
&\quad = -\log p(\mathbf{X}|\mathbf{C},\boldsymbol{\theta}) - \log p(\mathbf{C}|\boldsymbol{\theta}) \\
&\quad \stackrel{c}{=} \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})\right] \\
&\qquad + \sum_k\sum_{fn}\left[\log(h_{k,n}w_{k,f}) + \frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\quad = \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + \sum_k\log(h_{k,n}w_{k,f}) + \sum_k\frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\qquad + N\sum_f \operatorname{trace}\Big[\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&\qquad\qquad - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\Big]
\end{aligned} \tag{43}$$

with $\mathbf{R}_{\mathbf{xx},f}$, $\mathbf{R}_{\mathbf{xs},f}$, $\mathbf{R}_{\mathbf{ss},f}$, and $u_{k,fn}$ defined by (13) and (14). Thus, we have shown that the complete data log-likelihood can be represented in the following form:

$$\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) = \langle\eta(\boldsymbol{\theta}),\mathbf{T}(\mathbf{X},\mathbf{C})\rangle + \nu(\boldsymbol{\theta}) \tag{44}$$

where $\mathbf{T}(\mathbf{X},\mathbf{C})$ is a vector of all scalar elements of $\mathbf{t}(\mathbf{X},\mathbf{C}) \stackrel{\Delta}{=} \{\mathbf{R}_{\mathbf{xx},f},\mathbf{R}_{\mathbf{xs},f},\mathbf{R}_{\mathbf{ss},f},\{u_{k,fn}\}_{kn}\}_f$, and $\eta(\boldsymbol{\theta})$ and $\nu(\boldsymbol{\theta})$ are some vector and scalar functions of parameters. That means that the complete data pdfs $\{p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11], [29]) and complete data statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$ is a *natural (sufficient) statistics* [11], [29] for this family. To derive an EM algorithm in this special case one needs to 1) solve complete data ML criterion (thanks to (44) this solution can be always expressed as a function of natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$), and 2) replace in this solution $\mathbf{t}(\mathbf{X},\mathbf{C})$ by its conditional expectation $\hat{\mathbf{t}}(\mathbf{X},\boldsymbol{\theta}') \stackrel{\Delta}{=} \int\mathbf{t}(\mathbf{X},\mathbf{C})p(\mathbf{C}|\mathbf{X},\boldsymbol{\theta}')d\mathbf{C}$ using model $\boldsymbol{\theta}'$ estimated at the previous step of EM.

To solve the complete data ML criterion, we first compute the derivatives of $\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})$ (43) w.r.t. model parameters $\boldsymbol{\theta}$ (see [53] for issues regarding derivation w.r.t. complex-valued parameters), set them to zero and solve the corresponding equations (subject to the constraint that $\boldsymbol{\Sigma}_{\mathbf{b},f}$ is diagonal), and we have:[5]

$$\mathbf{A}_f = \mathbf{R}_{\mathbf{xs},f}\mathbf{R}_{\mathbf{ss},f}^{-1} \tag{45}$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{b},f} = \operatorname{diag}\big(&\mathbf{R}_{\mathbf{xx},f} - \mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&- \mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\big)
\end{aligned} \tag{46}$$

$$w_{fk} = \frac{1}{N}\sum_n\frac{u_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f\frac{u_{k,fn}}{w_{fk}}. \tag{47}$$

[5]Bayesian MAP estimation can be carried out instead of ML by simply adding a prior term $-\log p(\boldsymbol{\theta})$ to the right part of (43) and solving the corresponding complete data MAP criterion.

Our EM algorithm is strictly speaking only a *Generalized* EM algorithm [54] because it only ensures $Q(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m|\boldsymbol{\theta}^m)$. Indeed, in (47) $\mathbf{W}$ is still a function of $\mathbf{H}$, and reversely, $\mathbf{H}$ is a function of $\mathbf{W}$.

To finish derivation of our EM algorithm we need to compute conditional expectation of the natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$. It can be shown that given $\mathbf{x}_{fn}$ the source vector $\mathbf{s}_{fn}$ is a proper Gaussian random vector, i.e.,

$$p(\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}) = N_c\left(\mathbf{s}_{fn};\hat{\mathbf{s}}_{fn},\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}}\right) \tag{48}$$

with mean vector $\hat{\mathbf{s}}_{fn}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}}$ as follows:

$$\begin{aligned}
\hat{\mathbf{s}}_{fn} &= \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{x}_{fn}, \\
\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\text{post}} &= \boldsymbol{\Sigma}_{\mathbf{s},f} - \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}.
\end{aligned}$$

Computing conditional expectations of $\mathbf{R}_{\mathbf{xs},f}$ and $\mathbf{R}_{\mathbf{ss},f}$ using (48) leads to (16) and (17) of EM Algorithm 1. Very similar derivations can be done to compute the conditional expectations of $u_{k,fn}$. To that matter, one only needs to compute the posterior distribution of $\mathbf{c}_{fn}$ instead of $\mathbf{s}_{fn}$, using mixing equation (10) instead of mixing equation (3).

APPENDIX B
MU ALGORITHM DERIVATION OUTLINE

Let θ be a scalar parameter of the set $\{\mathbf{Q},\mathbf{W},\mathbf{H}\}$. The derivative of cost $C_2(\boldsymbol{\theta})$, given by (30), w.r.t. θ simply writes

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ifn}(\nabla_\theta\hat{v}_{i,fn})d'_{IS}(v_{i,fn}|\hat{v}_{i,fn}) \tag{49}$$

where $d'_{IS}(x|y)$ is the derivative of $d_{IS}(x|y)$ w.r.t. y given by

$$d'_{IS}(x|y) = \frac{1}{y} - \frac{x}{y^2}. \tag{50}$$

Using (49), we obtain the following derivatives:

$$\begin{aligned}
\nabla_{q_{ij,f}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{n=1}^N p_{j,fn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{w_{jfk}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{n=1}^N q_{ij,f}h_{j,kn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{h_{jkn}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{f=1}^F q_{ij,f}w_{j,fk}d'(v_{i,fn}|\hat{v}_{i,fn})
\end{aligned}$$

which can be written in the following matrix forms:

$$\begin{aligned}
\nabla_{\mathbf{q}_{ij}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \left(\hat{\mathbf{V}}_i^{\cdot-1}\mathbf{P}_j - \hat{\mathbf{V}}_i^{\cdot-2}.\mathbf{V}_i.\mathbf{P}_j\right)\mathbf{1}_{N\times1} \\
\nabla_{\mathbf{W}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\operatorname{diag}(\mathbf{q}_{ij})\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right)\mathbf{H}_j^T \\
\nabla_{\mathbf{H}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I(\operatorname{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right).
\end{aligned}$$

Hence, the update rules given in Algorithm 2, following the multiplicative update strategy described in Section III-B3.

ACKNOWLEDGMENT

The authors would like to thank S. Arberet for kindly sharing his implementation of DEMIX algorithm [41], all the organizers of SiSEC'08 for well-prepared evaluation campaign, as well as the anonymous reviewers for their valuable comments.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2003, pp. 177–180.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, 2007, pp. 65–68.

[4] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[6] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, Mar. 2006, pp. 666–673.

[7] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals Syst. Conf.*, Dublin, Ireland, Sep. 2005, pp. 8–12.

[8] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[9] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, Hong Kong, 2003, pp. 613–616.

[10] 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.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[12] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'97)*, Apr. 1997, pp. 3617–3620.

[13] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, pp. 803–851, 1999.

[14] H. Attias, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, pp. 297–300.

[15] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," in *Proc. Interspeech'08*, 2008, pp. 419–422.

[16] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'09)*, 2009, pp. 751–758.

[17] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. 11th Eur. Signal Process. Conf. (EUSIPCO'02)*, 2002, pp. 561–564.

[18] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, Mohonk, NY, Oct. 2005, pp. 78–81.

[19] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'97)*, New Paltz, NY, Oct. 1997, 4 pp..

[20] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE Int. Symp. Circuits Syst. (ISCAS'07)*, May 27–30, 2007, pp. 3247–3250.

[21] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Inf. Process. Syst. (NIPS 19)*, 2007, pp. 953–960.

[22] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, Oct. 2007, pp. 147–150.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, Apr. 2009, pp. 3137–3140.

[24] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[25] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Symp. Music Inf. Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[26] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, 2006, pp. 32–39.

[27] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[28] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Comput. Intell. Neurosci.*, vol. 2008, pp. 1–9, 2008.

[29] 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 Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[31] M. Goodwin, "The STFT, sinusoidal models, and speech modification," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York: Springer, 2008, ch. 12, pp. 229–258.

[32] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

[33] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recognition Lett.*, vol. 22, no. 12, pp. 1255–1261, 2001.

[34] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 792–799, ACM.

[35] Example Web Page [Online]. Available: http://www.irisa.fr/metiss/ozerov/demos.html#ieee_taslp09

[36] S. Hurley, Call for Remixes: Shannon Hurley [Online]. Available: http://www.ccmixter.org/shannon-hurley

[37] in *Signal Separation Evaluation Campaign (SiSEC 2008)*, 2008 [Online]. Available: http://www.sisec.wiki.irisa.fr

[38] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 552–559, Springer.

[39] D. Campbell, Roomsim Toolbox [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/5184

[40] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 430–437.

[41] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'06)*, 2006, pp. 536–543.

[42] P. D. O'Grady and P. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA)*, 2004, pp. 428–435.

[43] in *Stereo Audio Source Separation Evaluation Campaign (SASSEC 2007)*, 2007 [Online]. Available: http://www.sassec.gforge.inria.fr/

[44] E. Vincent, S. Araki, and P. Bofilld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'09)*, 2009, pp. 734–741 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_ICA09.pdf

[45] in *SiSEC 2008 Under-Determined Speech and Music Mixtures Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_underdetermined/

[46] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 414–421.

[47] in *SiSEC 2008 Professionally Produced Music Recordings Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_professional/

[48] S. Winter, H. Sawada, S. Araki, and S. Makino, "Hierarchical clustering applied to overcomplete BSS for convolutive mixtures," in *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Process. (SAPA 2004)*, Oct. 2004, pp. 652–660.

[49] "Audacity: The Free, Cross-Platform Sound Editor," [Online]. Available: http://www.audacity.sourceforge.net/

[50] J.-F. Cardoso and M. Martin, "A flexible component model for precision ICA," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 1–8.

[51] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 839–847, Mar. 2006.

[52] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop Signal Process. Adaptative Sparse Structured Representations (SPARS'05)*, Saint-Malo, France, Apr. 2009.

[53] A. van den Bos, "Complex gradient and Hessian," *IEE Proc. Vision, Image, Signal Process.*, vol. 141, pp. 380–382, 1994.

[54] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint-Petersburg State University, Saint-Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France, in 2006.

He worked towards the Ph.D. degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, From 1999 to 2002, he worked at Terayon Communication Systems as a R&D Software Engineer, first in Saint-Petersburg and then in Prague, Czech Republic. He was for one year (2007) in the Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) with TELECOM ParisTech / CNRS LTCI—Signal and Image Processing (TSI) Department. Currently, he is with the METISS team of IRISA/INRIA—Rennes as a Postdoctoral Researcher. His research interests include audio source separation, source coding, and automatic speech recognition.

Cédric Févotte received the State Engineering degree and the M.Sc. degree in control and computer science from École Centrale de Nantes, Nantes, France, in 2000, and the Ph.D. degree in 2003 from the University of Nantes.

From 2003 to 2006, he was a Research Associate with the Signal Processing Laboratory at the University of Cambridge, Cambridge, U.K., working on Bayesian approaches to audio signal processing tasks such as audio source separation, denoising, and feature extraction. From May 2006 to February 2007, he was a Research Engineer with the start-up company Mist-Technologies (Paris), working on mono/stereo to 5.1 surround sound upmix solutions. In March 2007, he joined CNRS LTCI/Telecom ParisTech, first as a Research Associate and then as a CNRS tenured Research Scientist in November 2007. His research interests generally concern statistical signal processing and unsupervised machine learning with audio applications.

Paper 2 (Ozerov, Vincent & Bimbot, *IEEE TASLP*, 2011)

A General Flexible Framework for the Handling of Prior Information in Audio Source Separation

Alexey Ozerov, *Member, IEEE,* Emmanuel Vincent, *Senior Member, IEEE,* and Frédéric Bimbot

***Abstract*—Most of audio source separation methods are developed for a particular scenario characterized by the number of sources and channels and the characteristics of the sources and the mixing process. In this paper we introduce a general audio source separation framework based on a library of structured source models that enable the incorporation of prior knowledge about each source via user-specifiable constraints. While this framework generalizes several existing audio source separation methods, it also allows to imagine and implement new efficient methods that were not yet reported in the literature. We first introduce the framework by describing the model structure and constraints, explaining its generality, and summarizing its algorithmic implementation using a generalized expectation-maximization algorithm. Finally, we illustrate the above-mentioned capabilities of the framework by applying it in several new and existing configurations to different source separation problems. We have released a software tool named** *Flexible Audio Source Separation Toolbox (FASST)* **implementing a baseline version of the framework in Matlab.**

***Index Terms*—Audio source separation, local Gaussian model, nonnegative matrix factorization, expectation-maximization**

I. Introduction

Separating audio sources from multichannel mixtures is still challenging in most situations. The main difficulty is that audio source separation problems are usually mathematically ill-posed and to succeed one needs to incorporate additional knowledge about the mixing process and/or the source signals. Thus, efficient source separation methods are usually developed for a particular source separation problem characterized by a certain *problem dimensionality*, e.g., determined or underdetermined, certain *mixing process characteristics*, e.g., instantaneous or convolutive, and certain *source characteristics*, e.g., speech, singing voice, drums, bass or noise [1]. For example, a source separation problem may be formulated as follows:

> *"Separate bass, drums, melody and the remaining instruments from a stereo professionally produced music recording."*

Given a source separation problem, one typically must introduce as much knowledge about this problem as possible into the corresponding separation method so as to achieve good separation performance. However, there is often no common formulation describing methods applied for different problems, and this makes it difficult to reuse a method for a problem it was not originally conceived for. Thus, given a new source separation problem, the common approach consists in (i) model design, taking into account problem formulation, (ii) algorithm design and (iii) implementation (see Fig. 1, top).

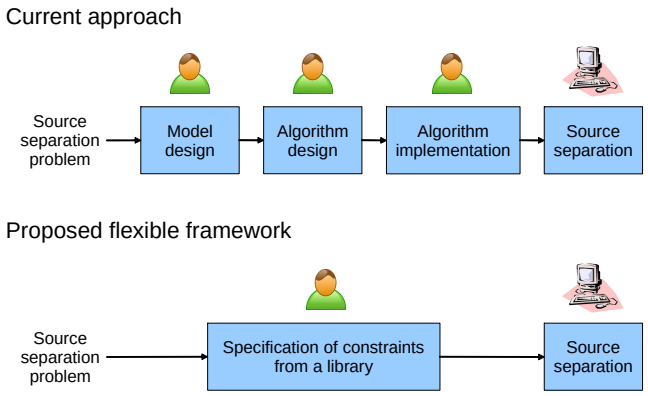


Fig. 1. Current way of addressing a new source separation problem (top) and the way of addressing it using the proposed flexible framework (bottom).

The motivation of this work is to improve over this time-consuming process by designing a general audio source separation framework that can be applied to virtually any source separation problem by simply selecting from a library of constraints suitable constraints accounting for the available information about that source (see Fig. 1, bottom). More precisely, we wish such a framework to be

- *general*, i.e., generalizing existing methods and making it possible to combine them,
- *flexible*, allowing easy incorporation of the *a priori* knowledge about a particular problem considered.

To achieve the property of generality, we need to find some common formulation for methods we would like to generalize. Many recently proposed methods for audio source separation and/or characterization [2]–[19] (see also [1] and references therein) are based on the same so-called *local Gaussian model* describing both the properties of the sources and of the mixing process. Thus, we chose this model as the core of our framework. To achieve flexibility, we fix the global structure of Gaussian covariances, and by means of a parametric model allow the introduction of knowledge about each individual source and its mixing characteristics via constraints on individual parameter subsets. The global

A. Ozerov and E. Vincent are with INRIA, Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes cedex, France (e-mails: alexey.ozerov@inria.fr, emmanuel.vincent@inria.fr).

F. Bimbot is with IRISA, CNRS - UMR 6074, Campus de Beaulieu, 35042 Rennes cedex, France (e-mail: frederic.bimbot@irisa.fr).

This work was partly supported by OSEO, the French State agency for innovation, under the Quaero program, and by the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research and the German Academic Exchange Service under project Procope 20142UD.

structure we consider corresponds to a generative model of the data that is motivated by the physics of the modeled processes, e.g., the source-filter model to represent a sound source and an approximation of the convolutive filter to represent its mixing characteristics. In summary, our framework generalizes the methods from [2]–[19], and, thanks to its flexibility, it becomes applicable in many other scenarios one can imagine.

We implement our framework using a generalized expectation-maximization (GEM) algorithm [20], where the M-step is solved by alternating optimization of different parameter subsets, taking the corresponding constraints into account and using multiplicative update (MU) rules inspired from the nonnegative matrix factorization (NMF) methodology (see, e.g., [9]) to update the nonnegative spectral parameters. Such an implementation is in fact possible thanks to the Gaussianity assumption leading to closed form update equations. The idea of mixing GEM algorithm with MU rules was already reported in [21] in the case of plain NMF spectral models and rank-1 spatial models, and we extend it here to the newly proposed structures. Our algorithmic contribution consists of (i) identifying the *GEM-MU* approach as suitable thanks to the implementability of the configurable framework, the simplicity of the update rules, the implicit verification of nonnegative constraints and its good convergence speed; and (ii) deriving of the update rules for the new model structures.

Our approach is in line with the *library of components* by Cardoso *et al* [22] developed for the separation of components in astrophysical images. However, we consider advanced audio-specific structures inspired by [1], [23] for source spectral power, as opposed to the unique block structure in [22] based on the assumption that source power is constant in some pre-defined region of time and space. In that sense, our framework is more flexible than [22]. Besides the framework itself, we propose a new structure for NMF-like decompositions of source power spectrograms, where the temporal envelope associated with each spectral pattern is represented as a nonnegative linear combination of time-localized temporal patterns. This structure can be used to ensure temporal continuity, but also to model more complex temporal characteristics, such as the attack or decay parts of a note. In line with time-localized patterns we include in our framework the so-called narrowband spectral patterns that allow constraining spectral patterns to be harmonic, inharmonic or noise-like. These structures were already reported in [14], [15], but only in case of harmonic constraints. Moreover, they were not applied for source separation so far. As compared to [24], where some preliminary aspects of this work were presented, we here present the framework in details, describe its implementation, and extend the experimental part illustrating the framework. Moreover, we propose an original mixing model formulation that allows the representation and the estimation of rank-1 [5] and full-rank [19] (actually any rank) spatial mixing models in a homogeneous way, thus enabling the combination of both models within a given mixture. Finally, we provide a proper probabilistic formulation of local Gaussian modeling for quadratic time-frequency representations [18] that supports and justifies the formulation given in [18].

We have also implemented and released a baseline version of the framework in Matlab. The corresponding software tool named *Flexible Audio Source Separation Toolbox (FASST)* is available at [25] together with a user guide, examples of usage (where the constraints are specified) and the corresponding audio examples. Given a source separation problem, one can choose one or few suitable constraint combinations based on his/her expertise and on the a priori knowledge, and then test all of them using FASST so as to select the best one.

In summary, the main contributions of this work include

- a general modeling structure,
- a general estimation algorithm,
- new spectral an temporal structures (time-localized patterns, narrowband spectral patterns),
- the implementation and distribution of a baseline version of the framework (the FASST toolbox [25]).

The rest of this paper is organized as follows. In Section II, existing approaches generalized by the proposed framework are discussed and an overview of the framework is given. Sections III and IV provide a detailed description of the framework and its algorithmic implementation. Thus, Section II is devoted to a reader interested in understanding the main principles of the framework and the physical meaning of the objects, and Sections III and IV to one willing to go deeper into the technical details. The results of a few source separation experiments are given in Section V to illustrate the flexibility of our framework and its potential performance improvement compared to individual approaches. Conclusions are drawn in Section VI.

II. Related existing approaches and framework overview

Source separation methods based on the local Gaussian model can be characterized by the following assumptions [1], [2], [5], [13], [19]:

1) *Gaussianity:* in some time-frequency (TF) representation the sources are modeled in each TF bin by zero-mean Gaussian random variables.
2) *Independence:* conditionally to their covariance matrices, these random variables are independent over time, frequency and between sources.
3) *Factorization of spectral and spatial characteristics:* for each TF bin, the covariance matrix of each source is expressed as the product of a *spatial covariance* matrix representing its spatial characteristics and a scalar *spectral power* representing its spectral characteristics.
4) *Linearity of mixing:* the mixing process translates into addition in the covariance domain.

A. State-of-the-art approaches based on the local Gaussian model

The state-of-the-art approaches [2]–[19] cover a wide range of source separation problems and models expressed via particular structures of local Gaussian covariances, including:

1) *Problem dimensionality:* Denoting by I and J, respectively, the number of channels of the observed

mixture and the number of sources to separate, the *single-channel* ($I = 1$) case is addressed in [6], and *underdetermined* ($1 < I < J$) and *(over-)determined* ($I \geq J$) cases are addressed in [5] and [2], respectively.

2) *Spatial covariance model: Instantaneous* and *convolutive* mixtures of point sources are modeled by *rank-1* spatial covariance matrices in [5] and [3], respectively. In [19] reverberant convolutive mixtures of point sources are modeled by *full-rank* spatial covariance matrices that, in contrast to rank-1 covariance matrices, can account for the spatial spread of each source induced by the reverberation.
3) *Spectral power model:* Several models were proposed for the spectral power, e.g., *unconstrained* models [10], *block constant* models [5], Gaussian mixture models (GMM) or hidden Markov models (HMM) [2], Gaussian scaled mixture models (GSMM) or scaled HMMs (S-HMM) [13], NMF [4] together with its variants, harmonic NMF [14] or temporal activation constrained NMF [9], and source-filter models [16]. These models are suitable for the representation of different types of sources, for example GSMM is rather suitable for a monophonic source, e.g., speech, and NMF for a polyphonic one, e.g., polyphonic musical instrument, [13].
4) *Input representation:* While the most of the considered methods use the short time Fourier transform (STFT) as the input TF representation, some of them, e.g., [14], [15], [18], use the auditory-motivated equivalent rectangular bandwidth (ERB) quadratic representation. More generally, we consider here both *linear representations*, where the signal is represented by a vector of complex-valued coefficients in each TF bin, as well as *quadratic representations*, where the signal is represented via its local covariance matrix in each TF bin [26].

Table I provides an overview of some of the local Gaussian model-based approaches considered here, where the specificities of each method are marked by crosses ×. We see from Table I that a few of these methods have already been combined together, for example GSMM and NMF were combined in [8], and NMF [9] was combined with rank-1 and full-rank mixing models in [13] and [17], respectively. However, many combinations have not yet been investigated. Indeed, assuming that each source follows one of the 3 spatial covariance models and one of the 8 spectral variance models from Table I, the total number of configurations equals to 2×24^J for J sources (in fact much more since each source can follow several spectral variance models at the same time), while Table I reports only 16 existing configurations.

B. Other related state-of-the-art approaches

While the local Gaussian model-based framework offers maximum of flexibility, there exist some methods that do not satisfy (fully or partially) the aforementioned assumptions and are thus not strictly covered by the framework. Nevertheless, our framework allows the implementation of similar structures. Let us give some examples. Binary masking-based source estimation [27], [28] does not satisfy the source independence assumption. However, it is known to perform poorly compared to local Gaussian model-based separation, as it was shown in [13], [18] for convolutive mixtures [1] and demonstrated through the signal separation evaluation campaigns SiSEC 2008 [30] and SiSEC 2010 [29], where for instantaneous mixtures local Gaussian model-based approaches gave better results than the *oracle* (using the ground truth) binary masks. The methods proposed in [31], [32] are also based on Gaussian models albeit in the time domain. Notably, time sample-based GMMs and time-varying autoregressive models are considered as source models in [31] and [32], respectively. However, the number of existing time-domain structures is fairly reduced. Our TF domain models make it possible to account for these structures by means of suitable constraints over spectral power, while allowing their combination with more advanced structures. There are also many works on NMF and its extensions [33]–[38] and on GMMs / HMMs [39], [40] based on nongaussian models of the complex-valued STFT coefficients. These models are essentially covered by our framework in the sense that we can implement similar or equivalent model structures, albeit under Gaussian assumptions. The benefit of local Gaussian modeling is that it naturally leads to closed-form expressions in the multichannel case and allows the modeling of diffuse sources [19], contrary to the models in [33]–[40]. Finally, according to Cardoso [41], nongaussianity and nonstationarity are alternative routes to source separation, such that nonstationary nongaussian models would offer little benefit compared to nonstationary Gaussian models in terms of separation performance despite considerably greater computation cost.

C. Framework overview

We now present an overview of the proposed framework focusing on the most important concepts. An exhaustive description is given in Sections III and IV.

The framework is based on a flexible model described by parameters $\theta = \{\theta_j\}_{j=1}^J$, where θ_j are the parameters of the j-th source ($j = 1, \ldots, J$). Each θ_j is split in turn into nine parameter subsets according to a fixed structure, as described below and summarized in Table II.

1) Model structure: The parameters of j-th source include a complex-valued tensor $\mathbf{A}_j$ modeling its spatial covariance, and eight nonnegative matrices $(\theta_{j,2}, \ldots, \theta_{j,9})$ modeling its spectral power over all TF bins.

The spectral power, denoted as $\mathbf{V}_j$, is assumed to be the product of an *excitation spectral power* $\mathbf{V}_j^{\mathrm{ex}}$, representing, e.g., the excitation of the glottal source for voice or the plucking of the string of a guitar, and a *filter spectral power* $\mathbf{V}_j^{\mathrm{ft}}$, representing, e.g., the vocal tract or the impedance of the guitar body [23], [35]. While such a model is usually called source-filter model, we call it here *excitation-filter model* in order to avoid possible confusions with the "sources" to be separated.

[1] Binary masking-based approaches can still be quite powerful for convolutive mixtures, as demonstrated in [29]. Thus, a good way to proceed is probably to use them to initialize local Gaussian model-based approaches, as it is done in [13], and as we do in the experimental part.

Reference		[7]	[6]	[8]	[16]	[4]	[14]	[15]	[9]	[5]	[11]	[13]	[19]	[18]	[17]	[3]	[2]
Problem dimensionality	single-channel	×	×	×	×	×	×	×									
	underdetermined									×	×	×	×	×	×		
	(over-)determined															×	×
Spatial covariance model	rank-1 instantaneous									×	×						
	rank-1 convolutive											×				×	×
	full-rank												×	×	×		
Spectral variance model	unconstrained												×	×			
	block constant									×						×	
	GMM / HMM	×									×						×
	GSMM / S-HMM		×	×	×												
	NMF			×	×	×						×			×		
	harmonic NMF						×	×									
	temp. constr. NMF							×	×								
	source-filter	×			×												
Input representation	linear	×	×	×	×	×			×	×	×	×	×		×	×	×
	quadratic						×	×						×			

TABLE I
SOME STATE-OF-THE-ART LOCAL GAUSSIAN MODEL-BASED APPROACHES FOR AUDIO SOURCE SEPARATION.

The excitation spectral power $\mathbf{V}_j^{\mathrm{ex}}$ is further decomposed as the sum of *characteristic spectral patterns* $\mathbf{E}_j^{\mathrm{ex}}$ modulated by *time activation coefficients* $\mathbf{P}_j^{\mathrm{ex}}$ [4], [9]. Each characteristic spectral pattern may be associated for instance with one specific pitch, so that the time activation coefficients denote which pitches are active on each time frame. In order to further constrain the fine structure of the spectral patterns, they are represented as linear combinations of *narrowband spectral patterns* $\mathbf{W}_j^{\mathrm{ex}}$ [14] with weights $\mathbf{U}_j^{\mathrm{ex}}$. These narrowband patterns may be for instance harmonic, inharmonic or noise-like and the weights determine the overall spectral envelope. Following the same idea, we propose here to represent the series of time activation coefficients $\mathbf{P}_j^{\mathrm{ex}}$ as sums of *time-localized patterns* $\mathbf{H}_j^{\mathrm{ex}}$ with weights $\mathbf{G}_j^{\mathrm{ex}}$. The time-localized patterns may represent the typical temporal shape of the notes while the weights encode their onset times. Different temporal fine structures such as continuity or specific rhythm patterns may also be accounted for in this way. Note that temporal models of the activation coefficients have been proposed in the state-of-the-art, using probabilistic priors [9], [34], note-specific Gaussian-shaped time-localized patterns [42], or unstructured TF patterns [33]. Our proposition is complementary to [9], [34] in that it accounts for temporal behaviour in the model structure itself in addition to possible priors on the model parameters. Moreover, it is more flexible than [9], [34], [42], since it allows the modeling of other characteristics than continuity or sparsity. Finally, while it can model similar TF patterns to [33], it involves much fewer parameters, which typically leads to more robust parameter estimation.

The filter spectral power $\mathbf{V}_j^{\mathrm{ft}}$ is similarly expressed in terms of characteristic spectral patterns $\mathbf{E}_j^{\mathrm{ft}}$ modulated by time activation coefficients [16], which are in turn decomposed into narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ft}}$ with weights $\mathbf{U}_j^{\mathrm{ft}}$ and time-localized patterns $\mathbf{H}_j^{\mathrm{ft}}$ with weights $\mathbf{G}_j^{\mathrm{ft}}$, respectively. In the case of speech or singing voice, each characteristic spectral pattern may represent the spectral formants of a given phoneme, while the plosiveness and the sequence of pronounced phonemes may be encoded by the time-localized patterns and the associated weights.

In summary, as it will be explained in details in Section III-E, the spectral power of each source obeys a three-level hierarchical nonnegative matrix decomposition structure (see equations (9), (10), (12), (13) and Figures 3 and 4 below) including at the bottom level the eight parameter subsets $\mathbf{W}_j^{\mathrm{ex}}$, $\mathbf{U}_j^{\mathrm{ex}}$, $\mathbf{G}_j^{\mathrm{ex}}$, $\mathbf{H}_j^{\mathrm{ex}}$, $\mathbf{W}_j^{\mathrm{ft}}$, $\mathbf{U}_j^{\mathrm{ft}}$, $\mathbf{G}_j^{\mathrm{ft}}$ and $\mathbf{H}_j^{\mathrm{ft}}$ (see Eq. (13)).

Parameter subsets		Size	Range
$\theta_{j,1} = \mathbf{A}_j$	mixing parameters	$I \times R_j \times F \times N$	$\in \mathbb{C}$
$\theta_{j,2} = \mathbf{W}_j^{\mathrm{ex}}$	ex. narrowband spectral patterns	$F \times L_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,3} = \mathbf{U}_j^{\mathrm{ex}}$	ex. spectral pattern weights	$L_j^{\mathrm{ex}} \times K_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,4} = \mathbf{G}_j^{\mathrm{ex}}$	ex. time pattern weights	$K_j^{\mathrm{ex}} \times M_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,5} = \mathbf{H}_j^{\mathrm{ex}}$	ex. time-localized patterns	$M_j^{\mathrm{ex}} \times N$	$\in \mathbb{R}_+$
$\theta_{j,6} = \mathbf{W}_j^{\mathrm{ft}}$	ft. narrowband spectral patterns	$F \times L_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,7} = \mathbf{U}_j^{\mathrm{ft}}$	ft. spectral pattern weights	$L_j^{\mathrm{ft}} \times K_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,8} = \mathbf{G}_j^{\mathrm{ft}}$	ft. time pattern weights	$K_j^{\mathrm{ft}} \times M_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,9} = \mathbf{H}_j^{\mathrm{ft}}$	ft. time-localized patterns	$M_j^{\mathrm{ft}} \times N$	$\in \mathbb{R}_+$

TABLE II
PARAMETER SUBSETS $\theta_{j,k}$ ($j = 1, \ldots, J$, $k = 1, \ldots, 9$) ENCODING THE STRUCTURE OF EACH SOURCE.

2) Constraints: Given the above fixed model structure, prior information about each source can now be exploited by specifying deterministic or probabilistic constraints over each parameter subset of Table II. Examples of such constraints are given in Table III. Each parameter subset can be fixed [2] (i.e., unchanged during estimation), adaptive (i.e., fully fitted to the mixture) or partially adaptive (only some parameters within the subset are adaptive). In the latter two cases, a probabilistic prior, such as a continuity prior [9] or a sparsity-inducing prior [4], can be specified over the parameters. The mixing parameters $\mathbf{A}_j$ can be time-varying or time-invariant (in Table III the latter case is only considered), frequency-dependent for convolutive mixtures or frequency-independent for instantaneous mixtures. Mixing parameters $\mathbf{A}_j$ can be given a probabilistic prior as well. E.g., it can be a Gaussian prior with the mean corresponding to the parameters of a presumed direction and with the covariance matrix representing

[2]The fixed parameters can be either set manually or learned beforehand from some training data. Learning is equivalent to model parameter estimation over the training data and can thus be achieved using our framework.

a degree of uncertainty about this direction. The rank R_j ($1 \leq R_j \leq I$) of the spatial covariance is specifiable via the size of tensor $\mathbf{A}_j$ (see Table II). Each parameter subset may also be constrained to have a limited number of nonzero entries. For instance, every column of $\mathbf{G}_j^{\text{ex}}$ and / or $\mathbf{G}_j^{\text{ft}}$ may be constrained to have a single nonzero entry accounting for a GSMM / S-HMM structure or a single nonzero entry equal to 1 accounting for a GMM / HMM structure.

Parameter subsets	Constraint	Value
$\mathbf{A}_j, \mathbf{W}_j^{\text{ex}}, \mathbf{U}_j^{\text{ex}}, \mathbf{G}_j^{\text{ex}}, \mathbf{H}_j^{\text{ex}}, \mathbf{W}_j^{\text{ft}}, \mathbf{U}_j^{\text{ft}}, \mathbf{G}_j^{\text{ft}}, \mathbf{H}_j^{\text{ft}}$	degree of adaptability	'fixed' 'part_adapt' 'adapt'
$\mathbf{A}_j$	mixing stationarity	'time_inv'
	mixing type	'conv' 'inst'
$\mathbf{G}_j^{\text{ex}}, \mathbf{G}_j^{\text{ft}}$	temporal constraint	'null' 'GMM', 'HMM' 'GSMM', 'SHMM'

TABLE III
EXAMPLES OF USER-SPECIFIABLE CONSTRAINTS OVER THE PARAMETER SUBSETS.

3) Estimation algorithm: Given the above model structure and constraints, source separation can be achieved in two steps as shown in Fig. 2. First, given initial parameter values, the model parameters θ are estimated from the mixture $\mathbf{X}$ using an iterative GEM algorithm, where the E-step consists in computing some quantity $\widehat{\mathbb{T}}$ called *conditional expectation of the natural statistics*, and the M-step consists in updating the parameters θ given $\widehat{\mathbb{T}}$ by alternating optimization of each of the $J \times 9$ parameter subsets. This allows taking any combination of constraints specified by user into account. Second, given the mixture $\mathbf{X}$ and the estimated model parameters θ, source estimates $\widehat{\mathbf{Y}}$ are computed using Wiener filtering.

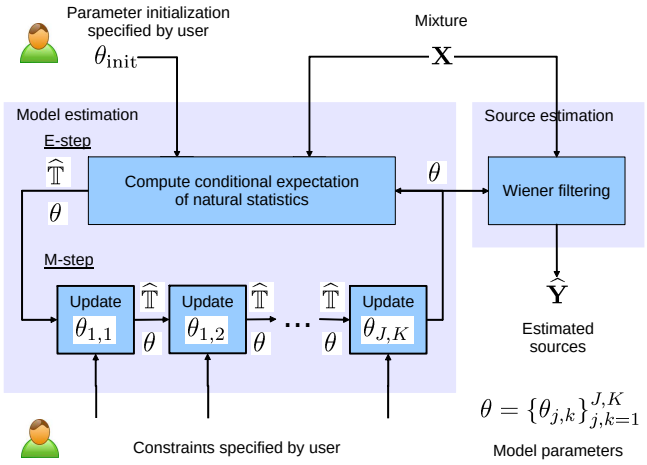


Fig. 2. Overview of the proposed general algorithm for parameter estimation and source separation.

D. FASST toolbox: Current baseline implementation

The FASST toolbox (released and available at [25]) implements so far a baseline version of the framework in Matlab that covers only the library of constraints summarized in Table III for mono or stereo recordings ($I = 1$ or $I = 2$). This restriction to up to $I = 2$ channels enables the use of a 2×2 matrix inversion trick described in [13] that leads to an efficient implementation in Matlab. However, the framework itself is neither restricted to the constraints in Table III nor to mono / stereo mixtures.

III. DETAILED STRUCTURE AND EXAMPLE CONSTRAINTS

In this section we describe in details the nine parameter subsets modeling each source and some example constraints. We also introduce the detailed notations to be used in the rest of the paper.

A. Formulation of the audio source separation problem

We assume that the observed I-channel time-domain signal, called *mixture*, $\tilde{\mathbf{x}}(t) \in \mathbb{R}^I$, $t = 1, \ldots, T$, is the sum of J multichannel signals $\tilde{\mathbf{y}}_j(t) \in \mathbb{R}^I$, called *spatial source images* [1], [22]:

$$\tilde{\mathbf{x}}(t) = \sum\nolimits_{j=1}^{J} \tilde{\mathbf{y}}_j(t). \tag{1}$$

The goal of source separation is to estimate the spatial source images $\tilde{\mathbf{y}}_j(t)$ given the mixture $\tilde{\mathbf{x}}(t)$. This now common formulation is more general than the convolutive formulation in [13], which is restricted to point sources [1], [22].

B. Input representation

Audio signals are usually processed in the TF domain, due to their sparsity in this domain. Two families of input representations are considered in the literature, namely *linear* [13] and *quadratic* [18] representations.

1) Linear representations: After applying a linear complex-valued TF transform, the mixture (1) becomes:

$$\mathbf{x}_{fn} = \sum\nolimits_{j=1}^{J} \mathbf{y}_{j,fn}, \tag{2}$$

where $\mathbf{x}_{fn} \in \mathbb{C}^I$ and $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ are I-dimensional complex-valued vectors of TF coefficients of the corresponding time-domain signals; and $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote respectively frequency bin and time-frame index. This formulation covers the STFT, that is the most popular TF representation used for audio source separation.

2) Quadratic representations: A few studies have relied on quadratic representations instead, where the signal is described in each TF bin by its empirical $I \times I$ covariance matrix [5], [10], [18]

$$\widehat{\mathbf{R}}_{\mathbf{x},fn} = \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H], \tag{3}$$

where $\widehat{\mathbb{E}}[\cdot]$ denotes *empirical expectation* computed, e.g., by local averaging of the STFT [5], [10] or of the input of an ERB filterbank [18]. Note that linear representations are special cases of quadratic representations with $\widehat{\mathbf{R}}_{\mathbf{x},fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$. Quadratic representations include additional information about the local correlation between channels which often increases the accuracy of parameters estimation [10]. In the following, we use the linear notations $\mathbf{x}_{fn}$ and $\mathbf{y}_{j,fn}$ for simplicity and include the empirical expectation when appropriate. A more rigorous derivation of the local Gaussian model for quadratic representations is given in Appendix A.

C. Local Gaussian model

We assume that in each TF bin, each source $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ is a proper complex-valued Gaussian random vector with zero mean and covariance matrix $\mathbf{\Sigma}_{\mathbf{y},j,fn} = v_{j,fn}\mathbf{R}_{j,fn}$

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right), \tag{4}$$

where the matrix $\mathbf{R}_{j,fn} \in \mathbb{C}^{I\times I}$ called *spatial covariance matrix* represents the spatial characteristics of the source and of the mixing setup, and the non-negative scalar $v_{j,fn} \in \mathbb{R}_+$ called *spectral power* represents the spectral characteristics of the source [1]. Moreover, the random vectors $\mathbf{y}_{j,fn}$ are assumed to be mutually independent given $\mathbf{\Sigma}_{\mathbf{y},j,fn}$.

D. Spatial covariance structure and example constraints

1) Structure: In the case of audio, it is mostly interesting to consider either rank-1 spatial covariances representing instantaneously or convolutively mixed point sources with low reverberation [13] or full-rank spatial covariances modeling diffuse or reverberated sources [19]. More generally, we assume covariances of any positive rank. Let $0 < R_j \leq I$ be the rank of covariance $\mathbf{R}_{j,fn}$. This matrix can then be non-uniquely represented as [3]

$$\mathbf{R}_{j,fn} = \mathbf{A}_{j,fn}\mathbf{A}_{j,fn}^H, \tag{5}$$

where $\mathbf{A}_{j,fn}$ is an $I \times R_j$ complex-valued matrix of rank R_j. Moreover, for every source j and for every TF bin (f,n) we introduce R_j independent Gaussian random variables $s_{jr,fn}$ $(r = 1,\ldots,R_j)$ distributed as

$$s_{jr,fn} \sim \mathcal{N}_c\left(0, v_{j,fn}\right). \tag{6}$$

With these notations the model defined by (2) and (4) is equivalent to the following mixture of $R = \sum_{j=1}^{J} R_j$ point *sub-sources* $s_{jr,fn}$:

$$\mathbf{x}_{fn} = \mathbf{A}_{fn}\mathbf{s}_{fn}, \tag{7}$$

where $\mathbf{s}_{fn} = [\mathbf{s}_{1,fn}^T,\ldots,\mathbf{s}_{J,fn}^T]^T$ is an $R \times 1$ vector of sub-source coefficients with $\mathbf{s}_{j,fn} = [s_{j1,fn},\ldots,s_{jR_j,fn}]^T$, and $\mathbf{A}_{fn} = [\mathbf{A}_{1,fn},\ldots,\mathbf{A}_{J,fn}]$ is an $I \times R$ mixing matrix. Thus, for a given TF bin (f,n) our model is equivalent to a complex-valued linear mixture of R sub-sources (7), where the sub-sources $s_{jr,fn}$ $(r = 1,\ldots,R_j)$ associated with the same source j share the same spectral power (6). We suppose that the rank R_j is specified for every source j.

2) Example constraints: In our baseline implementation we assume that the spatial covariances are time-invariant, i.e., $\mathbf{A}_{j,fn} = \mathbf{A}_{j,f}$. Moreover, we assume that for every source j the spatial parameters $\mathbf{A}_j$ can be either instantaneous (i.e., constant over frequency and real-valued: $\mathbf{A}_{j,fn} = \mathbf{A}_{j,n} \in \mathbb{R}^{I\times R_j}$) or convolutive (i.e., frequency-dependent), and either fixed, adaptive or partially adaptive. Some examples of constraints are given in Table III.

[3]Such an R_j-rank covariance matrix parametrization was inspired by [22], where $\mathbf{R}_{j,fn}$, intended to model correlated or multi-dimensional components, is parametrized as $\mathbf{R}_{j,fn} = \mathbf{A}_{j,fn}\mathbf{P}_{j,fn}\mathbf{A}_{j,fn}^H$, where $\mathbf{P}_{j,fn}$ is a full-rank $R_j \times R_j$ positive matrix. However, our parametrization (5) is less redundant and it is applied for audio source separation, and not for separation of components in astrophysical images, as in [22].

E. Spectral power structure and example constraints

To model spectral power we use nonnegative hierarchical audio-specific decompositions [23], thus all variables introduced in this section are assumed to be non-negative.

1) Excitation-filter model: We first model the spectral power $v_{j,fn}$ as the product of an excitation spectral power $v_{j,fn}^{\mathrm{ex}}$ and a filter spectral power $v_{j,fn}^{\mathrm{ft}}$ [23], [35]:

$$v_{j,fn} = v_{j,fn}^{\mathrm{ex}} \times v_{j,fn}^{\mathrm{ft}}, \tag{8}$$

that can be rewritten as

$$\mathbf{V}_j = \mathbf{V}_j^{\mathrm{ex}} \odot \mathbf{V}_j^{\mathrm{ft}}, \tag{9}$$

where $\odot$ denotes element-wise matrix multiplication and $\mathbf{V}_j \triangleq [v_{j,fn}]_{f,n}$, $\mathbf{V}_j^{\mathrm{ex}} \triangleq [v_{j,fn}^{\mathrm{ex}}]_{f,n}$, $\mathbf{V}_j^{\mathrm{ft}} \triangleq [v_{j,fn}^{\mathrm{ft}}]_{f,n}$.

Figure 3 gives an example of the excitation-filter decomposition (9) as applied to the spectral power of several guitar notes. In this example the filter $\mathbf{V}_j^{\mathrm{ft}}$ is time-invariant with lowpass characteristics, and the excitation $\mathbf{V}_j^{\mathrm{ex}}$ is a time-varying combination of few characteristic spectral patterns. However, in the most of realistic situations both the excitation and the filter are time-varying. Thus, the excitation-filter model with time-varying excitation and filter is a physically-motivated generative model that is suitable for many audio sources. While time-invariant filters were considered, e.g., in [7], [35], some approaches consider time-varying filters [16], [43]. We believe that our framework opens a door for further investigation of time-varying filters.

2) Excitation power structure: The excitation spectral power $[v_{j,fn}^{\mathrm{ex}}]_f$ is modeled as the sum of K_j^{ex} characteristic spectral patterns $[e_{j,fk}^{\mathrm{ex}}]_f$ modulated in time by $p_{j,kn}^{\mathrm{ex}}$, i.e., $v_{j,fn}^{\mathrm{ex}} = \sum_{k=1}^{K_j^{\mathrm{ex}}} p_{j,kn}^{\mathrm{ex}} e_{j,fk}^{\mathrm{ex}}$ [9]. Introducing the matrices $\mathbf{P}_j \triangleq [p_{j,kn}^{\mathrm{ex}}]_{k,n}$ and $\mathbf{E}_j^{\mathrm{ex}} \triangleq [e_{j,fk}^{\mathrm{ex}}]_{f,k}$ it can be rewritten as

$$\mathbf{V}_j^{\mathrm{ex}} = \mathbf{E}_j^{\mathrm{ex}}\,\mathbf{P}_j^{\mathrm{ex}}. \tag{10}$$

In order to further constrain the spectral fine structure of the spectral patterns, they are represented as linear combinations of L_j^{ex} narrowband spectral patterns $[w_{j,fl}^{\mathrm{ex}}]_f$ [14], i.e., $e_{j,fk}^{\mathrm{ex}} = \sum_{l=1}^{L_j^{\mathrm{ex}}} u_{j,lk}^{\mathrm{ex}} w_{j,fl}^{\mathrm{ex}}$, where $u_{j,lk}^{\mathrm{ex}}$ are non-negative weights. The series of time activation coefficients $p_{j,kn}^{\mathrm{ex}}$ are also represented as sums of M_j^{ex} time-localized patterns, i.e., $p_{j,kn}^{\mathrm{ex}} = \sum_{m=1}^{M_j^{\mathrm{ex}}} h_{j,mn}^{\mathrm{ex}} g_{j,km}^{\mathrm{ex}}$. Altogether we have:

$$v_{j,fn}^{\mathrm{ex}} = \sum\nolimits_{k=1}^{K_j^{\mathrm{ex}}} \sum\nolimits_{m=1}^{M_j^{\mathrm{ex}}} h_{j,mn}^{\mathrm{ex}} g_{j,km}^{\mathrm{ex}} \sum\nolimits_{l=1}^{L_j^{\mathrm{ex}}} u_{j,lk}^{\mathrm{ex}} w_{j,fl}^{\mathrm{ex}}, \tag{11}$$

and, introducing matrices $\mathbf{H}_j^{\mathrm{ex}} \triangleq [h_{j,mn}^{\mathrm{ex}}]_{m,n}$, $\mathbf{G}_j^{\mathrm{ex}} \triangleq [g_{j,km}^{\mathrm{ex}}]_{k,m}$, $\mathbf{U}_j^{\mathrm{ex}} \triangleq [u_{j,lk}^{\mathrm{ex}}]_{l,k}$ and $\mathbf{W}_j^{\mathrm{ex}} \triangleq [w_{j,fl}^{\mathrm{ex}}]_{f,l}$, this equation can be rewritten in matrix form as

$$\mathbf{V}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}\,\mathbf{U}_j^{\mathrm{ex}}\,\mathbf{G}_j^{\mathrm{ex}}\,\mathbf{H}_j^{\mathrm{ex}}. \tag{12}$$

Figure 4 shows an example of the excitation structure $\mathbf{V}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}\,\mathbf{U}_j^{\mathrm{ex}}\,\mathbf{G}_j^{\mathrm{ex}}\,\mathbf{H}_j^{\mathrm{ex}}$, as applied to six notes played on a xylophone. In this example, the narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ex}}$ include 66 harmonic patterns modeling the harmonic part of 11 notes and 9 smooth patterns modeling the attacks, and the matrix of weights $\mathbf{U}_j^{\mathrm{ex}}$ is very sparse so as to eliminate invalid combinations of narrowband spectral patterns (e.g., a

characteristic spectral pattern should not be a combination of narrowband spectral patterns with different pitches). The time-localized patterns $\mathbf{H}_j^{\text{ex}}$ include decreasing exponentials to model the decay part of the notes and discrete Dirac functions to model note attacks, and the matrix of weights $\mathbf{G}_j^{\text{ex}}$ is sparse so as not to allow the attacks (smooth spectral patterns) to be modulated by exponential temporal patterns and not to allow harmonic note parts (harmonic spectral patterns) to be modulated by Dirac temporal patterns. Such a structure is a simplified version of the conventional attack-decay-sustain-release model (see, e.g., [44]). More sophisticated structures, where, e.g., the sustain and release parts are modeled by exponentials with different decrease rates can be implemented as well within our framework.

3) Filter power structure: The filter spectral power $[v_{j,fn}^{\text{ft}}]_f$ is represented using exactly the same structure as in (11).

4) Total power structure: Altogether the spectral power structure can be represented by the following nonnegative matrix decomposition (see also Table II)

$$\mathbf{V}_j = \left(\mathbf{W}_j^{\text{ex}}\,\mathbf{U}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}\right) \odot \left(\mathbf{W}_j^{\text{ft}}\,\mathbf{U}_j^{\text{ft}}\,\mathbf{G}_j^{\text{ft}}\,\mathbf{H}_j^{\text{ft}}\right). \quad (13)$$

Each matrix in this decomposition is subject to specific constraints presented below.

5) Example constraints: Each matrix $\theta_{j,k}$ ($k = 2, \ldots, 9$) in (13) can be fixed, adaptive or partially fixed (see Tab. III). In the latter two cases, a probabilistic prior $p(\theta_{j,k}|\eta_{j,k})$, such as a time continuity prior [9] or a sparsity-inducing prior [4] can be set. We denote by $\eta_{j,k}$ the *hyperparameters* of the prior that can be fixed or adaptive as well.

To cover *discrete state-based models* such as GMM, HMM, and their scaled versions GSMM, S-HMM, every column $\mathbf{g}_{j,m}^{\text{ex}} = [g_{j,km}^{\text{ex}}]_k$ of matrix $\mathbf{G}_j^{\text{ex}}$ (and similarly for matrix $\mathbf{G}_j^{\text{ft}}$) may further be constrained to have either a single nonzero entry (for GSMM, S-HMM) or a single nonzero entry equal to 1 (for GMM, HMM). Let $q_{j,m}^{\text{ex}} \in \{1, \ldots, K_j^{\text{ex}}\}$ be the index of the corresponding nonzero entry and $\mathbf{q}_j^{\text{ex}} = [q_{j,m}^{\text{ex}}]_m$ the resulting *state sequence* [4]. The prior distribution of $\theta_{j,4} = \mathbf{G}_j^{\text{ex}}$ with hyperparameters $\eta_{j,4} = \mathbf{\Lambda}_j^{\text{ex}}$ is defined as

$$p(\theta_{j,4}|\eta_{j,4}) = p(\mathbf{q}_j^{\text{ex}}|\mathbf{\Lambda}_j^{\text{ex}}) = \prod\nolimits_{m=2}^{M_j^{\text{ex}}} \lambda_{j,q_{j,m-1}^{\text{ex}} q_{j,m}^{\text{ex}}}^{\text{ex}}, \quad (14)$$

where $\mathbf{\Lambda}_j^{\text{ex}} = [\lambda_{j,kk'}^{\text{ex}}]_{k,k'}$ ($\lambda_{j,kk'}^{\text{ex}} = \mathbb{P}(q_{j,m}^{\text{ex}} = k'|q_{j,m-1}^{\text{ex}} = k)$) denotes the $K_j^{\text{ex}} \times K_j^{\text{ex}}$ state transition probability matrix with $\lambda_{j,kk'}^{\text{ex}}$ being independent on k (i.e., $\lambda_{j,kk'}^{\text{ex}} = \lambda_{j,k'}^{\text{ex}}$) in the case of GMM or GSMM. As discussed in [12], the discrete state-based models are rather suitable for monophonic sources (e.g., singing voice or wind instruments), while the unconstrained NMF decompositions are more appropriate for polyphonic sources (e.g., piano or guitar).

F. Generality

It can be easily shown that the model structures considered in [2]–[19] are particular instances of the proposed general formulation. Let us give some examples.

Pham *et al* [3] assume rank-1 spatial covariances and constant spectral power over time-frequency regions of size 1 frequency bin $\times$ L frames. This structure can be implemented in our framework by choosing rank-1 adaptive spatial time-invariant covariances, i.e., $\mathbf{A}_j$ is an adaptive tensor of size $2 \times 1 \times F \times N$ subject to the time-invariance constraint, and constraining the spectral power to $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ being the $F \times F$ identity matrix, $\mathbf{G}^{\text{ex}}$ a $F \times \lceil N/L \rceil$ adaptive matrix, and $\mathbf{H}_j^{\text{ex}}$ the $\lceil N/L \rceil \times N$ fixed matrix with entries $h_{j,mn}^{\text{ex}} = 1$ for $n \in \mathcal{L}_m$ and $h_{j,mn}^{\text{ex}} = 0$ for $n \notin \mathcal{L}_m$, where $\mathcal{L}_m$ is the set of time frames belonging to the m-th block.

Multichannel NMF structures with point source (rank-1) [13] or diffuse source (full-rank) [17] models can be represented within our framework as $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ and $\mathbf{G}_j^{\text{ex}}$ being adaptive matrices of size $F \times K_j^{\text{ex}}$ and $K_j^{\text{ex}} \times N$, respectively, and $\mathbf{A}_j$ being an adaptive tensor of size $2 \times 1 \times F \times N$ or $2 \times 2 \times F \times N$, respectively, subject to the time-invariance constraint.

Excitation-filter model-based separation of the main melody vs. the background music from single-channel recordings by Durrieu *et al.* [16] can be represented within our framework as follows. Mixing parameters $\mathbf{A}_j$ ($j = 1, 2$) are assumed to form a tensor of size $1 \times 1 \times F \times N$ with all the entries fixed to 1. The background music spectral power $\mathbf{V}_1$ is modeled exactly as in the case of the multichannel NMF described in the previous paragraph. The main melody spectral power is constrained to $\mathbf{V}_2 = (\mathbf{W}_2^{\text{ex}}\,\mathbf{G}_2^{\text{ex}}) \odot (\mathbf{W}_2^{\text{ft}}\,\mathbf{G}_2^{\text{ft}})$ [5] with $\mathbf{W}_2^{\text{ex}}$ being fixed and $\mathbf{G}_2^{\text{ex}}$, $\mathbf{W}_2^{\text{ft}}$ and $\mathbf{G}_2^{\text{ft}}$ being adaptive. Without any supplementary constraints this model is equivalent to the model referred as *instantaneous mixture model* in [16], and applying GSMM constraints to both the matrices $\mathbf{G}_2^{\text{ex}}$ and $\mathbf{G}_2^{\text{ft}}$ this model is equivalent to the model referred as *GSMM* in [16].

IV. Estimation algorithm

In this section we describe in details the proposed algorithm for the estimation of the model parameters and subsequent source separation.

A. Model estimation criterion

To estimate the model parameters, we use the standard maximum *a posteriori* (MAP) where the log-likelihood $\log p(\mathbf{x}_{fn}|\theta)$ in every TF point is replaced by its empirical expectation $\widehat{\mathbb{E}}[\log p(\mathbf{x}_{fn}|\theta)]$ according to the empirical expectation operator $\widehat{\mathbb{E}}[\cdot]$ introduced in Section III-B2 [10], [18]. Mathematically rigorous derivation of this criterion is given in Appendix A. This criterion consists in maximizing the *modified log-posterior* $\widehat{\mathcal{L}}(\theta, \eta|\mathbf{X}) \triangleq \widehat{\mathbb{E}}[\log p(\theta, \eta|\mathbf{X})]$, where $\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n}$, over the model parameters θ and the hyperparameters $\eta = \{\eta_{j,k}\}_{j,k=1}^{J,9}$. This quantity can be rewritten,

[4]Note that we consider here the state sequence $\mathbf{q}_j^{\text{ex}}$ as a parameter to be estimated, and not as a latent variable one integrates over, as it is usually done for GMM / HMM parameter estimation. This is indeed to achieve the goal of generality by making the E-step of the GEM algorithm independent of the specified constraints.

[5]Note that any set of matrices can be virtually removed from the spectral power decomposition (13). For example, one can obtain $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$ by assuming that the matrices $\mathbf{W}_j^{\text{ft}}$, $\mathbf{U}_j^{\text{ft}}$, $\mathbf{G}_j^{\text{ft}}$ and $\mathbf{H}_j^{\text{ft}}$ are of sizes $F \times 1$, 1×1, 1×1, and $1 \times N$, and that all their entries are fixed to 1, and that $\mathbf{U}_j^{\text{ex}} = \mathbf{I}_{K_j^{\text{ex}}}$ is the $K_j^{\text{ex}} \times K_j^{\text{ex}}$ identity matrix.

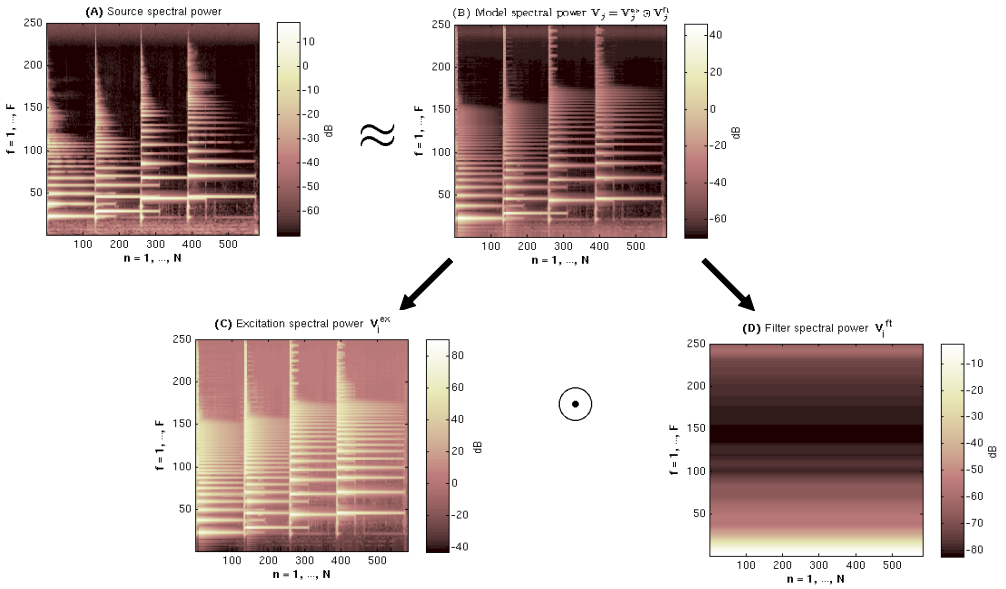


Fig. 3. Excitation-filter decomposition as applied to the spectral power of several guitar notes. **(A)**: source spectral power, **(B)**: model spectral power $\mathbf{V}_j = \mathbf{V}_j^{\text{ex}} \odot \mathbf{V}_j^{\text{ft}}$, **(C)**: excitation spectral power $\mathbf{V}_j^{\text{ex}}$, **(D)**: filter spectral power $\mathbf{V}_j^{\text{ft}}$.

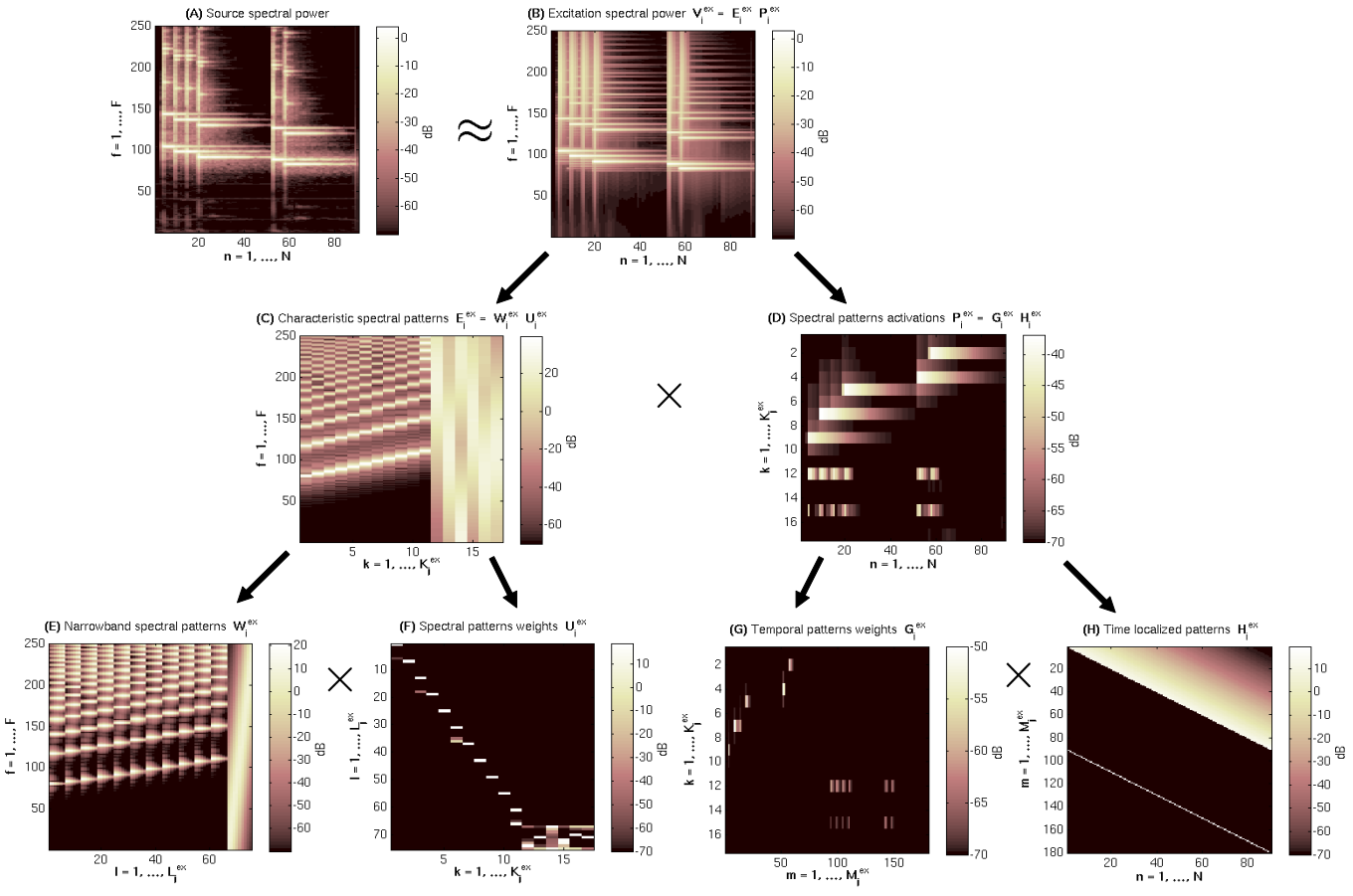


Fig. 4. Excitation power decomposition $\mathbf{V}_j^{\text{ex}} = \mathbf{W}_j^{\text{ex}}\,\mathbf{U}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$ as applied to the spectral power of several xylophone notes. **(A)**: source spectral power, **(B)**: excitation spectral power $\mathbf{V}_j^{\text{ex}} = \mathbf{E}_j^{\text{ex}}\mathbf{P}_j^{\text{ex}}$, **(C)**: characteristic spectral patterns $\mathbf{E}_j^{\text{ex}} = \mathbf{W}_j^{\text{ex}}\mathbf{U}_j^{\text{ex}}$, **(D)**: spectral pattern activations $\mathbf{P}_j^{\text{ex}} = \mathbf{G}_j^{\text{ex}}\mathbf{H}_j^{\text{ex}}$, **(E)**: narrowband spectral patterns $\mathbf{W}_j^{\text{ex}}$, **(F)**: spectral pattern weights $\mathbf{U}_j^{\text{ex}}$, **(G)**: temporal pattern weights $\mathbf{G}_j^{\text{ex}}$, **(H)**: time-localized patterns $\mathbf{H}_j^{\text{ex}}$.

using (2) and (4), as:

$$\widehat{\mathcal{L}}(\theta,\eta|\mathbf{X}) \stackrel{\text{c}}{=} \widehat{\mathcal{L}}(\mathbf{X}|\theta) + \log p(\theta|\eta) = \sum\nolimits_{f,n} \widehat{\mathbb{E}}[\log N_c(\mathbf{x}_{fn}|0,\boldsymbol{\Sigma}_{\mathbf{x},fn})] + \log p(\theta|\eta), \quad (15)$$

where $\boldsymbol{\Sigma}_{\mathbf{x},fn} \triangleq \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$, $\widehat{\mathcal{L}}(\mathbf{X}|\theta) \triangleq \widehat{\mathbb{E}}[\log p(\mathbf{X}|\theta)]$ is the *modified log-likelihood* and "$\stackrel{\text{c}}{=}$" denotes equality up to a constant. Using (3), the resulting criterion can be expressed as [13], [18]:

$$\theta^*,\eta^* = \arg\min_{\theta,\eta} \sum\nolimits_{f,n} \left[\text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn}\right) + \log|\boldsymbol{\Sigma}_{\mathbf{x},fn}|\right] - \sum\nolimits_{j,k=1}^{J,9} \log p(\theta_{j,k}|\eta_{j,k}). \quad (16)$$

We see that this criterion does not rely any more on the linear mixture representation $\mathbf{X}$, but only on the resulting empirical mixture covariances $\{\widehat{\mathbf{R}}_{\mathbf{x},fn}\}_{f,n}$.

B. Model estimation via a GEM algorithm

Given the model parameters $\theta = \{\theta_{j,k}\}_{j,k=1}^{J,9}$ specified in Table II and the hyperparameters $\eta = \{\eta_{j,k}\}_{j,k=1}^{J,9}$ together with user-defined constraints and initial values, we minimize the criterion (16) using a GEM algorithm [20] that consists in iterating the following expectation (E) and maximization (M) steps (see Fig. 2):

- *E-step:* Compute the conditional expectation of the so-called *natural (sufficient) statistics*, given the observations $\mathbf{X}$ and the current parameters θ, η.
- *M-step:* Given the expectation of the natural statistics, update the parameters θ, η so as to increase the conditional expectation of the modified log-posterior of the so-called *complete data* [20]. This step is implemented via a loop over all $J \times 9$ parameter subsets $\theta_{j,k}$ specified in Table II. Each subset, depending whether it is adaptive (partially adaptive) or fixed, is updated (partially updated) or not in turn using suitable update rules inspired by [9], [13], [14].

1) Preliminaries:

a) Additive noise and simulated annealing: As explained in [13], where a similar GEM algorithm is used, the mixing parameters $\mathbf{A}_{fn}$ (see Eq. (7)) updated via this GEM algorithm can become stuck into a suboptimal value. To overcome this issue, we use a form of *simulated annealing* proposed in [13], which consists in adding to (7) a noise term whose variance is decreased by a fixed amount at each iteration. Thus, we assume that there is a $J+1$-th source with full-rank time-invariant spatial covariance $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \sigma_f^2\mathbf{I}_I = \mathbf{R}_{J+1,fn}$ and trivial spectral power ($v_{J+1,fn} = 1$) that represents a controllable additive isotropic noise $\mathbf{b}_{fn} = \mathbf{y}_{J+1,fn}$. Introducing this noise component leads to considering the noise covariance $\boldsymbol{\Sigma}_{\mathbf{b},fn}$ as part of the model parameters θ and to adding it to the mixing equation (7):

$$\mathbf{x}_{fn} = \mathbf{A}_{fn}\mathbf{s}_{fn} + \mathbf{b}_{fn}. \quad (17)$$

b) Complete data log-posterior and natural statistics: We chose $\mathbf{Z} = \{\mathbf{X},\mathbf{S}\}$ as the complete data, where $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f,n}$, and the modified log-posterior of the complete data can be written as:

$$\begin{aligned}\widehat{\mathcal{L}}(\theta,\eta|\mathbf{X},\mathbf{S}) &\stackrel{\text{c}}{=} \widehat{\mathcal{L}}(\mathbf{X}|\mathbf{S};\theta) + \widehat{\mathcal{L}}(\mathbf{S}|\theta) + \log p(\theta|\eta)\\ &\stackrel{\text{c}}{=} -\sum_{f,n}\text{tr}\left[\boldsymbol{\Sigma}_{\mathbf{b},fn}^{-1}\left(\mathbf{R}_{\mathbf{x},fn} - \mathbf{A}_{fn}\mathbf{R}_{\mathbf{xs},fn}^H - \mathbf{R}_{\mathbf{xs},fn}\mathbf{A}_{fn}^H + \mathbf{A}_{fn}\mathbf{R}_{\mathbf{s},fn}\mathbf{A}_{fn}^H\right)\right] - \sum_{f,n}\log|\boldsymbol{\Sigma}_{\mathbf{b},fn}|\\ &\quad - \sum_j R_j\sum_{f,n} d_{IS}(\xi_{j,fn}|v_{j,fn}) + \sum_{j,k=1}^{J,9}\log p(\theta_{j,k}|\eta_{j,k}), \quad (18)\end{aligned}$$

where $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$ is the Itakura-Saito (IS) divergence [9], $v_{j,fn}$ are the entries of matrix $\mathbf{V}_j$ specified by (13), and $\mathbf{R}_{\mathbf{x},fn}$, $\mathbf{R}_{\mathbf{xs},fn}$, $\mathbf{R}_{\mathbf{s},fn}$ and $\xi_{j,fn}$ are defined as:

$$\mathbf{R}_{\mathbf{x},fn} \triangleq \widehat{\mathbf{R}}_{\mathbf{x},fn} = \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H], \quad \mathbf{R}_{\mathbf{xs},fn} \triangleq \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{s}_{fn}^H], \quad (19)$$

$$\mathbf{R}_{\mathbf{s},fn} \triangleq \widehat{\mathbb{E}}[\mathbf{s}_{fn}\mathbf{s}_{fn}^H], \quad \xi_{j,fn} \triangleq \frac{1}{R_j}\sum_{r=1}^{R_j}\widehat{\mathbb{E}}[|s_{jr,fn}|^2]. \quad (20)$$

It can be easily shown from (18) that the family of functions $\{\exp\widehat{\mathcal{L}}(\mathbf{X},\mathbf{S}|\theta)\}_\theta$ forms an *exponential family* [7], [20], and the set $\mathbb{T}(\mathbf{X},\mathbf{S}) = \{\mathbf{R}_{\mathbf{x},fn},\mathbf{R}_{\mathbf{xs},fn},\mathbf{R}_{\mathbf{s},fn}\}_{f,n}$ is a *natural (sufficient) statistics* [7] for this family. Given this result, we derive a GEM algorithm that is summarized below.

2) Conditional expectation of the natural statistics (E-step): The conditional expectations of the natural statistics $\mathbb{T}(\mathbf{X},\mathbf{S})$ are computed as follows:

$$\widehat{\mathbf{R}}_{\mathbf{xs},fn} = \widehat{\mathbf{R}}_{\mathbf{x},fn}\boldsymbol{\Omega}_{\mathbf{s},fn}^H, \quad (21)$$

$$\widehat{\mathbf{R}}_{\mathbf{s},fn} = \boldsymbol{\Omega}_{\mathbf{s},fn}\widehat{\mathbf{R}}_{\mathbf{x},fn}\boldsymbol{\Omega}_{\mathbf{s},fn}^H + (\mathbf{I}_R - \boldsymbol{\Omega}_{\mathbf{s},fn}\mathbf{A}_{fn})\boldsymbol{\Sigma}_{\mathbf{s},fn}, \quad (22)$$

where

$$\boldsymbol{\Omega}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}, \quad (23)$$

$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_{fn}\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^H + \boldsymbol{\Sigma}_{\mathbf{b},fn}, \quad (24)$$

$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \text{diag}\left([\phi_{r,fn}]_{r=1}^R\right), \quad (25)$$

and $\phi_{r,fn} = v_{j,fn}$ if and only if $r \in \mathcal{R}_j$, where $\mathcal{R}_j$ denotes the set of sub-source indices associated with source j in the vector $\mathbf{s}_{fn}$ (see section III-D).

3) Update of the spatial covariances (M-step):

a) Unconstrained time-invariant mixing parameters: We first consider the case where there are no probabilistic priors specified for the mixing parameters $\{\mathbf{A}_j\}_j$ and these parameters are time-invariant. Let $\mathcal{A},\mathcal{A}' \subset \{1,\ldots,R\}$ be subsets of indices of sizes $\mathcal{D} = \#(\mathcal{A})$ and $\mathcal{D}' = \#(\mathcal{A}')$, respectively. Below we denote by $\mathbf{A}_{fn}^{\mathcal{A}}$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}^{\mathcal{A}}$ and $\widehat{\mathbf{R}}_{\mathbf{s},fn}^{\mathcal{A}\mathcal{A}'}$ the matrices of respective sizes $I \times \mathcal{D}$, $I \times \mathcal{D}$ and $\mathcal{D} \times \mathcal{D}'$, that consist of the corresponding entries of the matrices $\mathbf{A}_{fn}$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}$ and $\widehat{\mathbf{R}}_{\mathbf{s},fn}$, i.e., $\mathbf{A}_{fn}^{\mathcal{A}} = [\mathbf{A}_{fn}(i,r)]_{i=1,r\in\mathcal{A}}^I$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}^{\mathcal{A}} = [\widehat{\mathbf{R}}_{\mathbf{xs},fn}(i,r)]_{i=1,r\in\mathcal{A}}^I$, and $\widehat{\mathbf{R}}_{\mathbf{s},fn}^{\mathcal{A}\mathcal{A}'} = [\widehat{\mathbf{R}}_{\mathbf{s},fn}(r,r')]_{r\in\mathcal{A},r'\in\mathcal{A}'}$. We also denote by $\overline{\mathcal{A}} = \{1,\ldots,R\}\backslash\mathcal{A}$ the complementary set. Let $\mathcal{C} \subset \{1,\ldots,R\}$ (resp. $\mathcal{I} \subset \{1,\ldots,R\}$) be the indices of convolutively (resp. instantaneously) mixed sources

with adaptive mixing parameters. With these conventions the mixing parameters are updated as follows [6]:

$$\mathbf{A}_{fn}^{\mathcal{C}} = \left[\sum_{\tilde{n}} \left\{\widehat{\mathbf{R}}_{\mathbf{xs},f\tilde{n}}^{\mathcal{C}} - \mathbf{A}_{f\tilde{n}}^{\overline{\mathcal{C}}} \widehat{\mathbf{R}}_{\mathbf{s},f\tilde{n}}^{\overline{\mathcal{C}}\mathcal{C}}\right\}\right] \left[\sum_{\tilde{n}} \widehat{\mathbf{R}}_{\mathbf{s},f\tilde{n}}^{\mathcal{C}\mathcal{C}}\right]^{-1}, \tag{26}$$

$$\mathbf{A}_{fn}^{\mathcal{I}} = \Re\left[\sum_{\tilde{f},\tilde{n}} \left\{\widehat{\mathbf{R}}_{\mathbf{xs},\tilde{f}\tilde{n}}^{\mathcal{I}} - \mathbf{A}_{\tilde{f}\tilde{n}}^{\overline{\mathcal{I}}} \widehat{\mathbf{R}}_{\mathbf{s},\tilde{f}\tilde{n}}^{\overline{\mathcal{I}}\mathcal{I}}\right\}\right] \left[\Re\left\{\sum_{\tilde{f},\tilde{n}} \widehat{\mathbf{R}}_{\mathbf{s},\tilde{f}\tilde{n}}^{\mathcal{I}\mathcal{I}}\right\}\right]^{-1} \tag{27}$$

b) Other constraints: Estimating time-varying mixing parameters without any priors does not make much sense in practice due to highly unconstrained nature of such the estimation. If the mixing parameters are given some Gaussian priors, closed-form updates similar to (26), (27) can be still derived, since the modified log-posterior (18) will be a quadratic form with respect to the mixing parameters. In case of nongaussian priors some Newton-like updates [22] can be derived.

4) Update of the spectral power parameters (M-step):

a) Unconstrained nonnegative matrices: Let $\mathbf{C}_j = \theta_{j,k}$ ($k = 2, \ldots, 9$) an adaptive or partially adaptive nonnegative matrix (see Tab II) with a uniform prior $p(\theta_{j,k}|\eta_{j,k}) = 1$. Whatever the matrix $\mathbf{C}_j$, it can be shown that the decomposition (13) can be rewritten as $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j$, where $\mathbf{B}_j$, $\mathbf{D}_j$ and $\mathbf{E}_j$ are some nonnegative matrices that are assumed to be fixed while $\mathbf{C}_j$ is updated. For example, if $\mathbf{C}_j = \mathbf{H}_j^{\mathrm{ft}}$ in (13), one can choose $\mathbf{B}_j = \mathbf{W}_j^{\mathrm{ft}}\,\mathbf{U}_j^{\mathrm{ft}}\,\mathbf{G}_j^{\mathrm{ft}}$, $\mathbf{D}_j = \mathbf{I}_N$ and $\mathbf{E}_j = \mathbf{W}_j^{\mathrm{ex}}\,\mathbf{U}_j^{\mathrm{ex}}\,\mathbf{G}_j^{\mathrm{ex}}\,\mathbf{H}_j^{\mathrm{ex}}$. With these notations it can be shown that the conditional expectation of the modified log-posterior (18) of the complete data is non-decreasing when the corresponding update for $\mathbf{C}_j$ does not increase the following cost function:

$$\mathcal{D}_{IS}(\mathbf{C}_j) = \sum\nolimits_{f,n} d_{IS}([\widehat{\boldsymbol{\Xi}}_j]_{f,n}|[\mathbf{V}_j]_{f,n}), \tag{28}$$

where $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j$ and $\widehat{\boldsymbol{\Xi}}_j = [\hat{\xi}_{j,fn}]_{f,n}$ with $\hat{\xi}_{j,fn}$ computed as follows:

$$\hat{\xi}_{j,fn} = \frac{1}{R_j} \sum\nolimits_{r\in\mathcal{R}_j} \widehat{\mathbf{R}}_{\mathbf{s},fn}(r,r), \tag{29}$$

where $\widehat{\mathbf{R}}_{\mathbf{s},fn}$ is computed in (22) and $\mathcal{R}_j$ is defined at the end of Section IV-B2. Applying some standard derivations (see, e.g., [9]), one can obtain the following nonnegative MU rule [7]

$$\mathbf{C}_j = \mathbf{C}_j \odot \frac{\mathbf{B}_j^T[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{E}_j \odot \{(\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j\}^{\cdot -2}]\mathbf{D}_j^T}{\mathbf{B}_j^T[\mathbf{E}_j \odot \{(\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j\}^{\cdot -1}]\mathbf{D}_j^T} \tag{30}$$

that guarantees non-increase of the cost function (28), and thus non-decrease of the conditional expectation of the modified log-posterior (18) of the complete data. These update rules, as applied to multichannel audio, are in fact a generalization of the GEM-MU algorithm proposed in [21], that has been shown to converge much more quickly than the GEM algorithm in [13].

b) Discrete state-based constraints: Let us now assume that $\theta_{j,4} = \mathbf{G}_j^{\mathrm{ex}}$ is subject to a discrete state-based constraint (similarly for $\theta_{j,8} = \mathbf{G}_j^{\mathrm{ft}}$). Note that when time-localized patterns $\mathbf{H}_j^{\mathrm{ex}}$ (or $\mathbf{H}_j^{\mathrm{ft}}$) have non-zero overlaps in time of maximum length L (see, e.g., Fig. 4) the model becomes equivalent to an HMM of the order L (in case of GMMs) or of the order $L+1$ (in case of HMMs). In order to avoid the complications of requiring consistency of overlapping patterns (which would introduce temporal constraints somewhat reminiscent of an HMM), in our baseline implementation and in the updates described below we only consider non-overlapping time-localized patterns $\mathbf{H}_j^{\mathrm{ex}} = \mathbf{I}_N$ in case of discrete state-based constraints. The updates are performed as follows:

1) Set $\tilde{\mathbf{G}}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}}$, and fill each entry of each column of $\tilde{\mathbf{G}}_j^{\mathrm{ex}}$ with the nonzero entry of the respective column of $\mathbf{G}_j^{\mathrm{ex}}$.
2) If $\mathbf{G}_j^{\mathrm{ex}}$ is adaptive, do for every $k = 1, \ldots, K_j^{\mathrm{ex}}$:
   - Set $\mathbf{C}_j = \tilde{\mathbf{G}}_j^{\mathrm{ex}}$, and set all the elements of $\mathbf{C}_j$ to zero, except the k-th row.
   - Update $\mathbf{C}_j$ using several iterations of (30) [8].
   - Set the k-th row of $\tilde{\mathbf{G}}_j^{\mathrm{ex}}$ equal to that of $\mathbf{C}_j$.
3) For every $k = 1, \ldots, K_j^{\mathrm{ex}}$ and $m = 1, \ldots, M_j^{\mathrm{ex}}$ set $\mathbf{C}_j = \tilde{\mathbf{G}}_j^{\mathrm{ex}}$, set all the elements of $\mathbf{C}_j$ to zero, except the (k,m)-th one, and compute the IS divergence $\mathcal{D}_{IS}(k,m)$ between $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j$ and $\widehat{\boldsymbol{\Xi}}_j$, as in (28).
4) Update the state sequence $\mathbf{q}_j^{\mathrm{ex}}$ using the Viterbi algorithm [45] to minimize the following criterion:
$$\mathbf{q}_j^{\mathrm{ex}} = \arg\min_{\mathbf{q}_j^{\mathrm{ex}}} \sum_{m=2}^{M_j^{\mathrm{ex}}} \mathcal{D}_{IS}(q_{j,m}^{\mathrm{ex}}, m) - \log p(\mathbf{q}_j^{\mathrm{ex}}|\boldsymbol{\Lambda}_j^{\mathrm{ex}}),$$
where $p(\mathbf{q}_j^{\mathrm{ex}}|\boldsymbol{\Lambda}_j^{\mathrm{ex}})$ is computed as in (14).
5) Set $\mathbf{G}_j^{\mathrm{ex}} = \tilde{\mathbf{G}}_j^{\mathrm{ex}}$ and set to zero all the entries of $\mathbf{G}_j^{\mathrm{ex}}$, except those corresponding to $\mathbf{q}_j^{\mathrm{ex}}$.
6) If $\boldsymbol{\Lambda}_j^{\mathrm{ex}}$ is adaptive, update the transition probabilities as $\lambda_{j,kk'}^{\mathrm{ex}} = \frac{\sum_{m=2}^{M_j^{\mathrm{ex}}} \mathbf{1}(q_{j,m-1}^{\mathrm{ex}}=k, q_{j,m}^{\mathrm{ex}}=k')}{(M_j^{\mathrm{ex}}-1)\sum_{m=2}^{M_j^{\mathrm{ex}}} \mathbf{1}(q_{j,m-1}^{\mathrm{ex}}=k)}$ in case of HMM or S-HMM or as $\lambda_{j,kk'}^{\mathrm{ex}} = \frac{1}{M_j^{\mathrm{ex}}-1}\sum_{m=2}^{M_j^{\mathrm{ex}}} \mathbf{1}(q_{j,m}^{\mathrm{ex}} = k')$ in case of GMM or GSMM.

c) Other constraints: We here discuss the updates that are not yet included in our current baseline implementation (see Sec. II-D).

An EM algorithm update rules for time pattern weights $\mathbf{G}_j^{\mathrm{ex}}$ or $\mathbf{G}_j^{\mathrm{ft}}$ with time continuity priors, such as inverse-Gamma or Gamma Markov chain priors, can be found in [9]. However, one cannot use these rules within our GEM algorithm, since we use a different, reduced, complete data set, as compared

[6] We see that the mixing parameters for different sources are updated jointly by Eqs. (26), (27), while we have claimed in the beginning of Section IV that they will be updated in an alternated manner. However, since we can here update parameters jointly without loss of flexibility, we do so, since joint optimization, as compared to the alternated one, leads in general to a faster convergence.

[7] In the case of partially adaptive matrix $\mathbf{C}_j$, only the adaptive matrix entries are updated with rule (30).

[8] Several iterations of update rule (30) are needed because all entries of $\tilde{\mathbf{G}}_j^{\mathrm{ex}}$ are initialized in step 1 from a particular sequence of gains carried by $\mathbf{G}_j^{\mathrm{ex}}$ and optimized for the current state sequence $\mathbf{q}_j^{\mathrm{ex}}$. Performing only one update of (30) would unfavor state sequence evaluation. However, to avoid all these issues, in our implementation we just keep matrix $\tilde{\mathbf{G}}_j^{\mathrm{ex}}$ in memory, skip step 1, and do only one iteration of (30).

to the one used in [9]. Nevertheless, one can always use some Newton-like updates [22] for these priors.

If a matrix $\theta_{j,k}$ ($k = 2, \ldots, 9$) is constrained with a sparsity-inducing prior [4], such as a Laplacian prior (corresponding to an l_1 norm penalty), it can be updated using the multiplicative updates described in [46], [47]. However, in such a case the renormalization described in the subsection below could not be applied, since it would change the value of the optimized criterion (16). At the same time, without any renormalization, the sparsity-inducing prior would loose its influence. To avoid that, all the other parameter subsets $\theta_{j,l}$ ($l \neq k$) should be constrained, e.g., to have a unitary (say l_1) norm, which can be handled using the gradient descent updates from [46] or the modified multiplicative updates from [47].

5) Renormalization: At the end of each GEM iteration, in order to avoid numerical (under/over-flow) problems, a renormalization of some parameters is done if needed, i.e., if these parameters are not already constrained by some priors that are not scale-invariant. This procedure is similar to the one described in [13], and it does not change the value of the optimized criterion (16). For example, the columns of matrix $\mathbf{U}_j^{\text{ex}}$ can be divided by their energies, and the rows of $\mathbf{G}_j^{\text{ex}}$ scaled accordingly (see (13)). Similar renormalization is applied in turn to each patameter subsets pairs $\theta_{j,k}, \theta_{j,k+1}$ ($k = 1, \ldots, 8$), and at the end of this operation the total energy is relegated into $\theta_{j,9}$.

C. Source estimation

Given the estimated model parameters θ, the sources can be estimated in the minimum mean square error (MMSE) sense via the Wiener filtering:

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}, \tag{31}$$

where $\boldsymbol{\Sigma}_{\mathbf{x},fn} = \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$. The counterpart of this equation for quadratic TF representations is given in Appendix A.

V. Experimental Illustrations

The goals of this experimental part are to illustrate on some examples how to specify the prior information in the framework, given a particular source separation problem, and to demonstrate that we can implement the existing and new methods within the framework. For that we first give an example of application of the framework to a music recording in a *non-blind* setting, i.e., when different sources are given different models according to the prior information. Second, we consider a few blind framework instances, corresponding to existing and new methods, and apply them for separation of underdetermined speech and music mixtures. Third, we describe how to apply the framework to solve the source separation problem mentioned in the beginning of the introduction, i.e., the separation of bass, drums and melody in music recordings. Finally, we briefly mention our application of the framework for speech separation in the context of noise robust speech recognition.

A. Non-blind separation of one music recording

1) Data: As an example stereo music recording to separate we took the 23-second snip of the song "Que pena tanto faz" by Tamy from the test dataset of the SiSEC 2008 [30] "Professionally produced music recordings" task. We know about this recording that there are two sources, a female singing voice and a guitar, that the voice is instantaneously mixed (panned) in the middle [9] and the guitar is possibly a non-point convolutive source.

2) Constraint specification and parameter initialization: To account for this information within our framework, we have chosen the following constraints. The singing voice mixing parameters $\mathbf{A}_1$ form a fixed tensor of size $2 \times 1 \times F \times N$ with all entries equal to 1. The guitar mixing parameters $\mathbf{A}_2$ form an adaptive tensor of size $2 \times 2 \times F \times N$ subject to the time-invariance constraint. The spectral powers $\mathbf{V}_j$ ($j = 1, 2$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\mathbf{U}_j^{\text{ex}}\mathbf{G}_j^{\text{ex}}\mathbf{H}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ and $\mathbf{H}_j^{\text{ex}}$ being fixed, and $\mathbf{U}_j^{\text{ex}}$ and $\mathbf{G}_j^{\text{ex}}$ being adaptive. The narrowband spectral patterns $\mathbf{W}_j^{\text{ex}}$ include $6 \times L$ harmonic patterns modeling the harmonic part of L pitches and 9 smooth patterns (see Fig. 4 (E) and [14]). The L pitches are chosen to cover the range of 77 - 1397 Hz (39 - 89 on the MIDI scale), which is enough for both the guitar and this particular singing. The time-localized patterns $\mathbf{H}_1^{\text{ex}}$ and $\mathbf{H}_2^{\text{ex}}$ are different. The singing voice time-localized patterns $\mathbf{H}_1^{\text{ex}}$ include half-Gaussians truncated at the left, i.e., only the right half is kept. The guitar time-localized patterns $\mathbf{H}_2^{\text{ex}}$ include decreasing exponentials to model the decay part of the notes and discrete Dirac functions to model note attacks (see Fig. 4 (H)). All adaptive parameters are initialized with random values. Finally, we used the ERB quadratic representation described in [18] as signal representation.

3) Results: After 500 iterations of the proposed GEM algorithm the separation results, measured in terms of the source to distortion ratio (SDR) [48], were 7.2 and 8.9 dB for voice and guitar, respectively. We have also separated the same mixture using all the blind settings described in the following section. The best results of 5.5 and 7.1 dB SDR were obtained by the unconstrained NMF spectral power model with the instantaneous rank-1 mixing, i.e., by the multichannel NMF for instantaneous mixtures [13].

4) Discussion: We see that our informed setting outperforms any blind setting by at least 1.7 dB SDR. This improvement is essentially due to the combination of rank-1 instantaneous and full-rank convolutive mixing models and the information about the position of one source. Moreover, while it is common in professionally produced music recordings that some sources are mixed instantaneously (panned) and others convolutively (e.g., live-recorded tracks or some artificial reverberation is added), in our best knowledge such hybrid models were not yet proposed for audio source separation, and it now becomes possible to implement them within our framework.

[9] This information can be for example obtained by subtracting the left channel from the right one and checking that the voice is cancelled.

B. Blind separation of underdetermined speech and music mixtures

1) Data: Here we evaluate several settings of our framework on the development dataset of the SiSEC 2010 [29] "Underdetermined-speech and music mixtures" task. This dataset include 10-seconds length instantaneous, convolutive and live-recorded stereo mixtures of three or four music and speech sources (see [29] for more details).

2) Constraint specification and parameter initialization: We consider eight blind settings of the framework that are specified by the following constraints. For all settings and for all sources $\mathbf{A}_j$ forms an adaptive tensor of size $2 \times R_j \times F \times N$ subject to the time-invariance constraint and subject to the frequency invariance constraint for instantaneous mixtures only. The spectral power of each source is structured as $\mathbf{V}_j = \mathbf{E}_j^{\mathrm{ex}} \mathbf{P}_j^{\mathrm{ex}}$ [5]. The eight settings are generated by all possible combinations of the following possibilities (see also Table IV):

- *Rank:* The rank R_j is either 1 or 2 (full-rank).
- *Spectral structure:* The characteristic spectral patterns $\mathbf{E}_j^{\mathrm{ex}}$ are either *unconstrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}$ with adaptive $\mathbf{W}_j^{\mathrm{ex}}$, or *constrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}}$ with fixed $\mathbf{W}_j^{\mathrm{ex}}$ being composed of harmonic and noise-like and smooth narrowband spectral patterns (see Fig. 4 (E) and [14]), and adaptive $\mathbf{U}_j^{\mathrm{ex}}$ (see Fig. 4 (F)) that is very sparse so as to eliminate invalid combinations of narrowband spectral patterns (e.g., patterns corresponding to different pitches should not be combined together).
- *Temporal structure:* The time activation coefficients $\mathbf{P}_j^{\mathrm{ex}}$ are either *unconstrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}}$ with adaptive $\mathbf{G}_j^{\mathrm{ex}}$, or *constrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}} \mathbf{H}_j^{\mathrm{ex}}$ with fixed $\mathbf{H}_j^{\mathrm{ex}}$ being composed of decreasing exponentials, as those on Fig. 4 (H), and adaptive $\mathbf{G}_j^{\mathrm{ex}}$.

The two settings with $R_j = 1$ and 2, and unconstrained $\mathbf{E}_j^{\mathrm{ex}}$ and $\mathbf{P}_j^{\mathrm{ex}}$ correspond to the state-of-the-art methods [13] and [17], respectively (see Section III-F), while the remaining six settings are new.

In line with [13], parameter estimation via GEM is sensitive to initialization for all the settings we consider. To provide our GEM algorithm with a "good initialization" we used for the instantaneous mixtures the DEMIX mixing matrix estimation algorithm [49] to initialize mixing parameters $\mathbf{A}_j$, followed by l_0 norm minimization (see e.g., [1]) and Kullback-Leibler (KL) divergence minimization (see [13]) to initialize the source power spectra $\mathbf{V}_j$. For synthetic convolutive and live recorded mixtures we first estimated the time differences of arrival (TDOAs) using the MVDRW estimation algorithm proposed in [50], that is based on a variance distortionless response (MVDR) beamformer. The estimated TDOAs were then used to initialize anechoic mixing parameters $\mathbf{A}_j$, followed by binary masking and KL divergence minimization (see [13]) to initialize the source power spectra $\mathbf{V}_j$. As signal representation we used the STFT.

3) Results: Source separation results in terms of average SDR after 200 iterations of the proposed GEM algorithm are summarized in Table IV together with results of the *baseline* used for initialization.

4) Discussion: As expected, in most cases rank-1 spatial covariances perform the best for instantaneous mixtures and full-rank spatial covariances perform the best for synthetic convolutive and live recorded mixtures. Moreover, in all the cases there is at least one of the six new methods that outperforms the state-of-the-art methods [13] and [17]. One can note that for music sources constraining the spectral structure does not improve the separation performance [10], however, constraining the temporal structure does improve it. For speech sources constraining both the spectral and the temporal structures improves the separation performance in most cases. This is probably because the unconstrained NMF is a poor model for speech. Indeed, as compared to simple music, speech includes much more different spectral patterns, notably due to a more pronounced vibrato effect (varying pitch). As a consequence, the unconstrained NMF model needs much more components to describe this variability, thus it cannot be estimated in a robust way from these quite short 10-second length mixtures. Introducing spectral and temporal constraints makes model estimation more robust.

C. Separation of bass, drums and melody in music recordings

Here we describe how to apply our framework to the separation of the bass, the drums, the melody and the remaining instruments from a stereo professionally produced music recording. This source separation problem is of great practical interest for music information retrieval and remastering (e.g., karaoke) applications.

1) State-of-the-art: The state-of-the-art approaches targeting this problem suffer from the following limitations. First, existing drum [52] and melody [16] separation algorithms have been designed for single-channel (mono) recordings and may fail to segregate the melody from the other harmonic sources despite the fact that they have different spatial directions. Second, blind source separation methods relying on joint use of spatial and spectral diversity, such as, e.g., the multichannel NMF [13], need some user input to label separated signals [21] and cannot separate sources mixed in the same direction, which is a very common situation, e.g., for singing melody and drums. Finally, no state-of-the-art approach treats this problem in a joint fashion and cascading the methods (e.g., separating the drums, then separating the melody, etc.) is clearly suboptimal. Thus, it is clear that an efficient solution to this problem should rely on:

- some prior knowledge about the source spectral characteristics (to label the sources automatically),
- the spatial diversity of different sources,
- some model describing harmonicity, and
- joint modeling of all sources.

2) Constraint specification, parameter initialization and reconstruction: Our framework satisfies these requirements, and in order to account for this information we have chosen the following constraints. The two-channel mixture is modeled as a sum of 12 sources: 4 sources ($j = 1, \ldots, 4$) representing

[10]The results for synthetic convolutive mixtures of music sources are not very informative because of the poor overall performance.

Mixing				instantaneous		synthetic convolutive				live recorded			
Sources				speech	music	speech		music		speech		music	
Microphone spacing				-	-	5 cm	1 m	5 cm	1 m	5 cm	1 m	5 cm	1 m
Number of 10 second-length mixtures				6	4	10	10	4	4	10	10	4	4
baseline (l_0 minimization [51] or binary masking)				8.6	12.4	1.0	1.4	-0.9	-0.7	1.1	1.4	2.5	0.3
Method	rank R_j	spectral struct.	temporal struct.										
[13]	1	unconstrained	unconstrained	8.8	17.2	1.6	2.1	-1.1	-1.2	2.2	2.5	3.2	0.4
[17]	2	unconstrained	unconstrained	8.9	17.0	1.8	2.7	-0.5	-0.2	2.0	3.0	3.5	0.8
new	1	constrained	unconstrained	**10.5**	13.6	1.9	2.5	-0.5	-0.5	2.2	2.8	3.0	0.5
new	2	constrained	unconstrained	10.4	13.0	**2.1**	3.1	-0.7	-0.4	2.3	3.2	3.2	0.8
new	1	unconstrained	constrained	8.9	**18.6**	1.5	2.2	-0.8	-0.5	2.4	2.6	3.4	0.9
new	2	unconstrained	constrained	8.7	15.4	1.8	2.6	-0.4	0.0	2.1	2.9	**4.5**	**1.8**
new	1	constrained	constrained	**10.5**	15.7	**2.1**	2.9	-1.2	**0.3**	**2.5**	3.9	3.2	0.4
new	2	constrained	constrained	10.2	13.8	**2.1**	**4.5**	**0.0**	-0.3	2.3	**5.0**	3.7	1.0

TABLE IV
AVERAGE SDRS ON SUBSETS OF SISEC 2010 "UNDERDETERMINED SPEECH AND MUSIC MIXTURES" TASK DEVELOPMENT DATASET.

the bass, 4 sources ($j = 5, \ldots, 8$) representing the drums [11], and the remaining 4 sources ($j = 9, \ldots, 12$) representing the melody and the other instruments. Each set of mixing parameters $\mathbf{A}_j$ ($j = 1, \ldots, 12$) form an adaptive tensor of size $2 \times 2 \times F \times N$ subject to the time-invariance constraint. The spectral powers $\mathbf{V}_j$ of the bass and the drums ($j = 1, \ldots, 8$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\mathrm{ex}} \mathbf{G}_j^{\mathrm{ex}}$ [5] with $\mathbf{G}_j^{\mathrm{ex}}$ being adaptive and $\mathbf{W}_j^{\mathrm{ex}}$ being fixed and pre-trained (using our framework) from isolated bass and drum samples from the RWC music database [53]. The spectral powers $\mathbf{V}_j$ of the melody and the remaining instruments ($j = 9, \ldots, 12$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}} \mathbf{G}_j^{\mathrm{ex}}$ [5] with $\mathbf{W}_j^{\mathrm{ex}}$ being fixed, and $\mathbf{U}_j^{\mathrm{ex}}$ and $\mathbf{G}_j^{\mathrm{ex}}$ being adaptive. The narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ex}}$ ($j = 9, \ldots, 12$) include $3 \times L$ harmonic patterns modeling the harmonic part of L pitches (see [14]). The L pitches are chosen to cover the range of 27 - 4186 Hz (21 - 108 on the MIDI scale), which is enough to cover the pitch range of most instruments. All adaptive parameters are initialized with random values, except the mixing parameters $\mathbf{A}_j$ ($2 \times 2 \times F \times N$ tensors) that are initialized with the same (random) $2 \times 2 \times N$ tensor for all frequency bins. We used the ERB quadratic representation in [18] as signal representation due to its higher low-frequency resolution than the STFT, which is desirable for the modeling of bass sounds. Once the GEM algorithm has run, the 12 sources are estimated via Wiener filtering. The bass and the drums are reconstructed by summing the corresponding source estimates, the melody is reconstructed by choosing the most energetic source among the corresponding four ($j = 9, \ldots, 12$) sources, and the remaining instruments by summing the other three sources.

3) Results: The corresponding source separation script together with one separation example are available from the FASST web page [25]. Note that this example is a difficult, real-world mixture, which involves several sources mixed in the center (bass, singing voice, certain drums) and several harmonic sources with comparable pitch range (singing voice, piano).

[11] The bass is modeled as a sum of 4 sources to facilitate initialization, since we do not know a priori its spatial direction. The drums are modeled as a sum of 4 sources for the same reason, but also because the drum track is often composed of several sources (e.g., snare, hi-hat, cymbals, etc) that can be mixed in different directions.

D. Separation of speech in multi-source environment for noise robust speech recognition

We have also applied the framework for the problem of speech separation in reverberant noisy multi-source environment. This was done for our submission to the 2011 CHiME Speech Separation and Recognition Challenge [12]. The corresponding description can be found in [54] and some separation examples are available from a demo web page at [13].

VI. CONCLUSION

We have introduced a general flexible audio source separation framework that generalizes several existing source separation methods, brings them into a common framework, and allows to imagine and implement new efficient methods, given the prior information about a particular source separation problem. Besides the framework itself, we proposed a new temporal structure for NMF-like decompositions and an original mixing model formulation combining rank-1 and full-rank spatial mixing models in a homogeneous way. Finally, we provided a proper probabilistic formulation of local Gaussian modeling for quadratic time-frequency representations.

In the experimental part we have illustrated how to specify the prior information about a particular source separation problem within the framework, and we have shown that the framework allows implementing existing and new efficient source separation methods. We have also demonstrated that in some situations our new propositions can improve the source separation performance, as compared to the state-of-the-art. As such combining instantaneous rank-1 and and convolutive full-rank can be useful for separation of professionally produced music recordings, and the newly proposed temporal structure for NMF-like decompositions brings some improvement for blind separation of underdetermined mixtures of speech and music sources.

As for further research, the following extensions could be introduced to the framework. In a similar fashion as for

[12] http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html
[13] http://www.irisa.fr/metiss/ozerov/chime_ssep_demo.html

spectral power, a flexible structure can be specified for the mixing parameters. E.g., the time-varying mixing parameters could be represented in terms of time-localized and locally time-invariant mixing parameter patterns, thus allowing the modeling of moving sources. Another interesting extension would be to introduce possible coupling between parameter subsets, thus allowing, e.g., the representation of the characteristic spectral patterns of different sources as linear combinations of eigenvoices [55] or eigeninstruments [56]. In fact, some parameter subsets corresponding to different sources can share common properties, and introducing such a coupling would make the estimation of these parameters more robust.

APPENDIX A
PROBABILISTIC FORMULATION OF THE LOCAL GAUSSIAN MODEL FOR QUADRATIC REPRESENTATIONS

Here we give a proper probabilistic formulation of the local Gaussian model (4) for quadratic representations, explaining the exact meaning of the empirical covariance (3) and a justification of the criterion (16).

A. Input representation

Following [10], [18], we assume that the considered quadratic TF representation is computed by local averaging of a linear TF representation such as a STFT or an ERB filterbank. We assume that the indexing of the considered linear TF complex-valued representation, hereafter noted as $m = 1, \ldots, M$, can be in general different from the indexing f, n of the quadratic representation (3). Such a formulation allows considering linear and quadratic representations with different TF resolutions, but also using linear TF representations that do not allow any uniform TF indexing, e.g., an ERB representation with different sampling frequencies in different frequency bands or a signal-adapted multiple-window STFT [57]. The mixing equation (1) now writes as

$$\mathbf{x}_m = \sum\nolimits_{j=1}^{J} \mathbf{y}_{j,m}, \tag{32}$$

and we re-define the empirical covariance (3) as

$$\widehat{\mathbf{R}}_{\mathbf{x},fn} = \sum\nolimits_{m} (\omega^{\mathrm{ana}}_{fn,m})^2 \mathbf{x}_m \mathbf{x}_m^H, \tag{33}$$

where $\omega^{\mathrm{ana}}_{fn,m} \geq 0$, satisfying $\sum_{f,n} (\omega^{\mathrm{ana}}_{fn,m})^2 = 1$, are the coefficients of a local bi-dimensional *analysis* window specifying a neighbourhood of the TF point (f, n) [10], [18].

B. Local Gaussian model

In this setting the local Gaussian model (4) is re-defined as follows. Each vector $\mathbf{y}_{j,m}$ is assumed to be distributed as

$$\mathbf{y}_{j,m} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right) \tag{34}$$

with probability $(\omega^{\mathrm{ana}}_{fn,m})^2$. In other words, $\mathbf{y}_{j,m}$ is a realization of a GMM. Moreover, the vectors $\{\mathbf{y}_{j,m}\}_j$ are assumed to be independent only conditionally on the same GMM state. More precisely, the joint probability density function of $\{\mathbf{y}_{j,m}\}_j$ is defined as

$$p(\mathbf{y}_{1,m}, \ldots, \mathbf{y}_{J,m}) \triangleq \sum\nolimits_{fn} (\omega^{\mathrm{ana}}_{fn,m})^2 \prod\nolimits_j N_c\left(\mathbf{y}_{j,m}; \bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right). \tag{35}$$

C. Model estimation criterion

Under the above-presented assumptions (see (32) and (35)), the log-posterior $\log p(\theta, \eta|\mathbf{X})$, maximized by the MAP criterion, writes

$$\log p(\theta, \eta|\mathbf{X}) \stackrel{\mathrm{c}}{=} \log p(\mathbf{X}|\theta) + \log p(\theta|\eta) = \sum_{f,n} \log \sum_{m} (\omega^{\mathrm{ana}}_{fn,m})^2 N_c(\mathbf{x}_m; \bar{0}, \mathbf{\Sigma}_{\mathbf{x},fn}) + \log p(\theta|\eta), \tag{36}$$

where $\mathbf{\Sigma}_{\mathbf{x},fn} = \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$. Log-posterior (36) is difficult to optimize, due to summations in log-domain. Thus, following the EM methodology [20], we replace $\log p(\theta, \eta|\mathbf{X})$ by its lower bound

$$\sum_{f,n} \sum_{m} (\omega^{\mathrm{ana}}_{fn,m})^2 \log N_c(\mathbf{x}_m; \bar{0}, \mathbf{\Sigma}_{\mathbf{x},fn}) + \log p(\theta|\eta), \tag{37}$$

using Jensen's inequality [20], and we get the criterion (16) with empirical covariances $\widehat{\mathbf{R}}_{\mathbf{x},fn}$ computed as in (33). Thus, the criterion (16) maximizes a lower bound of the log-posterior (36).

Note, that with this formulation we could obtain exactly the same updates as those presented in Section IV-B by deriving a GEM algorithm for the MAP criterion (36). This is because the computing of the lower bound (37) is based on the EM methodology. However, we prefer to keep the criterion (16), since it makes the formulation more compact and links it to quadratic representations and to the existing works [10], [18].

D. Source estimation

The sources can be estimated as follows [10], [18]:

$$\hat{\mathbf{y}}_{j,m} = \sum\nolimits_{f,n} \omega^{\mathrm{syn}}_{fn,m} \omega^{\mathrm{ana}}_{fn,m} v_{j,fn} \mathbf{R}_{j,fn} \mathbf{\Sigma}^{-1}_{\mathbf{x},fn} \mathbf{x}_m, \tag{38}$$

where $\omega^{\mathrm{syn}}_{fn,m} \geq 0$ is a so-called *synthesis* window satisfying $\sum_{f,n} \omega^{\mathrm{syn}}_{fn,m} \omega^{\mathrm{ana}}_{fn,m} = 1$. This estimator becomes the MMSE estimator when $\omega^{\mathrm{syn}}_{fn,m} = \omega^{\mathrm{ana}}_{fn,m}$.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] E. Vincent, M. 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*. IGI Global, 2010, ch. 7, pp. 162–185.

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

[3] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proceedings of the 7th International Symposium on Signal Processing and its Applications*, 2003, pp. II–73–76.

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

[5] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05)*, Mohonk, NY, USA, Oct. 2005, pp. 78–81.

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

[7] 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, July 2007.

[8] R. Blouet, G. Rapaport, I. Cohen, and C. Févotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, Apr. 2008, pp. 37 – 40.

[9] 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, no. 3, pp. 793–830, Mar. 2009.

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

[11] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'09)*, 2009, pp. 751–758.

[12] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, Oct. 18–21, 2009, pp. 121–124.

[13] A. Ozerov and C. Févotte, "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, March 2010.

[14] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[15] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[16] 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, 2010.

[17] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *10th Int. Conf. on Information Sciences, Signal Proc. and their applications (ISSPA'10)*, 2010, pp. 1–4.

[18] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sep. 27-30 2010, pp. 73–80.

[19] ——, "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.

[20] 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. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[21] 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 (ICASSP'11)*, Prague, Czech Republic, May 2011, pp. 257 – 260.

[22] J.-F. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule, and G. Patanchon, "Component separation with flexible models — Application to multichannel astrophysical observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 5, pp. 735–746, 2008.

[23] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience. Hindawi Publishing Corp*, vol. 2008, 2008.

[24] 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 (LVA/ICA'10)*, Saint-Malo, France, Sep. 27-30 2010, pp. 33–40.

[25] ——, "Flexible Audio Source Separation Toolbox (FASST)." [Online]. Available: http://bass-db.gforge.inria.fr/fasst/

[26] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[27] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[28] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[29] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. 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 (LVA/ICA'10)*, Saint-Malo, France, Sep. 2010, pp. 114–122.

[30] E. Vincent, S. Araki, and P. Bofilld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA'09)*, 2009, pp. 734–741.

[31] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, April 1997, pp. 3617 – 3620.

[32] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69 – 84, 2010.

[33] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs." in *Fifth International Conference on Independent Component Analysis*, Granada, Spain, Sep. 2004, pp. 494–499.

[34] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[35] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'07)*, vol. 1, 2007, pp. 53 – 56.

[36] I. Lee, T. Kim, and T.-W. Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind speech separation*. Springer, 2007, pp. 169–192.

[37] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'08)*, 2008, pp. 1833–1836.

[38] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, no. Article ID 785152, 2009.

[39] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 793–799.

[40] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[41] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'01)*, San Diego, USA, Dec. 2001, pp. 1–6.

[42] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[43] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Transactions on*

Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 744–753, 2011.
[44] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *Proc. Int. Conf. on Spoken Language Processing*, 1998.
[45] 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, 1989.
[46] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
[47] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'04)*, 2004, pp. 2529–2533.
[48] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[49] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121 –133, jan. 2010.
[50] C. Blandin, E. Vincent, and A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, Czech Republic, May 2011, pp. 2616 – 2619.
[51] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07)*, 2007, pp. 430–437.
[52] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.
[53] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound databases," in *5th International Symposium on Music Information Retrieval (ISMIR)*, 2004, pp. 229–230. [Online]. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/
[54] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, Florence, Italy, September 2011, pp. 86–87.
[55] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.
[56] G. Grindlay and D. Ellis, "Multi-voice polyphonic music transcription using eigeninstruments," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, 2009, pp. 53–56.
[57] L. Benaroya, R. Blouet, C. Févotte, and I. Cohen, "Single sensor source separation using multiple-window STFT representation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 12–14 2006.

Alexey Ozerov holds a Ph.D. in Signal Processing from the University of Rennes 1 (France). He worked towards this degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, he received an M.Sc. degree in Mathematics from the Saint-Petersburg State University (Russia) in 1999 and an M.Sc. degree in Applied Mathematics from the University of Bordeaux 1 (France) in 2003. From 1999 to 2002, Alexey worked at Terayon Communicational Systems (USA) as a R&D software engineer, first in Saint-Petersburg and then in Prague (Czech Republic). He was for one year (2007) in Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) in TELECOM ParisTech / CNRS LTCI - Signal and Image Processing (TSI) Department. Now he is with METISS team of IRISA / INRIA - Rennes as a Post-Doc researcher. His research interests include audio source separation, source coding, and automatic speech recognition.

Emmanuel Vincent (M'07 - SM'10) is a Research Scientist with the French National Institute for Research in Computer Science and Control (INRIA, Rennes, France). Prior to that he received the Ph.D. degree in music signal processing from IRCAM (Paris, France) in 2004 and worked as a Research Assistant with the Centre for Digital Music at Queen Mary, University of London (London, U.K.) 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 the founding chair of the annual Signal Separation Evaluation Campaign (SiSEC). Dr. Vincent is an Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing.

Frédéric Bimbot received in 1988 a PhD in signal processing (speech synthesis using temporal decomposition), after graduating as a telecommunication engineer in 1985 (ENST, Paris, France). He also obtained in 1987 a B.A. in Linguistics (Sorbonne Nouvelle University, Paris III).

In 1990, he joined CNRS (French National Center for Scientific Research) as a permanent researcher, worked with ENST for 7 years and then moved to IRISA (CNRS & INRIA), in Rennes. He also repeatedly visited AT&T – Bell Laboratories between 1990 and 1999. He is now a Senior Researcher with CNRS.

He is heading the (IRISA/CNRS & INRIA/Rennes) METISS research group, dedicated to selected topics in speech and audio processing and his research is focused on speech and audio analysis, speaker recognition, music content modeling and audio source separation. He is also in charge of the coordination of D5 Department (Digital Signals and Images, Robotics) at IRISA.

He has been involved in a number of research projects, among which the ongoing QUAERO project in multimedia information retrieval.

Paper 3 (Ozerov, Liutkus, Badeau & Richard, *IEEE TASLP*, 2013)

Coding-based Informed Source Separation: Nonnegative Tensor Factorization Approach

Alexey Ozerov, Antoine Liutkus, *Member, IEEE,* Roland Badeau, *Senior Member, IEEE,* and Gaël Richard, *Senior Member, IEEE.*

***Abstract*—Informed source separation (ISS) aims at reliably recovering sources from a mixture. To this purpose, it relies on the assumption that the original sources are available during an *encoding* stage. Given both sources and mixture, a *side-information* may be computed and transmitted along with the mixture, whereas the original sources are not available any longer. During a *decoding* stage, both mixture and side-information are processed to recover the sources. ISS is motivated by a number of specific applications including active listening and remixing of music, karaoke, audio gaming, etc. Most ISS techniques proposed so far rely on a source separation strategy and cannot achieve better results than oracle estimators. In this study, we introduce Coding-based ISS (CISS) and draw the connection between ISS and source coding. CISS amounts to encode the sources using not only a model as in source coding but also the observation of the mixture. This strategy has several advantages over conventional ISS methods. First, it can reach any quality, provided sufficient bandwidth is available as in source coding. Second, it makes use of the mixture in order to reduce the bitrate required to transmit the sources, as in classical ISS. Furthermore, we introduce Nonnegative Tensor Factorization as a very efficient model for CISS and report rate-distortion results that strongly outperform the state of the art.**

***Index Terms*—Informed source separation, spatial audio object coding, source coding, constrained entropy quantization, probabilistic model, nonnegative tensor factorization.**

I. Introduction

AUDIO compression has been a very active field of research for several decades due to the tremendous demand for transmitting, or storing, digital audio signals at reduced rates. Audio compression can either be *lossless* (the original signal can be exactly recovered) or *lossy* (the original signal can only be approximately recovered). The latter scheme which reaches much higher compression ratios usually exploits psychoacoustic principles to minimize the perceptual loss. A large variety of methods were developed amongst which some have been standardized. MPEG1-Layer 3 (e.g. mp3) [1] or Advanced Audio Coding (AAC) [2] are probably amongst the most widely popular standardized lossy audio compression schemes. It is generally admitted that most coding schemes either rely on a parameterized signal model (e.g. as in *parametric coding approaches*), or on a direct quantization of the signal (as in *waveform* or *transform coding*), but also in some cases on a combination of both [3].

Concurrently, the domain of source separation (and audio source separation in particular) has also seen a great interest from the community but with little or no interaction with the audio compression sphere [4]. The general problem of source separation can be described as follows: assume J signals (*the sources*) $\mathbf{S}$ have been mixed through I channels to produce I signals (*the mixtures*) $\mathbf{X}$. The goal of source separation is to estimate the sources $\mathbf{S}$ given their mixtures $\mathbf{X}$. Many advances were recently made in the area of audio source separation [5], [6]. However, the problem remains challenging in the undetermined setting ($I < J$), including the single-channel case ($I = 1$), and for convolutive mixtures [7].

It is now quite clear that audio source separation performances strongly depend on the amount of available prior information about the sources and the mixing process one can introduce in the source separation algorithm. In unsupervised source separation, this information can be under the form of a specific source model (as for example the source/filter model used in [8] for singing voice separation or more generally a composite model from a library of models [6]). However, this information can also be provided by a user [9], [10] or by a partial transcription in the case of music signals (see for example [11]). In the extreme case, this information can be the sources themselves. In these cases, we refer to *informed source separation (ISS)*.

Such so-called ISS schemes were recently developed for the case where both the sources and the mixtures are assumed known during an *encoding* stage [12]–[15]. This knowledge enables the computation of any kind of *side-information* that should be small and should help the source separation at the *decoding* stage, where the sources are no longer assumed to be known. The side-information can be either embedded into the mixtures using watermarking methods [14] or just kept aside. ISS is motivated by a number of specific applications including active listening and remixing of music, karaoke, audio gaming, etc.

Note that the performances of *source separation* and the above-mentioned *conventional ISS* methods, depending on the underlying models and assumptions, are bounded by those

Alexey Ozerov is with Technicolor Research & Innovation, 975 avenue des champs blancs, CS 17616, 35576 Cesson Sévigné, France, e-mail: alexey.ozerov@technicolor.com.

Antoine Liutkus, Roland Badeau, and Gaël Richard are with Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, 37-39, rue Dareau, 75014 Paris, France, email: firstname.lastname@telecom-paristech.fr.

This work was partly supported by the European Commission under contract FP7-ICT-287723 - REVERIE and by the ANR through the DReaM project (ANR-09-CORD-006-03).

of *oracle estimators* [16] [1]. Indeed, since the majority of conventional ISS methods [13], [14] are source separation-inspired and thus fall into the category of parametric coding approaches, they cannot achieve distortions that are better (below) the *oracle distortions* provided by the corresponding oracle estimators whatever the chosen bitrate [2]. In order to outperform the oracle estimators, some hybrid approaches have been developed, which involve waveform source coding. In [15], some sources are encoded using a source coding method and the remaining sources are recovered by a conventional ISS method. However, such a straightforward hybridization does not allow overcoming the above-mentioned drawbacks that are still valid for individual sources.

With regard to the above description, it is quite clear that ISS shares many similarities with the recently introduced Spatial Audio Object Coding (SAOC) (see [17]–[19] and [20] for the ISO/MPEG SAOC standard version). Developed as a multichannel audio compression scheme, SAOC also aims at recovering so called *sound objects* at the decoding side from a transmitted downmix signal and side information about the audio objects. In the literature, different kinds of side information were also considered in the framework of Spatial Audio Coding (SAC), such as the inter and intra-channel correlation [21], spatial coherence cues [22], source localization parameters [23], or a sinusoids plus noise model of the sources [24]. In SAOC [20], high quality remixing is guaranteed by also transmitting perceptually-encoded residual signals resulting from an imperfect object extraction at the encoding side (therefore jointly exploiting waveform coding and parametric coding principles). However, this scheme has a major drawback which limits its potential. Indeed, in SAOC the "separation step" (sound object extraction) is independent of the "residual compression step" while this could be done jointly.

The purpose of this paper is then:

1) to further develop and to present in an even more general manner the novel concept of *Coding-based ISS (CISS)* recently introduced in [25], [26] and to highlight its main theoretic advantage against the approaches followed in both conventional ISS [13], [14] and SAOC [20];
2) to extend the previous "proof of concept" model used in [25] by integrating a more elaborate model based on Non-Negative Tensor Factorization (NTF). [3] We also discuss how the proposed approach relates to other relevant state of the art methods such as non-negative matrix factorization (NMF) or NTF-based coding methods [27], [28], but to the best of our knowledge this is the first attempt of using NTF models with waveform coding principles.
3) and to show that the proposed scheme allows for a smooth transition between low rate object-based parametric coding and high-rate waveform coding relying on the same object-based model (here the NTF model), thus exploiting long-term redundancy.

It is also important to underline that although our model is presented in the ISS framework, it is directly applicable to traditional audio coding or multichannel audio coding (that is without assuming the mixture to be known at the decoder side).

The paper is organized as follows: Section II introduces the general concept of CISS and thoroughly discusses its relation to the state of the art. Then, its particular variant based on NTF (CISS-NTF) is described in details and analyzed in section III in the case of single-channel mixtures with the Mean Squared Error (MSE) criterion for optimisation.

Experimental results are presented in section IV and the conclusions and perspectives are drawn in the final section.

II. Coding-based Informed Source Separation

The general probabilistic framework introduced herein for ISS is called coding-based ISS (CISS). This approach consists in quantizing the sources, as in waveform source coding, while using the *a posteriori* source distribution, given the mixture and some generative probabilistic source model, as in source separation. The quantization can be performed by optimizing the MSE or some perceptually-motivated distortion driven by a perceptual model. In this section the framework is presented in a very general manner, i.e., it is not limited to a particular problem dimensionality (e.g., multichannel or single-channel mixtures), mixing type (e.g., linear instantaneous or convolutive mixture), source model or perceptual model. A particular instance of the framework will be described in the following section III and evaluated in section IV.

Fig. 1 and 2 give very high-level presentations of the state of the art approaches, notably the conventional ISS approaches [13], [14] and the SAOC [17]–[20], where all audio objects are enhanced.[4] In the conventional ISS approaches (Fig. 1), at the encoding stage, a source model parameterized by $\hat{\theta}$ is estimated, given the sources $\mathbf{S}$ and the mixtures $\mathbf{X}$. It is then encoded and transmitted as a side-information yielding its quantized version $\bar{\theta}$. At the decoding stage, the model parameter $\bar{\theta}$ is reconstructed, and the sources $\hat{\mathbf{S}}$ are reconstructed in turn, given $\bar{\theta}$ and the mixture $\mathbf{X}$ (e.g., by Wiener filtering, as in [13]). However, as mentioned in the introduction, the best achievable distortion of such parametric coding approaches is inherently limited.

At a very high level view, the parametric coding part of SAOC approaches (Fig. 2) follows exactly the same scheme as the conventional ISS (Fig. 1), except that the parametric model, called *SAOC parameters*, is different. To achieve a higher quality at the expense of a higher transmission rate, the residuals $\mathbf{S}_r$ of the parametric SAOC reconstruction $\hat{\mathbf{S}}_p$

[1] Given a measure of source separation performance (i.e., a distortion) and a class of source separation approaches (e.g., binary time-frequency masking approaches [16]) specified by some separation parameters, the oracle estimator of the separation parameters is the one leading to the best possible performance (see [16] for more details).

[2] This remark does not concern [12], where the distortion can be always decreased by increasing the size of the corresponding molecular dictionary, which would lead, however, to an excessive rate needed to transmit such a dictionary.

[3] While an NTF model for CISS was already considered in a short study [26] in the multichannel case, here we consider the single-channel case and conduct a more thorough evaluation. Moreover, we provide some theoretical support to the results that were used in [25], [26].

[4] Within this paper, if the contrary is not stated, we always consider SAOC with enhanced audio objects as in [19], [20].

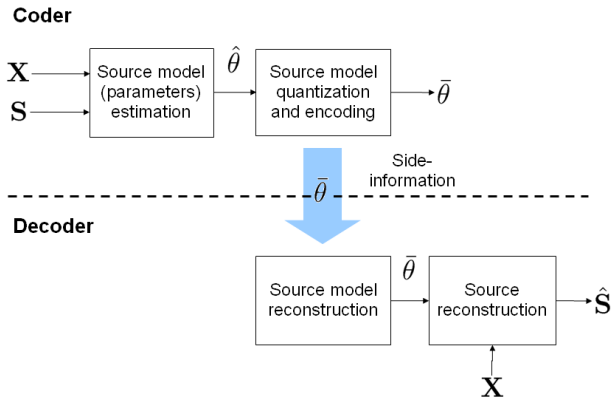


Fig. 1. High level presentation of the conventional ISS [13], [14].

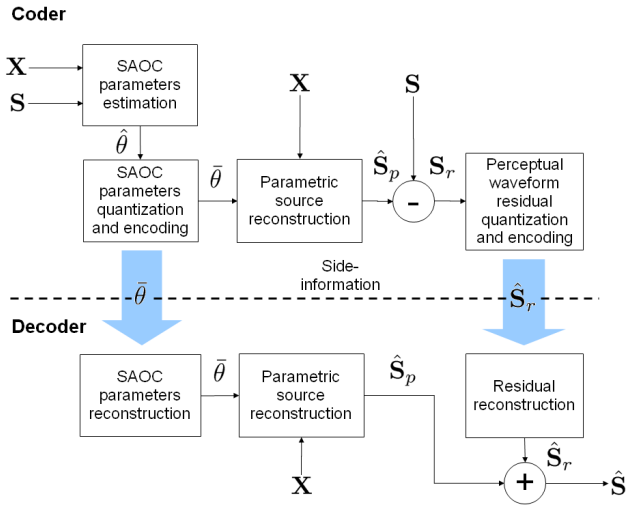


Fig. 2. High level presentation of SAOC (all objects are enhanced) [17]–[20].

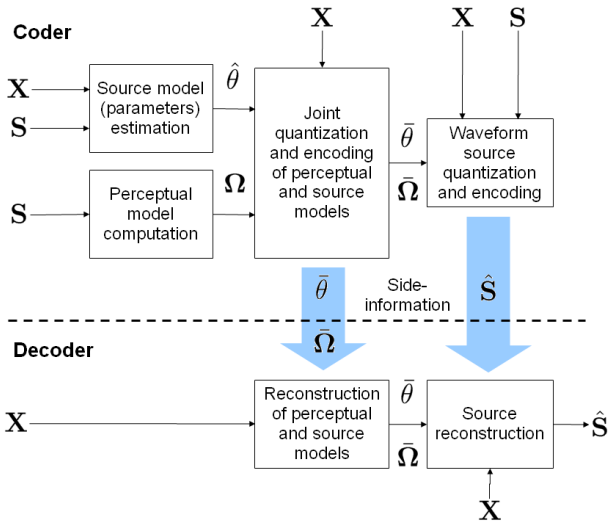


Fig. 3. High level presentation of CISS (proposed).

can be encoded using a perceptual waveform coder yielding $\hat{\mathbf{S}}_r$. However, as we see in Fig. 2, the parametric and waveform coding steps are performed independently and using different models. This is suboptimal since there is no evidence that the residual encoding should be independent of the parametric source encoding.

Fig. 3 gives a high-level representation of the proposed CISS approach. At the encoding stage, the model parameter $\hat{\theta}$ specifying the posterior distribution $p(\mathbf{S}|\mathbf{X},\hat{\theta})$ from a particular family of distributions is estimated, given the sources $\mathbf{S}$ and the mixtures $\mathbf{X}$. A perceptual model $\mathbf{\Omega}$ can be optionally computed as well. $\hat{\theta}$ and $\mathbf{\Omega}$ are then jointly encoded and transmitted [5] as a side-information yielding their quantized versions $\bar{\theta}$ and $\bar{\mathbf{\Omega}}$. This encoding can optionally use the knowledge of the mixtures $\mathbf{X}$. Finally, using the posterior $p(\mathbf{S}|\mathbf{X},\bar{\theta})$ and a perceptual distortion measure driven by $\bar{\mathbf{\Omega}}$ the sources $\mathbf{S}$ are waveform encoded and transmitted as a side-information. This is achieved using a probabilistic model-based quantization and encoding under high-rate theory assumptions, as in [30], [31]. At the decoding stage, the quantized parameters $\bar{\theta}$ and $\bar{\mathbf{\Omega}}$, and then the quantized sources $\hat{\mathbf{S}}$ are reconstructed.

Thus, in contrast to the conventional ISS methods, the CISS framework allows the distortion being unbounded below as in waveform source coding (see Fig. 3 vs. Fig. 1). In other words, CISS can achieve any desirable distortion, given a sufficient bitrate, and in that sense the notion of oracle estimators [16] cannot be extended to CISS. Moreover, in contrast to SAOC, CISS permits, as we will see below, to use more advanced source models that better exploit the redundancy of audio signals, and to use the knowledge of the mixture and model parameters to encode the residuals (see Fig. 3 vs. Fig. 2).

In this work we propose a particular instance of the general CISS framework, referred herein as *CISS-NTF*, that is based on an (object-based) probabilistic NTF source model. Moreover, CISS-NTF is designed for the single-channel case and for the MSE distortion criterion. Investigation of distortions driven by more advanced perceptual models (e.g., those considered in [32]–[34]) is left for a further study.

The major differences of the proposed CISS-NTF approach compared to the state of the art can then be highlighted as follows:

- In contrast to conventional ISS methods [12]–[14], it is based on waveform coding, thus potentially leading to much superior quality for moderate and high rates, as it was already mentioned for CISS in general.
- In contrast to SAOC [20], based on some local parameters (e.g., intra-channel correlation [21] or spatial coherence cues [22]), it exploits advanced source models, i.e., NTF. First, this allows using long-term redundancy of audio

[5]For example, inspired by what is done in the AMR-WB speech coder [29], one way of reconstructing the perceptual model $\mathbf{\Omega}$ at the decoder would be to estimate it, given the source model and the mixture. This approach does not require any extra rate for perceptual model transmission. However, other approaches exist, thus we are more generally speaking about joint encoding and transmission of perceptual and source models.

signals for coding. Second, the parameters used for parametric coding (as in the earlier version of SAOC [18]) and those used for waveform coding (as in [20]) are all computed from the NTF source model and the mixture. Thus, these parameters are coupled (or jointly encoded), while in SAOC [20] they are encoded separately. Moreover, the proposed method exploits posterior correlations between sources (given the mixture), while in SAOC the residuals of the enhanced audio objects are encoded independently.

- In the NMF / NTF-based methods [27], [28] the signal short-time Fourier transform (STFT) (a redundant signal representation) amplitudes are encoded by approximating them with an NMF / NTF decomposition. In [27] the STFT phase is then entropy encoded and the rate (between phase and amplitude encoding) is allocated empirically. Also, the rate between different NMF / NTF model parameters is empirically allocated.
  Besides the fact that we consider a different coding problem, the proposed approach has the following possible advantages over [27], [28]. First, we consider a probabilistic NTF applied to the modified discrete cosine transform (MDCT) or STFT of the sources. As such, we do not split amplitude and phase, but encode them jointly via waveform coding within the corresponding time-frequency representation, while minimizing a target distortion under the constrained entropy. Thus, we consider our approach as a waveform coding-based within the NTF framework. Second, our probabilistic NTF formulation and quantization under high-rate theory assumptions, allows us deriving (under some approximations) analytical expressions for rate allocation between different NTF model parameters which allows avoiding time-consuming empirical parameter optimization. Third, MDCT being a critically sampled signal representation, we show its great advantage over redundant STFT within this application. To our best knowledge NMF / NTF models were not so far applied to MDCT signal representations for compression purposes.

III. Single-channel CISS-NTF with MSE

In this section, we investigate the proposed approach in the case of single-channel mixtures ($I = 1$) using the NTF source model and MSE distortion criterion.

All signals are represented in a real-valued or complex-valued (here, respectively, MDCT or STFT) time-frequency domain. In the time-frequency domain the mixing equation writes

$$x_{fn} = \sum\nolimits_{j=1}^{J} s_{jfn} + b_{fn}, \tag{1}$$

where $j = 1, \ldots, J$, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote, respectively, the source index, the frequency index and the time-frame index; and x_{fn}, s_{jfn} and b_{fn} denote, respectively, the time-frequency coefficients of the mixture, of the sources and of an additive noise. Depending on the particular configuration this additive noise can represent any combination of the following distortions:

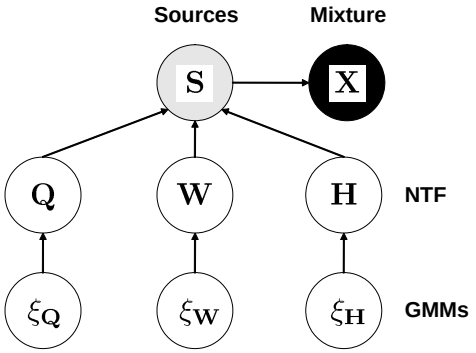


Fig. 4. High-level graphical representation of CISS-NTF probabilistic hierarchical modeling. Shadings of nodes: variables observed at both coder and decoder sides (black), variables observed at the coder side, quantized and transmitted (gray), and parameters estimated at the coder side, quantized and transmitted (white).

1) a background or recording noise, if $\mathbf{X} = \{x_{fn}\}_{f,n}$ is an unquantized mixture of sources $\mathbf{S}_j = \{s_{jfn}\}_{f,n}$ ($j = 1, \ldots, J$),
2) a quantization noise if $\mathbf{X}$ is a quantized version of its clean version $\mathbf{X}^{\text{clean}}$, i.e., $b_{fn} = x_{fn} - x_{fn}^{\text{clean}}$ (e.g., as in SAOC [18], [19]),
3) additional sources $\{\mathbf{S}_j\}_{j=J+1}^{J^*}$ if one is only interested to encode J sources among J^* ($J < J^*$) sources in the mixture.

Fig. 4 gives a high-level graphical representation of CISS-NTF probabilistic hierarchical modeling described in details below. It includes mixture, sources, NTF parameters (Sec. III-A) and Gaussian mixture models (GMM) used to encode these parameters (Sec. III-D2c).

A. NTF source model

As a source model we use the NTF model previously used for source separation in [10] and for ISS in [14]. Its main idea is to assume that the spectrograms of the sources can be considered as the activation over time of some *spectral templates*. To avoid a pre-defined choice for the number of spectral templates for each source, a refinement of the model is to consider a common *pool* of spectral templates jointly approximating all spectrograms of the sources. Such a strategy permits to reduce the number of parameters of the model and to share the same templates for several sources, which may be of interest when there is some kind of redundancy among sources.

Formally, the NTF model can be described as follows. First, the source and noise time-frequency coefficients s_{jfn} and b_{fn} are assumed mutually independent, i.e., over j, f and n, and distributed as follows:

$$s_{jfn} \sim \mathcal{N}_{r/c}(0, v_{jfn}), \qquad b_{fn} \sim \mathcal{N}_{r/c}(0, \sigma_{b,fn}^2), \tag{2}$$

where the distribution $\mathcal{N}_{r/c}(\cdot, \cdot)$ is the standard Gaussian distribution if s_{jfn} is real-valued, or the circular complex Gaussian distribution if it is complex-valued. The source variances v_{jfn} are structured as

$$v_{jfn} = \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk}, \tag{3}$$

with $q_{jk}, w_{fk}, h_{nk} \geq 0$ and the noise variances $\sigma^2_{b,fn}$ are assumed to be known. The noise variances can be either constant and fixed ($\sigma^2_{b,fn} = \sigma^2_b$) to represent a background noise or have a structure similar to that of the source variances to represent a nonstationary noise.

We here assume the noise variances to be constant and fixed. This model can be parameterized as follows

$$\theta = \left\{\mathbf{Q}, \mathbf{W}, \mathbf{H}, \sigma^2_b\right\}, \tag{4}$$

with $\mathbf{Q} = \{q_{jk}\}_{j,k}$, $\mathbf{W} = \{w_{fk}\}_{f,k}$ and $\mathbf{H} = \{h_{nk}\}_{n,k}$ being, respectively, $J \times K$, $F \times K$ and $N \times K$ nonnegative matrices (see Fig. 4).

This model is in fact an object-based approximation of the 3-valence tensor of source power spectra

$$\mathbf{P} \triangleq \{p_{jfn}\}_{j,f,n} \quad (p_{jfn} \triangleq |s_{jfn}|^2) \tag{5}$$

consisting of K objects (rank-1 tensors) that represent individual sounds. Whereas each column of $\mathbf{W}$ stands for one spectral template, its activation over time is given by the corresponding column of $\mathbf{H}$. Finally, the columns of $\mathbf{Q}$ model the possible couplings between both the spectral templates (columns of $\mathbf{W}$) and their temporal activations (columns of $\mathbf{H}$), i.e., different sources can share the same templates together with the corresponding activations. One can see from the example on Fig. 5 (detailed just below) that several of the 9 components are involved in the modeling of the spectral templates and temporal activations of all the 3 sources. Exploiting redundancies over time and over sources appears to be an important feature of the NTF model.

An illustrative example of this NTF modeling is given in Fig. 5, where the first row shows MDCT power spectrograms p_{jfn} (Eq. (5)) of three sources (drums, guitar and singing voice), the second raw shows their structured approximations v_{jfn} (Eq. (3)), and the third raw includes NTF matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$. First, by investigating matrix $\mathbf{Q}$ one can note that among the $K = 9$ components (in average 3 components per source) 7 components were automatically assigned (dark brown color) to each source, while sharing the 6-th component between drums and voice and sharing the 9-th component between all three sources. This last component can be interpreted as the background noise floor that is common to all three sources. Second, one can note that while this is a good approximation (MDCT power spectrograms and their structured approximations look very similar), it drastically reduces the dimensionality (i.e., the number of parameters to be transmitted). Indeed, for this example, instead of $J \times F \times N$ = $3 \times 1024 \times 421$ = 1293312 coefficients p_{jfn}, one have only $(J + F + N) \times K$ = $(3 + 1024 + 421) \times 9$ = 13032 entries of NTF matrices, which divides the number of parameters by 100.

B. Prior and posterior distributions

We give here the expressions for prior and posterior (i.e., given the mixture) source distributions assuming the NTF source model presented above. The posterior distribution is then used for source encoding and the prior one is needed for some derivations presented in section III-D below.

Since the source time-frequency coefficients are modeled as distributed with respect to independent Gaussian distributions, the additive noise is as well assumed Gaussian, and the mixing (1) is linear, the posterior distribution of the sources given the observed mixture is Gaussian, and analytical expression of this distribution is readily obtained. Let $\mathbf{s}_{fn} = [s_{1fn}, \ldots, s_{Jfn}]^T$ be the vector containing the time-frequency coefficients of all sources at bin (f, n). Provided all parameters (4) are available, the prior and posterior distributions of $\mathbf{s}_{fn}$ write, respectively, as [6]

$$p(\mathbf{s}_{fn}|\theta) = N_{r/c}\left(\mathbf{s}_{fn}; \boldsymbol{\mu}^{\mathrm{pr}}_{fn}, \boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\right), \tag{6}$$

$$p(\mathbf{s}_{fn}|x_{fn};\theta) = N_{r/c}\left(\mathbf{s}_{fn}; \boldsymbol{\mu}^{\mathrm{pst}}_{fn}, \boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn}\right), \tag{7}$$

where $N_{r/c}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability density function (pdf) of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ for either real-valued or complex-valued cases; and prior and posterior covariance matrices ($\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}$ and $\boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn}$) and means ($\boldsymbol{\mu}^{\mathrm{pr}}_{fn}$ and $\boldsymbol{\mu}^{\mathrm{pst}}_{fn}$) from (6) and (7) are computed as follows:

$$\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn} = \mathrm{diag}\left[\{v_{jfn}\}_j\right], \qquad \boldsymbol{\mu}^{\mathrm{pr}}_{fn} = 0, \tag{8}$$

$$\boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn} = \left(\mathbf{I}_J - \mathbf{g}_{fn}\mathbf{1}_J\right)\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}, \tag{9}$$

$$\boldsymbol{\mu}^{\mathrm{pst}}_{fn} = \mathbf{g}_{fn}x_{fn}, \tag{10}$$

$$\mathbf{g}_{fn} = \boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\mathbf{1}_J^T\left(\mathbf{1}_J\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\mathbf{1}_J^T + \sigma^2_b\right)^{-1}, \tag{11}$$

with $\mathbf{g}_{fn}$ being the Wiener filter gain, v_{jfn} being NTF source model variances defined by (3), and $\mathbf{I}_J$ and $\mathbf{1}_J$ denoting, respectively, the $J \times J$ identity matrix and the J-length row vector of ones.

C. Source encoding and reconstruction

In this section we explain how the posterior source distribution presented in the previous section is used to encode the sources within the proposed CISS framework.

Given the Gaussian NTF source model outlined above, source coding would amount to encode each source vector $\mathbf{s}_{fn}$ according to its prior distribution (6). The main idea of CISS is to employ exactly the same techniques as in source coding, but to use instead its *posterior* distribution (7).

In the Gaussian case, such an encoding is readily performed through constrained entropy quantization relying on scalar quantization in the mean-removed Karhunen-Loeve transform (KLT) domain, as described in [31]. We summarize below its main steps.

Let $\boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn} = \mathbf{U}_{fn}\boldsymbol{\Lambda}_{fn}\mathbf{U}^H_{fn}$ be the eigenvalue decomposition of the covariance matrix, where $\mathbf{U}_{fn}$ is an orthogonal matrix ($\mathbf{U}^H_{fn}\mathbf{U}_{fn} = \mathbf{I}_J$) and $\boldsymbol{\Lambda}_{fn} = \mathrm{diag}\{\lambda_{1fn}, \ldots, \lambda_{Jfn}\}$ is a diagonal matrix of eigenvalues. The linear transform $\mathbf{U}^H_{fn}$ decorrelating $\mathbf{s}_{fn}$ is the KLT. Assuming the MSE distortion, uniform quantization is asymptotically optimal for the constrained entropy case [35]. Thus, we consider here scalar uniform quantization with a fixed step size Δ in the mean-removed KLT domain, which can be summarized as follows:

1) Remove the mean and apply the KLT

$$\mathbf{y}_{fn} = \mathbf{U}^H_{fn}(\mathbf{s}_{fn} - \boldsymbol{\mu}^{\mathrm{pst}}_{fn}). \tag{12}$$

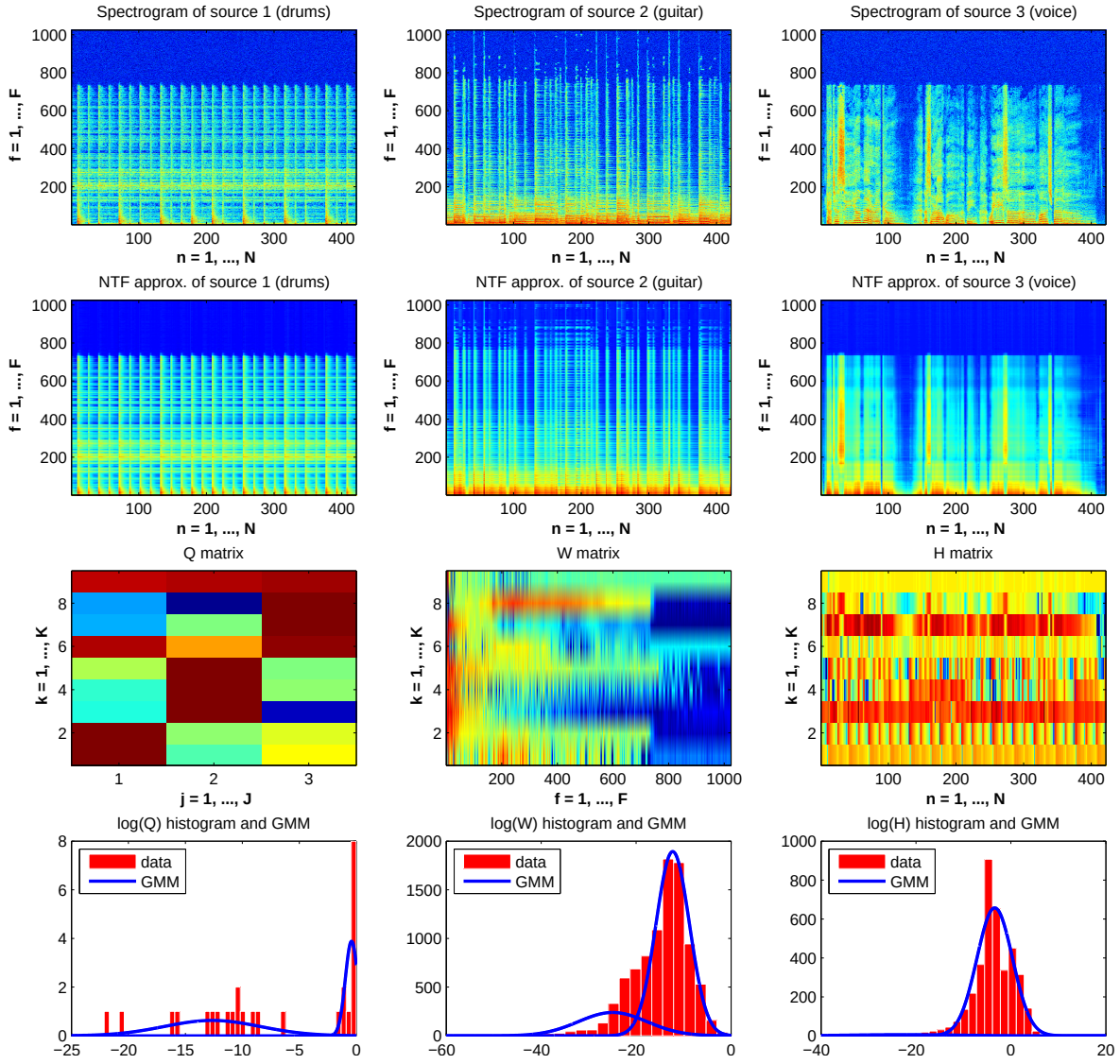


Fig. 5. Source MDCT power spectrograms p_{jfn} (5) (first row), their structured approximations v_{jfn} (3) (second row), NTF matrices (third row), histograms of NTF log-coefficients (*bars*) and two-state GMMs (*solid line*) modeling them (fourth row). In this example $J = 3$, $F = 1024$, $N = 421$ and $K = 9$.

2) In the real-valued case, quantize each dimension of $\mathbf{y}_{fn} = [y_{1fn}, \ldots, y_{Jfn}]^T$ with a uniform scalar quantizer $Q_\Delta : y_{jfn} \rightarrow \hat{y}_{jfn}$ having a constant step size Δ. In the complex-valued case, the same quantization is applied independently to real and imaginary parts of y_{jfn}. Using an arithmetic coder as an entropy coder [31], the effective codeword length (in bits) is given by

$$L(\mathbf{s}_{fn}|x_{fn};\theta) = -\sum_{j=1}^{J} \log_2 \int_{y-\hat{y}_{jfn}\in\mathcal{A}(\Delta)} N_{r/c}(y;0,\lambda_{jfn})dy. \quad (13)$$

where in the real valued case $\mathcal{A}(\Delta) \triangleq [-\Delta/2, \Delta/2]$, and in the complex-valued case $\mathcal{A}(\Delta) \triangleq \{z \in \mathbb{C} | \max(|\Re z|, |\Im z|) \leq \Delta/2\}$.

3) Reconstruct the quantized source vector $\hat{\mathbf{s}}_{fn}$

$$\hat{\mathbf{s}}_{fn} = \mathbf{U}_{fn}\hat{\mathbf{y}}_{fn} + \boldsymbol{\mu}_{fn}^{\text{pst}}. \quad (14)$$

D. Model estimation and encoding

In this section we first detail the strategy for the estimation and quantization of the NTF parameters θ (see Fig. 3). Our derivations mostly follow those from [30]. However, they are applied here to the NTF model instead of the autoregressive model considered in [30]. As highlighted above, the optimal approach would consider posterior distribution (7) for the model estimation and encoding. However, the derivation of the corresponding estimation strategy is overly complex and did not permit us to obtain a simple solution. To simplify this analysis we then assume that the sources were quantized using the prior distribution (6) instead of the posterior one (7). This choice leads us to an optimization strategy based upon some standard algorithms, and we leave the more optimal case of the posterior optimization for further study. Note however that if the posterior distributions are not used in the analysis of model estimation and encoding they are indeed used below for the

residual sources encoding.

1) Model estimation: Under high-rate theory assumptions and given the model parameter θ, the total rate (in bits) required to encode the sources $\mathbf{S} = \{s_{jfn}\}_{j,f,n}$ is [30]

$$R(\mathbf{S}|\theta) = -\log_2 p(\mathbf{S}|\theta) - \frac{\mathcal{M}(J,F,N)}{2}\log_2\frac{D}{C_s}, \quad (15)$$

where $D = C_s\Delta^2$ is the mean distortion (per real-valued dimension), defined as $D \triangleq \mathbb{E}[|\hat{s}_{jfn} - s_{jfn}|^2]$ in the real-valued case and as $D \triangleq (1/2)\mathbb{E}[|\hat{s}_{jfn} - s_{jfn}|^2]$ in the complex-valued case; $C_s = 1/12$ is the coefficient of scalar quantization, and $\mathcal{M}(J,F,N)$ denotes the total number of real-valued coefficients in $\mathbf{S}$, i.e., $\mathcal{M}(J,F,N) \triangleq JFN$ in the real-valued case and $\mathcal{M}(J,F,N) \triangleq 2JFN$ in the complex-valued case. Thus, the model parameter θ should be estimated in the maximum likelihood (ML) sense, as follows

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{S}|\theta) \quad (16)$$

that, in the case of the NTF model, can be shown equivalent to [14], [36]

$$\widehat{\mathbf{Q}}, \widehat{\mathbf{W}}, \widehat{\mathbf{H}} = \arg\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{jfn} d_{IS}\left(p_{jfn}\left|\sum_{k=1}^{K} q_{jk}w_{fk}h_{nk}\right.\right), \quad (17)$$

where p_{jfn} is defined by (5) and $d_{IS}(x|y) = x/y - \log(x/y) - 1$ is the Itakura-Saito (IS) divergence. The optimization of criterion (17) can be achieved by iterating the following multiplicative updates [10], [14], [37]:

$$q_{jk} \leftarrow q_{jk}\left(\frac{\sum_{f,n} w_{fk}h_{nk}p_{jfn}v_{jfn}^{-2}}{\sum_{f,n} w_{fk}h_{nk}v_{jfn}^{-1}}\right), \quad (18)$$

$$w_{fk} \leftarrow w_{fk}\left(\frac{\sum_{j,n} h_{nk}q_{jk}p_{jfn}v_{jfn}^{-2}}{\sum_{j,n} h_{nk}q_{jk}v_{jfn}^{-1}}\right), \quad (19)$$

$$h_{nk} \leftarrow h_{nk}\left(\frac{\sum_{j,f} w_{fk}q_{jk}p_{jfn}v_{jfn}^{-2}}{\sum_{j,f} w_{fk}q_{jk}v_{jfn}^{-1}}\right). \quad (20)$$

2) Model quantization and encoding:

a) Criterion for quantization: Assuming the model parameter quantized and transmitted (Fig. 3), the total rate required to encode the sources becomes [30]

$$R(\mathbf{S}) = \psi(\bar{\theta}, \hat{\theta}, \mathbf{S}) + R(\mathbf{S}|\hat{\theta}), \quad (21)$$

where

$$\psi(\bar{\theta}, \hat{\theta}, \mathbf{S}) \triangleq R(\bar{\theta}) + \log_2\left(p(\mathbf{S}|\hat{\theta})/p(\mathbf{S}|\bar{\theta})\right) \quad (22)$$

is the *index of resolvability* [30] involving the rate required to encode the model $R(\bar{\theta})$ and a term representing the loss in the rate for source encoding due to the usage of the quantized $\bar{\theta}$ model instead of the ideal ML model $\hat{\theta}$. Relying on some realistic approximations (see below) this term can be shown independent of $\mathbf{S}$, and, denoted by $\Psi(\bar{\theta}, \hat{\theta}) \triangleq \log\left(\frac{p(\mathbf{S}|\hat{\theta})}{p(\mathbf{S}|\bar{\theta})}\right)$, while omitting a constant multiplicative term $1/\log(2)$, it can be expressed as

$$\begin{aligned}
\Psi(\hat{\theta}, \bar{\theta}) &= \frac{1}{2}\sum_{j,f,n}\left(\frac{p_{jfn}}{\bar{v}_{jfn}} - \frac{p_{jfn}}{\hat{v}_{jfn}} - \log\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}}\right) && (23)\\
&= \frac{1}{2}\sum_{j,f,n}\left(\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - \log\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - 1\right) + \\
&\quad \frac{1}{2}\sum_{j,f,n}\left(\frac{p_{jfn} - \hat{v}_{jfn}}{\hat{v}_{jfn}}\frac{\hat{v}_{jfn} - \bar{v}_{jfn}}{\bar{v}_{jfn}}\right) && (24)\\
&\approx \frac{1}{2}\sum_{j,f,n}\left(\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - \log\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - 1\right) && (25)\\
&\approx \frac{1}{4}\sum_{j,f,n}\left(\log\hat{v}_{jfn} - \log\bar{v}_{jfn}\right)^2, && (26)
\end{aligned}$$

where approximation (25) follows from a reasonable assumption that the relative error of modeling $(p_{jfn} - \hat{v}_{jfn})/\hat{v}_{jfn}$ and that of quantization $(\hat{v}_{jfn} - \bar{v}_{jfn})/\bar{v}_{jfn}$ are uncorrelated [30] and at least one of these errors is zero-mean.

The last approximation (26) is obtained using the following second order Taylor expansion $u \approx 1 + \log(u) + \frac{1}{2}\log(u)^2$ in the neighborhood of $u = 1$ (with $u = \hat{v}_{jfn}/\bar{v}_{jfn}$), as in [30], [38]. Note that we find again the IS divergence in the expression (25), and the last approximation (26) indicates that the NTF model variances $\hat{v}_{jfn}$, structured as in (3), should be quantized by minimizing the MSE of their logarithms.

This result is quite similar to what was done in [14], where the log-spectrograms were compressed using the JPEG image coder. However, while [14] does not justify this particular choice, we provide here a theoretical explanation of its appropriateness.

b) NTF parameters quantization: Although the criterion (26) is quite simple, it does not give yet any precise idea of how to quantize individual NTF model parameters, i.e., matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$. Using (3), the criterion (26) can be rewritten as

$$\Psi(\hat{\theta}, \bar{\theta}) \approx \frac{1}{4}\sum_{j,f,n}\left(\log\sum_{k=1}^{K}\hat{q}_{jk}\hat{w}_{fk}\hat{h}_{nk} - \log\sum_{k=1}^{K}\bar{q}_{jk}\bar{w}_{fk}\bar{h}_{nk}\right)^2. \quad (27)$$

We see that there are quite complicated dependencies between elements of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ in this criterion. To simplify this expression we consider the following criterion

$$\Phi(\hat{\theta}, \bar{\theta}) = \frac{1}{4}\sum_{j,f,n}\sum_{k}\left(\log\hat{q}_{jk}\hat{w}_{fk}\hat{h}_{nk} - \log\bar{q}_{jk}\bar{w}_{fk}\bar{h}_{nk}\right)^2 \quad (28)$$

that is in fact an upper bound of (27), i.e.,

$$\Psi(\hat{\theta}, \bar{\theta}) \leq \Phi(\hat{\theta}, \bar{\theta}), \quad (29)$$

which can be shown by applying Lemma A.1 from Appendix A with $c = 1$ and $f(u) = \log(u)^2$. Note however that this upper bound is not very tight, as it can be seen from the proof of Lemma A.1.

Now, assuming that the entries of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are quantized independently the cross-terms in (28) will be canceled in average (if $K \times \min(J,F,N)$ is big enough), due to the fact that the quantization noise of say $\mathbf{Q}$ will be independent

of (thus decorrelated with) that of say $\mathbf{W}$. Thus, (28) can be rewritten

$$\Phi(\hat{\theta},\bar{\theta}) = \frac{1}{4}\sum_{j,f,n}\sum_{k}\Big[(\log\hat{q}_{jk}-\log\bar{q}_{jk})^2 + (\log\hat{w}_{fk}-\log\bar{w}_{fk})^2 + \Big(\log\hat{h}_{nk}-\log\bar{h}_{nk}\Big)^2\Big] = \frac{JFN}{4}\sum_{k}\Bigg[\frac{1}{J}\sum_{j}(\log\hat{q}_{jk}-\log\bar{q}_{jk})^2 + \frac{1}{F}\sum_{f}(\log\hat{w}_{fk}-\log\bar{w}_{fk})^2 + \frac{1}{N}\sum_{n}\Big(\log\hat{h}_{nk}-\log\bar{h}_{nk}\Big)^2\Bigg]. \quad (30)$$

Under all approximations above, we conclude that, if we choose to independently quantize NTF coefficients under an entropy constraint, we should use scalar quantizers of their logarithms. Thus, we opt for a logarithmic compressor, followed by a scalar quantizer and an exponential expander. It is interesting to note that Nikunen *et al.* [27], [28] use μ-law compressor and expander to quantize NTF / NMF coefficients, and the μ-law compressor also acts as logarithmic for high values. Note finally that our NTF model has a different goal than the one presented in [27], [28]. The NTF considered in [27], [28] models both the source and the perception, while our goal is to model source distribution only, and we propose addressing perceptual aspects separately (see Fig. 3). Thus, given different modeling goals the ways the NTF parameters are quantized may be different as well.

We see that squared log-differences of different NTF parameters appear with different weights in the summation of (30). Thus, in order to have the MSE over all parameters, the parameters, up to the same uniform quantization, should be divided by the square roots of these weights, or, equivalently, they should be quantized with different step-sizes $\Delta_{\mathbf{Q}}$, $\Delta_{\mathbf{W}}$ and $\Delta_{\mathbf{H}}$ (respectively, to quantize logarithms of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$) computed as follows

$$\Delta_{\mathbf{Q}} = \sqrt{J/(J+F+N)}\cdot\Delta_\theta, \quad (31)$$

$$\Delta_{\mathbf{W}} = \sqrt{F/(J+F+N)}\cdot\Delta_\theta, \quad (32)$$

$$\Delta_{\mathbf{H}} = \sqrt{N/(J+F+N)}\cdot\Delta_\theta, \quad (33)$$

where Δ_θ is some global model quantization step-size governing the rate-distortion trade-off. We see that within our framework (that is based on high-rate theory) we are able to find an analytical solution for the allocation of the rate between different NTF parameters, while in [27], [28] such an allocation was established experimentally. Thus, our approach has the following advantages over [27], [28]. First, it permits to considerably reduce the number of parameters to be optimized experimentally. Second, we show that the rate allocation between NTF parameters depends on the NTF dimensions J, F and N, and, as a consequence it depends, e.g., on the length of the signal to be encoded and on the number of sources. Thus, we show that even if an experimental optimization of this rate allocation is followed, it should be performed again every time one of these parameters (e.g., signal length) changes.

c) NTF parameters encoding by GMMs: In order to quantize each of the three NTF matrices we model the distribution of its log-coefficients by a two-state Gaussian mixture model (GMM) (see the fourth row of Fig. 5). GMMs are denoted $\xi_{\mathbf{Q}}$, $\xi_{\mathbf{W}}$ and $\xi_{\mathbf{H}}$ (see Fig. 4) and optimized in the ML sense for each matrix, thus their parameters must be transmitted resulting in a very small extra rate (there are only 15 parameters, i.e., 5 parameters per matrix: two means, two variances and one weight). As an alternative the Huffman coding can be used as well, as it is done in [27], [28]. There are pros and cons for using Huffman coding. From the one hand, it is optimal. From the other hand, it requires transmitting a codebook to the decoder, which can be more costly, as compared to transmitting just the five parameters of a GMM.

E. Operational rate-distortion function and parameter optimization

Now we write a so-called *operational rate-distortion function (RDF)* [39] that is accurate for high rates and gives a practical relation between rate and distortion for our CISS coding scheme. Considering (15), but now with posterior $p(\mathbf{S}|\mathbf{X},\theta)$ instead of prior $p(\mathbf{S}|\theta)$, and adding to it the rate required to encode the model parameter $R(\bar{\theta})$, one can show that the total rate (in bits) R_{tot} relates to the mean distortion (per dimension) as

$$R_{\mathrm{tot}} = -\frac{\mathcal{M}(J,F,N)}{2}\log_2\frac{D}{C_s} + \eta(\mathbf{S},\mathbf{X},\bar{\theta}), \quad (34)$$

with

$$\eta(\mathbf{S},\mathbf{X},\bar{\theta}) \triangleq R(\bar{\theta}) - \log_2 p(\mathbf{S}|\mathbf{X},\bar{\theta}), \quad (35)$$

that is independent [6] of the rate R_{tot} and distortion D. Thus, in order to optimize operational RDF (34) for any high rate, one needs to minimize (35).

The only free parameters we need to optimize experimentally are the model quantization step-size Δ_θ (determining the model rate $R(\bar{\theta})$) and the number of NTF components K. We optimize these parameters so as to minimize $\eta(\mathbf{S},\mathbf{X},\bar{\theta})$ from (35). These parameters can be either optimized globally for a set of signals, or they can be re-optimized for each signal to be encoded. In the last case the parameters must be quantized and transmitted to the decoder.

IV. Experiments

In this section we evaluate the proposed single-channel CISS-NTF method for both STFT and MDCT representations. This evaluation includes the optimization of different parameters and the comparison with relevant state of the art methods.

A. State of the art methods

As for conventional ISS, we consider two state of the art methods proposed in [14], [40]. Both methods are based on a parametric reconstruction of the sources via Wiener filtering in the STFT domain, while the source spectrograms (the variances used to compute Wiener filter) are encoded

[6] We know from [30] that, under high-rate theory assumptions, the optimal model rate $R(\bar{\theta})$ is constant, thus independent on the total rate R_{tot}.

differently. In the first method, referred to as *Wiener-JPEG*, the images of source log-spectrograms are encoded by the JPEG lossy coder. In the second method, referred to as *Wiener-NTF*, source spectrograms are approximated by exactly the same NTF model as the one considered here. Parvaix *et al.* [12] introduced another conventional ISS method that is suitable for single channel mixtures. This method is based on binary masking of sources in the MDCT domain, while it is known [16] that oracle bounds of binary masking-based methods are lower than those of Wiener filter-based methods [14], [40]. Another conventional ISS method that is suitable for single channel mixtures is the ISS using iterative reconstruction (ISSIR) by Sturmel and Daudet [41] (see also [42]). ISSIR permits to benefit from phase consistency constraints in the case of STFT representations to reach better performance than Wiener filtering in the case of mono mixtures. However, in the case of MDCT, there is no such constraint that can be exploited to improve performance of filtering techniques and we have thus chosen not to include ISSIR in our evaluation. Thus, we here consider only Wiener filter-based methods for comparison.

B. Testing methodology

1) Data: We considered seven single-channel mixtures of several musical sources such as singing voice, bass, guitar, piano, distorted guitars, etc ... The number of sources J varies from 3 to 6, and the duration of each mixture is about 20 seconds. All signals are sampled at either 48kHz or 44.1kHz. For each mixture the sources were obtained by summing up stereo source images from the QUASI database [7] and by restricting them to a desired time duration. Sources from the same artist were never included into different mixtures.

2) Parameters: MDCT and STFT were computed with frames of 2048 samples and 50 % overlap for STFT. Note however that due to STFT redundancy, as compared to MDCT, this representation includes twice as many real-valued coefficients $\mathcal{M}(J,F,N)$ to be encoded.

3) Evaluation metrics: Since ISS is an emerging research area lying in between source separation and lossy audio coding, we used evaluation metrics coming from these two fields. Notably, we used signal-to-distortion ratio (SDR) [43] usually used to evaluate source separation algorithms and perceptual similarity measure (PSM) of PEMO-Q [44] usually used to evaluate perceptual quality of lossy audio coding schemes. We used the implementation provided by [45] for this purpose.

In most experiments presented below we do not consider directly SDR and PSM but rather the improvements of these measures, denoted as $\delta\mathrm{SDR}$ and $\delta\mathrm{PSM}$, over the corresponding measures computed for the oracle Wiener filtering source estimates [8] in the STFT domain. These oracle performances are shown in Fig. 8 for each mixture from test dataset.

[7] http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

[8] The oracle Wiener filtering source estimates are computed by equation (10), where the structured prior source variances v_{jfn} in (8) are replaced by the true source power spectrograms $p_{jfn} = |s_{jfn}|^2$.

C. Simulations

1) High-rate optimal parameters: As it is explained in section III-E, for high rates, the optimal model quantization step size Δ_θ and the optimal number of NTF components K must be constant, i.e., independent of the total rate. To find these optimal parameters in the case of the STFT representation we have computed $\eta(\mathbf{S},\mathbf{X},\bar{\theta})$ from (35) for each mixture for different combinations of model quantization step sizes $\Delta_\theta = [1.8, 0.5, 0.13, 0.04, 0.01]$ and numbers of NTF components per source $K/J = [2,3,4,5,10,15,20,30]$, and we averaged the result over all mixtures. We observed that the average $\eta(\mathbf{S},\mathbf{X},\bar{\theta})$ reaches its minimum for $\Delta_\theta = 0.13$ and $K/J = 4$, which are thus in average the optimal parameters for high rates. These results, i.e. in average 4 NTF components per source, are in fact consistent with what was found in [46], where a similar modeling was considered for conventional source separation.

2) CISS-NTF with STFT and different ways of optimizing the parameters: The parameters Δ_θ and $K/J = 4$, that have been found optimal in the previous section, are only optimal for high rates and in average. Thus, first, it could be that for some low rates (that can be attractive in practice) the optimal parameters are different. Second, it could be that the optimal parameters, especially the optimal number of NTF components per source K/J, varies from one mixture to another. Indeed, intuitively it seems that a mixture composed of "simple" sources (e.g., triangle) should require less NTF components than a mixture composed of "complex" sources (e.g., organ). The goal of the following experiments is to clarify these points by first evaluating the proposed CISS-NTF for different parameters and over a range of rates, and then by investigating and comparing the optimal parameters for low/high rates and for different mixtures.

We first consider CISS-NTF in the STFT domain, and address the MDCT domain later. This is because the state of the art approaches were designed for STFT domain, and we would like to investigate the possible advantage of CISS-NTF over the state of the art besides the change of the signal representation considered. We have evaluated the CISS-NTF over the same different parameters Δ_θ and K/J as in the previous section, and over a wide range of rates by using 10 logarithmically-spaced values for the source quantization step size as $\Delta = \mathrm{logspace}\,(-0.15, 2.5, 10)$. The source quantization step size $\Delta = +\infty$ has also been tested and corresponds to simply omitting the "waveform source encoding" block in CISS (Fig. 3), so that it essentially becomes a conventional ISS approach (Fig. 1). However, this scheme is still different from Wiener-NTF approach of [14], [40], since in our approach NTF parameters are quantized in log-domain with any step size Δ_θ, while in [14], [40] it was proposed to quantize NTF parameters in the linear domain with a fixed small step size.

The simulations described above gave us many (rate, $\delta\mathrm{SDR}$) pairs, for which we have also computed $\delta\mathrm{PSM}$. Then, for each small range of rates we have chosen (under certain constraints, as described below) the pairs corresponding to the highest $\delta\mathrm{SDR}$. The resulting points in (rate, $\delta\mathrm{SDR}$) and (rate, $\delta\mathrm{PSM}$) planes were then smoothed using the locally

weighted scatterplot smoothing (LOESS) method to produce the rate/performance curves. We have computed the following curves:

- **[Opt-HR-avg]** the same parameters (i.e., Δ_θ and K/J) for all rates and all mixtures optimized for high-rates (i.e., exactly as in section IV-C1),
- **[Opt-LR-avg]** the same parameters for all rates and all mixtures optimized for low-rates (0.5-2 kbps per source),
- **[Opt-HR-mix]** parameters constant over rates, but optimized for each particular mixture for high-rates,
- **[Opt-LR-mix]** parameters constant over rates, but optimized for each particular mixture for low-rates,
- **[Opt-System]** parameters systematically optimized to a particular rate and a particular mixture,

and we have plotted them in Fig. 6. This figure includes as well the results of Wiener-NTF [14], [40] state of the art method and the results of the so called *Wiener-NTF-log-quant* method that is similar to Wiener-NTF, but using newly-proposed log-domain NTF parameters quantization, i.e., with $\Delta = +\infty$.

One can note from Fig. 6 that Wiener-NTF-log-quant outperforms Wiener-NTF for the SDR metric for all rates. That shows the advantage of the proposed log-quantization of NTF parameters over the state of the art [14], [40]. Moreover, waveform source quantization of CISS brings further a great advantage over Wiener-NTF, outperforming it by a large margin for all rates. Also, it outperforms the oracle Wiener results (zero levels of $\delta\mathrm{SDR}$ and $\delta\mathrm{PSM}$ measures) starting from 1-2 kbps per source for SDR, and starting from 7-10 kbps per source for PSM. Note also that the performances of CISS-NTF obtained with parameters optimized for each mixture and/or each particular rate are not much better than the performances with fixed parameters (optimized in average for low or high rates). This is a very good news for a practical coder implementation. Indeed, that means that one does not need to adjust Δ_θ and K/J to each particular mixture, and can just keep them fixed. Finally, it should be noted that for PSM, high-rate optimized parameters (in terms of SDR) are better for low-rates than low-rate optimized parameters (in terms of SDR). This observation indicates a possible use of the distribution preserving quantization (DPQ) [47] to better model perceptual quality.

3) CISS-NTF with MDCT and STFT vs. the state-of-the-art: We have performed for CISS-NTF with MDCT exactly the same simulations as for CISS-NTF with STFT. The qualitative behavior of the results with different ways of optimizing the parameters was exactly the same as for CISS-NTF with STFT, as reported in the previous section. Thus, for these results we can draw exactly the same conclusions as in the previous section for STFT, and we here show in Fig. 7 the results with average parameters optimized for high/low-rates (**[Opt-HR-avg]** and **[Opt-LR-avg]**) for both STFT and MDCT. We have also added the results of the two state of the art methods: Wiener-NTF and JPEG-NTF [14], [40]. We see that CISS-NTF with MDCT outperforms CISS-NTF with STFT for very low rates. This improvement is mostly due to the fact that the MDCT representation is critically sampled, i.e., includes as many coefficients as the time-domain signal, while the STFT is redundant. However, for higher bitrates CISS-NTF with STFT becomes superior, and we explain that as follows. If the signal is a real stationary Gaussian process, then both the MDCT and STFT spectral coefficients are asymptotically independent and distributed with respect to a centered Gaussian distribution. Since MDCT is critically sampled, its performance should be superior to that of STFT. Still, this was not observed during our experiments, since STFT seems to be more efficient at high bitrates. One interpretation of this phenomenon is that MDCT is not shift invariant and may hence be more sensible to the use of short frames than STFT when computing an estimate of the power spectral density. Still, this is only a hypothesis for now and we are currently investigating on this issue.

In any case, using CISS-NTF with STFT is attractive in the multichannel case, and this is what we have done in [26]. Indeed, most of probabilistic multichannel models in source separation involving convolutive mixing [5], [6] are specified in the STFT domain.

4) Summary of results: Fig. 8 gives a summary of the results for each mixture obtained by the oracle Wiener filtering, two state of the art methods (Wiener-NTF and JPEG-NTF), the proposed CISS-NTF with MDCT, and a version of the AAC standard coder [2] available at [9] that was applied independently to each source. The results are now presented in terms of SDR and PSM absolute values (not their increments δ as before) and for an average bitrate of 6 kbps per source, which is attractive for practical applications. We observe on this figure that the proposed method largely outperforms state of the art, while it uses a smaller bitrate. Note also that the proposed method outperforms for all experts the AAC coder, which does not rely on the mixture information, while using a more than twice as small bitrate (3.7 kbps/source instead of 8.3 kbps/source for AAC).

V. Conclusion

We have introduced CISS, a general probabilistic framework for ISS and SAOC. We have further detailed and evaluated in the single-channel mixture case its particular instance called CISS-NTF based on a probabilistic NTF source representation. This approach relates at the same time to different state of the art areas, notably ISS [13], [14], SAOC [17]–[19] and NTF / NMF model-based audio compression [27], [28]. We have discussed possible advantages of CISS in general and of its particular instance, CISS-NTF, over all these state of the art approaches. In summary, without going into details, the main advantages of CISS and CISS-NTF are:

1) waveform quantization based on a structural probabilistic source model (NTF) allowing modeling long-term redundancy in audio signals;
2) in contrast to the conventional ISS and SAOC methods, the parameters used for parametric and waveform coding are jointly encoded within this probabilistic model;
3) the proposed probabilistic formulation allows using NTF / NMF models specified over critically sampled signal representations such as the MDCT, which are known more efficient for compression;

[9] http://www.nero.com/enu/technologies-aac-codec.html

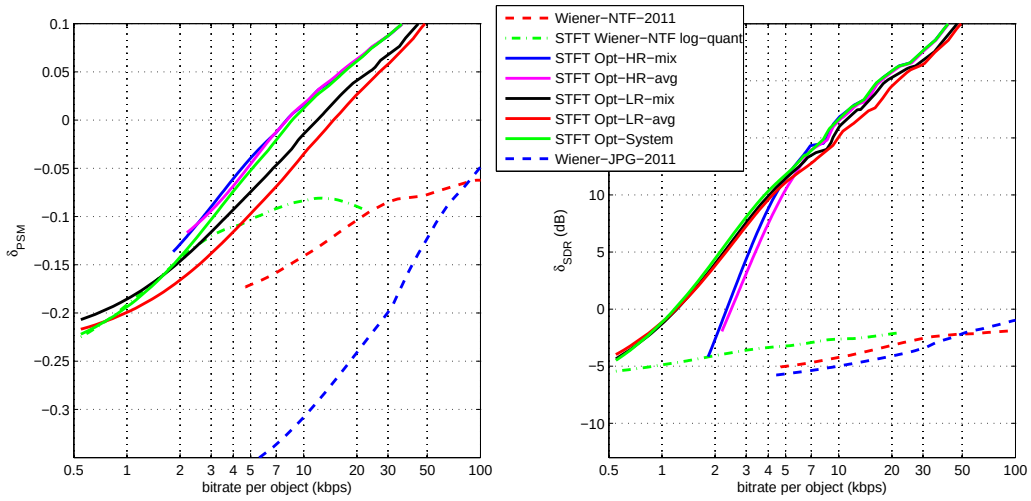


Fig. 6. CISS-NTF with STFT and different ways of optimizing parameters, compared to state of the art. δSDR and δPSM denote the improvements over the corresponding measures computed for the oracle Wiener filtering source estimates in the STFT domain.

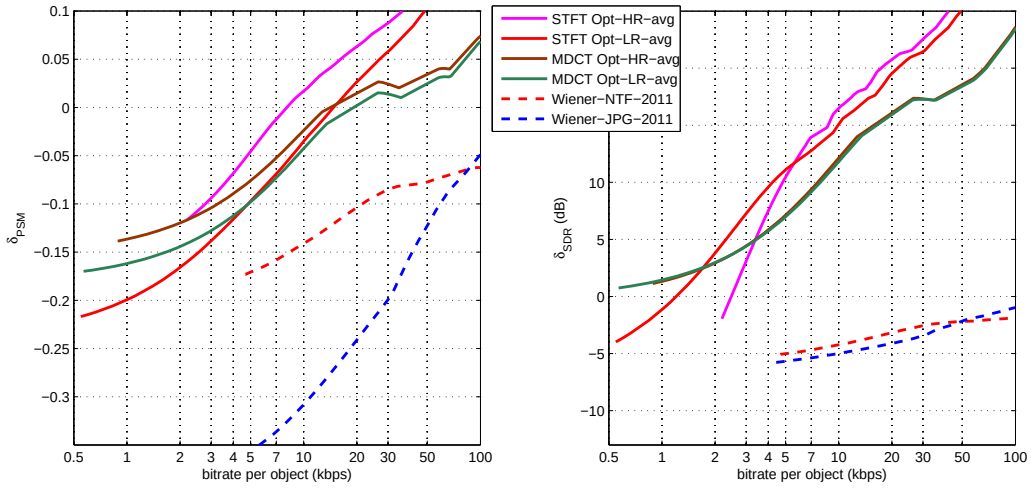


Fig. 7. CISS-NTF with MDCT and STFT vs. the state-of-the-art. Evaluation was performed using both STFT and MDCT transforms. δSDR and δPSM denote the improvements over the corresponding measures computed for the oracle Wiener filtering source estimates in the STFT domain.

4) in contrast to the conventional ISS methods, even if it was not yet implemented, the proposed CISS allows using advanced perceptual models for enhanced perceived quality.

Our extensive experimental evaluation has shown a great advantage of the proposed CISS-NTF approach over the state of the art conventional ISS methods.

This work opens the doors for various further investigations. First, given that most music recordings nowadays are at least stereo, CISS-NTF should be extended to the multichannel case [48] in order to improve its efficiency due to the spatial source diversity. Recent work covering punctual and low-reverberant sources and using STFT signal representation being already done in this direction [26], some questions remain still open. Notably, how to model non-punctual and highly reverberant sources and how to cope with STFT redundancy that is undesirable within compression applications either by reducing STFT overlap or by resorting to critically sampled transforms such as MDCT (see discussion in Sec IV-C3). Second, perceptual modeling should be integrated within CISS-NTF and it should be compared with SAOC through both objective measures and listening tests, when an optimized encoder for this emerging standard is available. The sensitivity matrix approach [34] combined with the newly introduced distribution preserving quantization (DPQ) [47] (see also discussion in Sec IV-C2) seem to be good candidates for modeling perception within this Gaussian model-based approach. Third, remember that in order to simplify the optimization we have chosen here a *generative* model estimation approach optimizing the prior distribution (6) instead of a *discriminative* model estimation optimizing the posterior (7), which is optimal (see Sec. III-D). Thus, new model estimation algorithms should be

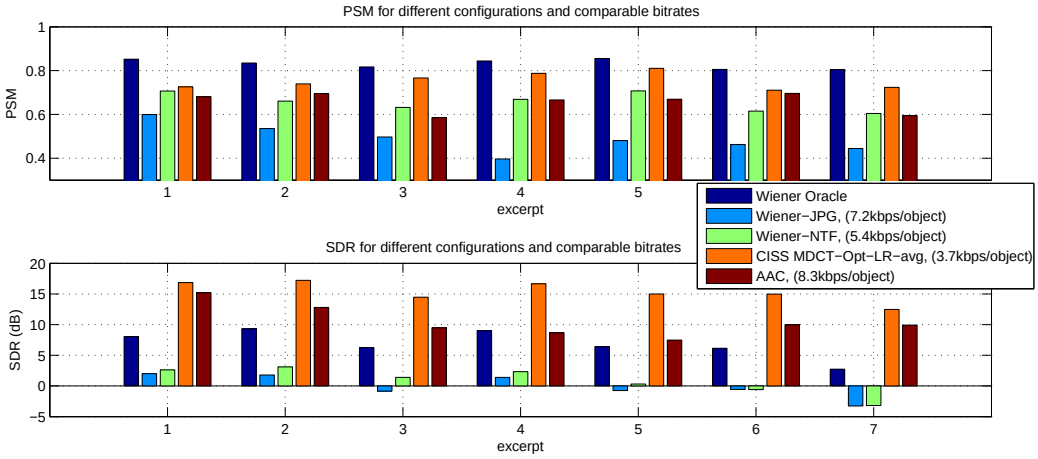


Fig. 8. Summary of results for all 7 excerpts of the database. For each excerpt, the SDR and PSM scores of Oracle source separation is compared to those of state of the art and of the proposed method. CISS-NTF largely outperforms all other techniques, for a smaller bitrate.

proposed to implement the discriminative approach. Moreover, the NTF source model can be replaced by possibly better structured probabilistic models to improve coding efficiency. In fact, any model from those implementable by a general source separation framework presented in [6] can be used in principle. Finally, while SAOC is able to encode and decode sources online, the proposed CISS-NTF requires the whole audio sequence to be analysed for encoding and only decoding can be performed online. This drawback could be overcome by using incremental NMF approaches [49] or other approaches suitable for online audio source separation [50].

More generally, besides ISS and SAOC applications, and in line with [27], [28], the proposed NTF-based approach (with some modifications) could be applied for regular and multichannel audio coding. Moreover, our approach is related to the context-based adaptive entropy coding schemes used for audio and video compression [51], [52]. However, our approach seems to be "more locally adaptive", since each frame is encoded by its own arithmetic coder having a distribution derived from local signal statistics. In other words, each frame has its own context. Thus, it would be interesting to extend such kind of advanced statistical model-based approaches for image or video compression.

APPENDIX A
ONE LEMMA

Lemma A.1. *Let $K \in \mathbb{N}$ and $c \in \mathbb{R}_+^*$. Let $f : \mathbb{R}_+^* \to \mathbb{R}$ a continuous function, that is strictly decreasing on $]0, c[$ and strictly increasing on $]c, +\infty[$.*

Then $\forall\ \hat{x}_1 \dots \hat{x}_K,\ \bar{x}_1 \dots \bar{x}_K \in \mathbb{R}_+^*$,

$$f\left(\frac{\sum_{k=1}^K \hat{x}_k}{\sum_{k=1}^K \bar{x}_k}\right) \leq \sum_{k=1}^K f\left(\frac{\hat{x}_k}{\bar{x}_k}\right). \tag{36}$$

Proof: We assume that $\forall k \in \{1 \dots K\}$, $u_k = \frac{\hat{x}_k}{\bar{x}_k}$, $\lambda_k = \frac{\bar{x}_k}{\sum_{k'=1}^K \bar{x}_{k'}}$ and $u = \sum_{k=1}^K \lambda_k u_k = \frac{\sum_{k=1}^K \hat{x}_k}{\sum_{k=1}^K \bar{x}_k}$. With these notations we need to prove that $f(u) \leq \sum_{k=1}^K f(u_k)$.

Since f is continuous, strictly decreasing on $]0, c[$ and strictly increasing on $]c, +\infty[$, it is clear that it reaches its maximum on any interval of the form $[a, b]$ (with $0 < a < b < +\infty$), and this maximum is reached either in a or in b.

We then define $a = \min(u_1 \dots u_K)$ and $b = \max(u_1 \dots u_K)$. Since $\forall k$, $\sum_{k=1}^K \lambda_k = 1$, it is clear that $u \in [a, b]$. Thus, we conclude that $f(u) \leq \max(f(a), f(b)) \leq \sum_{k=1}^K f(u_k)$. ■

REFERENCES

[1] MPEG-1 Audio, Layer III, "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – Part 3: Audio," *ISO/IEC 11172-3:1993*, 1993.

[2] MPEG-2 Advanced Audio Coding, AAC, "Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio," *ISO/IEC 13818-3:1998*, 1998.

[3] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, Apr. 2009, pp. 1 –4.

[4] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*. Academic Press, 2010.

[5] E. Vincent, M. 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*. IGI Global, 2010, ch. 7, pp. 162–185.

[6] 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, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[7] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[8] 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.

[9] P. Smaragdis and G. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Oct. 2009, pp. 69 –72.

[10] 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 (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[11] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[12] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1464–1475, 2010.

[13] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[14] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[15] M. Parvaix, L. Girin, L. Daudet, J. Pinel, and C. Baras, "Hybrid coding/indexing strategy for informed source separation of linear instantaneous under-determined audio mixtures," in *20th International Congress on Acoustics (ICA 2010)*, Sydney, Australia, Aug. 2010.

[16] E. Vincent, R. Gribonval, and M. Pumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933 – 1950, Aug. 2007.

[17] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in *IEEE International Conference on Multimedia and Expo (ICME 2007)*, Beijing, China, Jul. 2007, pp. 1894–1897.

[18] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," in *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008.

[19] C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," in *13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, Sep. 2010.

[20] O. Hellmuth, H. Purnhagen, J. Koppens, J. Herre, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG spatial audio object coding - The ISO/MPEG standard for efficient coding of interactive audio scenes," in *129th Audio Engineering Society Convention (AES 2010)*, 2010.

[21] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High-fidelity multichannel audio coding with Karhunen-Loève transform," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 365–380, Jul. 2003.

[22] C. Faller, "Parametric multichannel audio coding: Synthesis of coherence cues," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 299–310, Jan. 2006.

[23] B. Cheng, C. Ritz, and I. Burnett, "Encoding independent sources in spatially squeezed surround audio coding," in *8th Pacific Rim Conference on Multimedia (PCM'07)*, Hong Kong, China, Dec. 2007, pp. 804–813.

[24] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "A multichannel sinusoidal model applied to spot microphone signals for immersive audio," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 8, pp. 1483–1497, Nov. 2009.

[25] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New Paltz, New York, USA, Oct. 2011, pp. 257–260.

[26] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *EUSIPCO, 20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[27] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *128th Audio Engineering Society Convention (AES 2010)*, London, UK, May 2010.

[28] J. Nikunen, T. Virtanen, and M. Vilermo, "Multichannel audio upmixing based on non-negative tensor factorization representation," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New Paltz, New York, USA, Oct. 2011, pp. 33–36.

[29] 3GPP TS 26.190, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec," 2005, technical specification.

[30] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, New Paltz, New York, USA, Oct. 2007, pp. 243–246.

[31] D. Y. Zhao, J. Samuelsson, and M. Nilsson, "On entropy-constrained vector quantization using Gaussian mixture models," *IEEE Trans. Commun.*, vol. 56, no. 12, pp. 2094–2104, Dec. 2008.

[32] T. Dau, D. Pušchel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, pp. 3615–3622, 1996.

[33] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[34] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 310–319, 2007.

[35] R. M. Gray, *Source coding theory*. Kluwer Academic Press, 1990.

[36] 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, no. 3, pp. 793–830, Mar. 2009.

[37] 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 (CMMR 2010)*, 2010.

[38] A. Buzo, A. Gray, R. Gray, and J. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 5, pp. 562–574, 1980.

[39] A. Ozerov and W. B. Kleijn, "Asymptotically optimal model estimation for quantization," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 1031–1042, Apr. 2011.

[40] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, St Malo, France, 2010.

[41] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 178 – 185, 2013.

[42] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation: a comparative study," in *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[43] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[44] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902 – 1911, 2006.

[45] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.

[46] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[47] M. Li, J. Klejsa, and W. B. Kleijn, "Distribution preserving quantization with dithering and transformation," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, 2010.

[48] 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 condition," in *Proc. of the 132th Audio Engineering Society Conv.*, Budapest, Hungary, 2012.

[49] S. Bucak and B. Gunsel, "Incremental subspace learning via non-negative matrix factorization," *Pattern Recognition*, vol. 42, no. 5, pp. 788–797, May 2009.

[50] L. S. R. Simon and E. Vincent, "A general framework for online audio source separation," in *International conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, Mar. 2012.

[51] K. Lakhdhar and R. Lefebvre, "Context-based adaptive arithmetic encoding of EAVQ indices," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1473 – 1481, Jul. 2012.

[52] D. Marpe, H. Schwarz, and T.Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 620–644, Jul. 2003.

Alexey Ozerov holds a Ph.D. in Signal Processing from the University of Rennes 1 (France). He worked towards this degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, he received an M.Sc. degree in Mathematics from the Saint-Petersburg State University (Russia) in 1999 and an M.Sc. degree in Applied Mathematics from the University of Bordeaux 1 (France) in 2003. From 1999 to 2002, Alexey worked at Terayon Communicational Systems (USA) as a R&D software engineer, first in Saint-Petersburg and then in Prague (Czech Republic). He was for one year (2007) in Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, for one year and half (2008-2009) in Télécom ParisTech / CNRS LTCI - Signal and Image Processing (TSI) Department, and for two years (2009 - 2011) with METISS team of IRISA / INRIA - Rennes. Now he is with Technicolor Research & Innovation at Rennes, France. His research interests include audio source separation, source coding, and automatic speech recognition.

Antoine Liutkus was born in France on February 23rd, 1981. He received the State Engineering degree from Telecom ParisTech, France, in 2005, along with the M.Sc. degree in acoustics, computer science and signal processing applied to music from the Université Pierre et Marie Curie (Paris VI). He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and obtained his Ph.D. degree in the Department of Signal and Image Processing, Télécom ParisTech in 2012. His research interests include statistical signal processing, source separation, inverse problems and machine learning methods applied to signal processing.

Roland Badeau (M'02-SM'10) was born in Marseille, France, in 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 degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010.

In 2001, he joined the Department of Signal and Image Processing of Télécom ParisTech, CNRS LTCI, 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 focus on statistical modeling of non-stationary signals (including adaptive high resolution spectral analysis and Bayesian extensions to NMF), with applications to audio and music (source separation, multipitch estimation, automatic music transcription, audio coding, audio inpainting). He is a co-author of 21 journal papers, over 50 international conference papers, and 2 patents. He teaches in the Master of Engineering of Télécom ParisTech and in the Master of Sciences and Technologies of UPMC. He is also a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state) and an Associate Editor of the EURASIP Journal on Audio, Speech, and Music Processing.

Gaël Richard (SM'06) received the State Engineering degree from Télécom 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, Télécom 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 120 papers and inventor in a number of patents and is also one of the experts of the European commission in the field of audio signal processing and man/machine interfaces. He was an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing between 1997 and 2011 and one of the guest editors of the special issue on "Music Signal Processing" of IEEE Journal on Selected Topics in Signal Processing (2011). He currently is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, member of the EURASIP and AES and senior member of the IEEE.

Paper 4 (Bilen, Ozerov & Pérez, *IEEE TSP*, 2018)

Solving Time Domain Audio Inverse Problems using Nonnegative Tensor Factorization

Çağdaş Bilen, Alexey Ozerov, and Patrick Pérez

***Abstract*—Nonnegative matrix and tensor factorizations (NMF and NTF) are important tools for modeling nonnegative data, which gained increasing popularity in various fields, a significant one of which is audio processing. However there are still many problems in audio processing, for which the NMF (or NTF) model has not been successfully utilized. In this work we propose a new algorithm based on NMF (and NTF) in the short-time Fourier domain for solving a large class of audio inverse problems with missing or corrupted time domain samples. The proposed approach overcomes the difficulty of employing a model in the frequency domain to recover time domain samples with the help of probabilistic modeling. Its performance is demonstrated for the following applications: Audio declipping and declicking (never solved with NMF/NTF modeling prior to this work); Joint audio declipping/declicking and source separation (never solved with NMF/NTF modeling or any other method prior to this work); Compressive sampling recovery and compressive sampling-based informed source separation (an extremely low complexity encoding scheme that is possible with the proposed approach and has never been proposed prior to this work).**

I. Introduction

Nonnegative matrix factorization (NMF) [1] and nonnegative tensor factorization (NTF) [2] decompositions have recently found great success in applications to audio modeling, notably for source separation [3]–[5], compression [6], [7], music transcription [8], [9] and audio inpainting [10]–[12]. It is now well-established in the audio signal processing community that spectrograms of natural audio signals exhibit a low-rank NMF (or NTF in case of multi-source signals) structure. They are indeed composed of relatively few characteristic spectral patterns modulated in time (*e.g.*, harmonic combs) that are well approximated by rank-1 nonnegative matrices/tensors. Within all these applications the power-spectrograms of single-channel or multichannel audio signals (usually powers of their short-time Fourier transforms (STFT)) are decomposed using NMF or NTF models.

However, these methods address quite poorly the situations when some chunks or samples of audio signals are missing in time domain, as for example in the situations of audio declipping or declicking, as described in a general audio inpainting paper [13]. Indeed, the NMF/NTF-based audio inpainting methods [10]–[12] assume that the audio data is missing directly in the corresponding time-frequency domain, usually the STFT domain. This is in fact the most convenient situation since the modeling itself is formulated in the STFT domain, and thus it becomes quite easy to take properly into account the missing values. In the case of audio with missing samples in the time domain, one can convert the missing information into an STFT domain formulation by simply assuming that all the STFT frames corresponding to missing time samples are missing in entirety. However, this will often lead to the loss of a huge amount of available information. In the case of a clipped audio for example, every STFT frame may be clipped, thus this naive solution would lead to considering the whole signal to be missing, even though there is perhaps only 20 % of the signal that is clipped in the time domain. Another problem of NMF/NTF-based audio inpainting methods [10]–[12] which consider fully-missing STFT coefficients is that NMF/NTF models are phase-invariant and thus they only allow estimating the magnitudes of the missing coefficients. As a result, the phase information, which is very important for audio perceptual quality, still needs to be reconstructed somehow. A popular approach by Griffin and Lim [14] is usually used for the phase reconstruction, but it performs quite poorly in many situations. As an alternative, a so-called high resolution NMF (HR-NMF) approach was proposed [15], [16]. This approach extends the NMF to model temporal dependencies between time-frequency bins, which yields better phase estimates. However, for the moment this approach is quite computationally expensive and it is limited to harmonic sounds. At the same time, when some samples are missing in the time domain and one manages to estimate properly the phase-invariant NMF model and the missing samples from these observations, the resulting phase estimates should be better than those obtained via Griffin and Lim's approach [14], since missing samples in time domain does not mean completely discarding the phase information in the STFT domain.

In this work, we propose a new approach allowing the estimation of lost time domain audio samples of audio sources and/or their mixture via applying a low-rank NMF/NTF model to latent power-spectrograms of the signals in the time frequency domain. The proposed method uses Itakura Saito (IS) divergence [4] for measuring how well the given NMF/NTF model parameters estimate the signal variances while using all the information available from all of the known time domain samples from the sources and/or the mixture. The model parameters are estimated using a generalized expectation-maximization (GEM) algorithm [17] and Wiener filtering [18] is used to recover the unknown signals. Unlike some other approaches that directly apply NMF/NTF model on the STFT coefficient magnitudes or powers, the proposed

Ç. Bilen is with Audio Analytic Ltd., UK, e-mail: contact@cagdasbilen.com. A. Ozerov is with Technicolor, France, e-mail: alexey.ozerov@technicolor.com. P. Pérez is with Valeo.ai, France, e-mail: patrick.perez@valeo.com

This work was partially supported by ANR JCJC program MAD (ANR-14-CE27-0002).

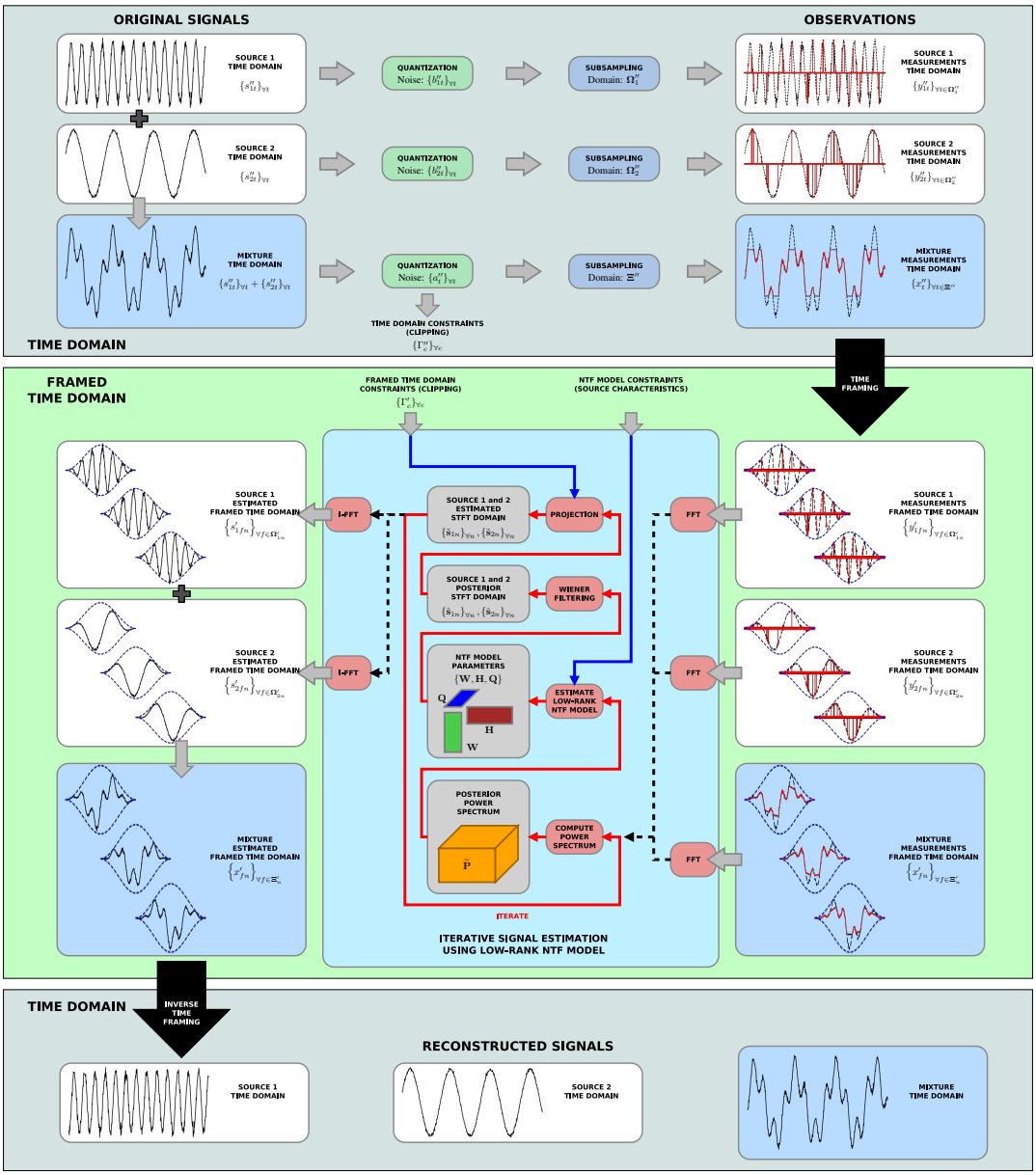


Fig. 1: The general framework of the proposed algorithm illustrating recovery of a mixture signal and its sources from a subset of the quantized samples of the sources and/or the mixture. The top section displays the time domain signals and an illustration of the generalized time domain audio inverse problem (recovering the sources from the measurements). The middle section illustrates the framed time domain variables, and the setup of the framed-time domain audio inverse problem (recovering the framed sources from the framed measurements). The middle section also illustrates a summary of the proposed algorithm steps. The bottom section illustrates the final output of the algorithm in time domain.

approach is formulated as a probabilistic Gaussian model on the complex-valued STFT coefficients. This enables us to estimate the NMF/NTF model in a *maximum likelihood (ML)* sense directly from the time domain observations, thus avoiding sub-optimally converting this missing information into the STFT domain. Furthermore, thanks to the flexibility of the NMF/NTF representations, the proposed framework can take into account mixtures of several sources, where both the sources and the mixtures can be partially or fully-missing in time domain. Last but not least, when the observed signals are not only partially lost but also corrupted, such as by noise or quantization, these corruptions can also be taken into account in the proposed approach. Within this general formulation the proposed framework is not limited to audio inpainting, but also becomes useful for different new applications related to audio compression, enhancement and source separation. This work builds on several previous conference/workshop publications [19]–[24] by the authors. Particular instances of the proposed approach for some specific applications have been presented in [19]–[21] and summarized in [22]. In this paper, we provide a

generalized formulation that is highly flexible to be adapted to various applications. We also present more comprehensive and extended experimental results, notably new experiments on compressive sampling recovery. The audio source separation approach presented in [23] falls within this general formulation as well, but it is not considered here for conciseness. Finally, while we here formulate the framework in the case of single-channel audio mixtures, its extension to the multichannel case is straightforward, which has been demonstrated in [24] for the declipping application.

More specifically, our general framework allowing recovering audio sources from partially observed and possibly quantized time domain audio samples of the sources and/or their mixture is applied here to the following existing or new applications:

- *Time domain audio inpainting and audio declipping* [13], [19], [25]–[28], where the mixture consisting of just one source is partially observed due to, *e.g.*, clipping. This is an existing application and we propose a new method to solve it.
- *Joint audio inpainting and source separation* [20], where the mixture consisting of several latent sources is partially observed due to, *e.g.*, clipping. The problem itself exists, but to the best of our knowledge, it was never addressed in a direct and systematic manner.
- *Compressive sampling recovery* [29], where the mixture consisting of just one source is partially observed due to a random sub-sampling. This is an existing application and we propose a new method to solve it.
- *Compressive sampling-based informed source separation* [21], where the mixture is observed and it consists of several latent sources that are partially observed after a random sub-sampling and quantization. This is a new informed source separation [7], [30] scheme resulting in an extremely fast encoder and a slow decoder.

The rest of this paper is organized as follows. The problem is formally defined in Section II and the proposed algorithm to solve it is described in Section III. Experiment results for various applications are given in Section IV and lastly final remarks and conclusions are presented in Section V. Readers willing to understand better and in detail the applications, before diving into the theoretical framework in Sections II and III, are invited to go through Section IV first.

II. Problem Definition

Let us consider a single-channel[1] mixture that is composed of J sources, among which each of the sources and/or the mixture might be fully, or partially observed and/or corrupted with noise (*e.g.*, quantization noise). For a mixture of length T, the mixture samples, x''_t,[2] are measured at a subset $\mathbf{\Xi}'' \subset [\![1, T]\!]$ of the entire time domain. Hence, the measured mixture samples can be represented in terms of unknown source samples, $s''_{jt}, j \in [\![1, J]\!]$, as

$$x''_t = \sum_{j=1}^{J} s''_{jt} + a''_t, \quad \forall t \in \mathbf{\Xi}'', \tag{1}$$

where a''_t represents the noise on the measurement sample due to various effects such as quantization. Furthermore, the individual sources may also be sampled at known subsets of the support $\mathbf{\Omega}''_j \subset [\![1, T]\!], j \in [\![1, J]\!]$ to obtain measured source samples, y''_{jt}, such that

$$y''_{jt} = s''_{jt} + b''_{jt}, \quad \forall t \in \mathbf{\Omega}''_j, \forall j \in [\![1, J]\!] \tag{2}$$

where b''_{jt} represents the noise for samples of each source. Lastly, for some problems such as declipping, we may also be given a set of C constraints, $\Gamma''_c(s''), c \in [\![1, C]\!]$, where each constraint, $\Gamma''_c(s'')$, is in one of the following forms:

$$s''_{j_c t_c} \geq \gamma''_c,\ s''_{j_c t_c} \leq \gamma''_c,\ \sum_{j=1}^{J} s''_{j t_c} \geq \gamma''_c,\ \sum_{j=1}^{J} s''_{j t_c} \leq \gamma''_c \tag{3}$$

in all of which γ''_c is a known constant and t_c and j_c are known time and source indices respectively.

Generalized Time Domain Audio Inverse Problem: *Given all of the above definitions, we define the generalized audio inverse problem in time domain as that of recovering the sources, $\{s''_{jt}\}_{\forall t,j}$ (and hence their mixture), given the noisy and incomplete measurements, $\{y''_{jt}\}_{\forall t \in \mathbf{\Omega}''_j, \forall j}$ and $\{x''_t\}_{\forall t \in \mathbf{\Xi}''}$, such that the constraints, $\{\Gamma''_c(s'')\}_{\forall c}$, are satisfied.*

III. Proposed Approach

A simple illustration of the known and unknown signals in the generalized time domain audio inverse problem is shown in the top section of the Figure 1, whereas the proposed algorithm in this work to solve this problem is illustrated in the middle section in the same figure. The individual steps of the proposed approach are explained in detail through the following subsections.

A. Redefining the Problem for a Frequency Domain Solution

The problem defined in Section II deals with constraints and unknowns in time domain, and as a result solving it with an approach that utilizes STFT domain constraints (such as the NTF model that will be introduced in Section III-B) can be computationally heavy and even intractable. To rectify this issue, we will introduce the framed-time domain and STFT domain notations, using which, we will define a modified problem that is much easier to handle.

The framed-time domain (or sometimes called windowed-time domain) is the representation of the time domain signal after it is split into (often overlapping) frames of fixed length, F, and multiplied by a fixed windowing function. Assuming that the total number of frames is N, the notations $x'_{fn}, y'_{jfn}, s'_{jfn}, a'_{fn}, b'_{jfn}, \mathbf{\Xi}'_n \subset [\![1, F]\!], \mathbf{\Omega}'_{jn} \subset [\![1, F]\!]$ represent the framed-time domain counterparts of the time domain notations defined in Section II for the source $j \in [\![1, J]\!]$, the intra-frame index $f \in [\![1, F]\!]$ within the frame $n \in [\![1, N]\!]$.

[1]For sake of simplicity, we only consider the single-channel case here. The proposed algorithm in this paper can be readily extended to multichannel case in a similar way as it is done in [24] for the declipping application.

[2]Throughout this paper the time domain signals will be denoted by letters with two primes, *e.g.*, x'', the framed-time domain signals by letters with one prime, *e.g.*, x', and complex-valued STFT coefficients by letters with no prime, *e.g.*, x.

The relationships between the framed-time domain variables are similar to that of the time domain counterparts such that[3]

$$x'_{fn} = \sum_{j=1}^{J} s'_{jfn} + a'_{fn}, \quad \forall f \in \Xi'_n, \forall n \tag{4}$$

$$y'_{jfn} = s'_{jfn} + b'_{jfn}, \quad \forall f \in \Omega'_{jn}, \ \forall j, n \tag{5}$$

We represent the STFT coefficients of the source signals simply by $\{s_{jfn}\}_{\forall j,f,n}$. Note that, the STFT coefficients are simply the Fourier transforms of the framed-time domain signals, such that $\mathbf{s}_{jn} = [s_{jfn}]_{\forall f} = \mathbf{U}\mathbf{s}'_{jn} \ \forall j, n$, where $\mathbf{s}'_{jn} \triangleq [s'_{jfn}]_{\forall f}$ and $\mathbf{U}$ is the normalized Fourier transform matrix satisfying $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}$.[4]

We now define a modified version of the initial problem using the framed-time domain variables and constraints, all of which can easily be computed from the time domain counterparts. This new definition of the problem has more relaxed conditions from the original problem in the sense that the problem is moved to a larger over-complete domain, and the correlation between the information within different frames is no longer defined. In the rest of this paper, we shall focus on solving this relaxed problem rather than the initial one.

Framed-Time Domain Audio Inverse Problem: *We define our problem as that of recovering the sources in framed-time domain, $\{s'_{jfn}\}_{\forall j,f,n}$ (or equivalently in STFT domain $\{s_{jfn}\}_{\forall j,f,n}$ since they are related with a unitary transform) given the noisy and incomplete framed-time measurements, $\{y'_{jfn}\}_{\forall f \in \Omega'_{jn}, \forall j,n}$ and $\{x'_{fn}\}_{\forall f \in \Xi'_n, \forall n}$, such that the constraints, $\{\Gamma'_c(s')\}_{\forall c}$, are satisfied.*

B. Applying NTF Model estimated via a GEM Algorithm

In order to make the problem described in Section III-A easier to solve, we make a number of assumptions:

Assumption 1. ***The noise is independently Gaussian distributed with known variance:*** *The noise time samples for the observations, $\{a'_{jfn}\}_{\forall j,f,n}$ and $\{b'_{fn}\}_{\forall f,n}$, are independently distributed with zero mean Gaussian with known variances, $\{\sigma^2_{a,jfn}\}_{\forall j,f,n}$ and $\{\sigma^2_{b,fn}\}_{\forall f,n}$ respectively, i.e.*

$$a'_{jfn} \sim \mathcal{N}_c(0, \sigma^2_{a,jfn}), \quad b'_{fn} \sim \mathcal{N}_c(0, \sigma^2_{b,fn}), \quad \forall j, f, n. \tag{6}$$

Assumption 2. ***The sources are independently Gaussian distributed:*** *Similarly, the unknown STFT coefficients of the sources, $\{s_{jfn}\}_{\forall j,f,n}$, are also independently distributed with zero mean complex valued Gaussian with variance $\{v_{jfn}\}_{\forall j,f,n}$, i.e.*

$$s_{jfn} \sim \mathcal{N}_c(0, v_{jfn}), \quad \forall j, f, n. \tag{7}$$

Even though it is known that the noise in practice (such as quantization noise) is not always Gaussian, modeling the noise as Gaussian is still known to be a good enough approximation that provides significant computational advantage. Similarly the assumption of Gaussian distribution for the sources is also very common in audio community and accepted as a good approximation. It is noted when dealing with non-stationary signals that the assumption of gaussianity in the sources often results in very little loss in the source separation performance with the added benefit of much lower computational requirements [31]. Without further assumptions the variances v_{jfn} in (7) would be difficult to estimate, since there are as many parameters (variances) as the observations. Hence in this work we will also assume that the variances v_{jfn} are structured via a low-rank nonnegative tensor.

Assumption 3. ***Variances of the sources form a low rank NTF structure:*** *The tensor of source variances, $[v_{jfn}]_{j,f,n}$, is represented as the sum of few rank-1 nonnegative tensors, i.e.*

$$v_{jfn} = \sum_{k=1}^{K} q_{jk} w_{fk} h_{nk}, \quad \forall j, f, n \tag{8}$$

with number of components, K, sufficiently small. This so-called PARAFAC/CANDECOMP [32] NTF model can be parametrized by $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$, such that $\mathbf{Q} = [q_{jk}]_{j,k} \in \mathbb{R}_+^{J\times K}$, $\mathbf{W} = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} = [h_{nk}]_{n,k} \in \mathbb{R}_+^{N\times K}$.

The assumption of a low rank NTF structure on the joint variances of audio sources is well known in the audio source separation community and it is shown to be an accurate model for audio signals in practice [7], [30], [33]. Please note that when the signal is treated as a single source (*i.e.* without source separation and $J = 1$), the tensor of source variances reduces to a matrix and the decomposition is simply a low-rank NMF representation.

We can define the observed mixture vector at frame n, $\mathbf{x}'_n$, and the observed source vector at frame n for source j, $\mathbf{y}'_{jn}$, as

$$\mathbf{x}'_n \triangleq \left[x'_{fn}\right]_{\forall f \in \Xi'_n} \in \mathbb{R}^{|\Xi'_n|\times 1}, \tag{9}$$

$$\mathbf{y}'_{jn} \triangleq \left[y'_{jfn}\right]_{\forall f \in \Omega'_{jn}} \in \mathbb{R}^{|\Omega'_{jn}|\times 1}. \tag{10}$$

Hence for each frame we can define the observed data vector, $\mathbf{o}'_n$, and each unknown source vector, $\mathbf{s}'_{jn}$, as

$$\mathbf{o}'_n \triangleq \left[\mathbf{y}'^{\,T}_{1n}, \cdots, \mathbf{y}'^{\,T}_{Jn}, \mathbf{x}'^{\,T}_n\right]^T. \tag{11}$$

Given the three assumptions above, we propose estimating the NTF model $\boldsymbol{\theta}$ in the ML sense as

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}'} p(\{\mathbf{o}'_n\}_{\forall n} | \boldsymbol{\theta}'). \tag{12}$$

To achieve that we employ a GEM algorithm [17], while considering as latent data the totality of in general missing source STFT coefficients $\mathbf{S} = [s_{jfn}]_{\forall j,f,n}$. The algorithm iteratively alternates between an expectation step (E-step) for estimating the posterior power spectra of the signal and a maximization step (M-step) for updating the NTF model parameters. These two main steps can be summarized as follows:

- **E-step**: Estimate conditional expectations of source power spectra $|s_{jfn}|^2$, given the current model $\boldsymbol{\theta}$ and the

[3] From this point on, we shall use simply $\forall n$ to denote $\forall n \in [\![1, N]\!]$, $\forall f$ to denote $\forall f \in [\![1, F]\!]$ and $\forall j$ to denote $\forall j \in [\![1, J]\!]$, unless a subset of these sets is specified, *e.g.* Ξ'_n.

[4] $\mathbf{x}^T$ and $\mathbf{x}^H$ represent the non-conjugate transpose and the conjugate transpose of the vector (or matrix) $\mathbf{x}$ respectively.

observations:

$$\hat{p}_{jfn} = \mathbb{E}\left[\,|s_{jfn}|^2 \big|\, \mathbf{o}'_n; \boldsymbol{\theta}\right], \quad \forall j, f, n. \tag{13}$$

- **M-step**: Re-estimate NTF model parameters such that the 3-valence tensor of the NTF model approximation, $\mathbf{V} = [v_{jfn}]_{\forall j,f,n}$, is as close to the 3-valence tensor of estimated source power spectra, $\widehat{\mathbf{P}} = [\hat{p}_{jfn}]_{\forall j,f,n}$, as possible with respect to the IS divergence [4]

$$D_{IS}(\widehat{\mathbf{P}}\|\mathbf{V}) = \sum_{\forall j,f,n} d_{IS}(\hat{p}_{jfn}\|v_{jfn}), \tag{14}$$

where $d_{IS}(x\|y) = x/y - \log(x/y) - 1$, and $\hat{p}_{jfn}$ and v_{jfn} are as specified respectively by (13) and (8).

The details of the E-step and the M-step are given in Sections III-C and III-D. In certain problems additional steps might also be required to satisfy certain constraints for the time domain signal or the NTF model parameters. It is described in Section III-E how these additional constraints can be handled by the proposed algorithm. A summary of the overall algorithm is given in Algorithm 1.

C. E-Step: Estimating Posterior Statistics

Following our assumptions of independently Gaussian distributed signals, we can write the posterior distribution of each source frame $\mathbf{s}_{jn}$ given the corresponding observed data $\mathbf{o}'_n$ and the NTF model $\boldsymbol{\theta}$ (or equivalently $\mathbf{V} = [v_{jfn}]_{\forall j,f,n}$ with v_{jfn} defined in (8)) as $\mathbf{s}_{jn}|\mathbf{o}'_n; \boldsymbol{\theta} \sim \mathcal{N}_c(\hat{\mathbf{s}}_{jn}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}})$ with $\hat{\mathbf{s}}_{jn}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$ being, respectively, posterior mean and posterior covariance matrix of the STFT coefficients, $\mathbf{s}_{jn}$. These terms can be computed respectively by Wiener filtering as [18]

$$\hat{\mathbf{s}}_{jn} = \boldsymbol{\Sigma}^H_{\mathbf{o}'_n\mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\mathbf{o}'_n\mathbf{o}'_n} \mathbf{o}'_n, \tag{15}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} - \boldsymbol{\Sigma}^H_{\mathbf{o}'_n\mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\mathbf{o}'_n\mathbf{o}'_n} \boldsymbol{\Sigma}_{\mathbf{o}'_n\mathbf{s}_{jn}}, \tag{16}$$

given the definitions of the covariance matrices

$$\boldsymbol{\Sigma}_{\mathbf{o}'_n\mathbf{o}'_n} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}'_{1n}\mathbf{y}'_{1n}} & \cdots & \underline{\mathbf{0}} & \boldsymbol{\Sigma}^H_{\mathbf{x}'_n\mathbf{y}'_{1n}} \\ \vdots & \ddots & \vdots & \vdots \\ \underline{\mathbf{0}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{y}'_{Jn}\mathbf{y}'_{Jn}} & \boldsymbol{\Sigma}^H_{\mathbf{x}'_n\mathbf{y}'_{Jn}} \\ \boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{y}'_{1n}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{y}'_{Jn}} & \boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{x}'_n} \end{bmatrix}, \tag{17}$$

$$\boldsymbol{\Sigma}_{\mathbf{o}'_n\mathbf{s}_{jn}} = \left[\underline{\mathbf{0}}^T_{L_{1,jn}\times F}, \boldsymbol{\Sigma}^T_{\mathbf{y}'_{jn}\mathbf{s}_{jn}}, \underline{\mathbf{0}}^T_{L_{2,jn}\times F}, \boldsymbol{\Sigma}^T_{\mathbf{x}'_n\mathbf{s}_{jn}}\right]^T, \tag{18}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}'_{jn}\mathbf{y}'_{jn}} = \mathbf{U}(\boldsymbol{\Omega}'_{jn})^H \mathrm{diag}\left([v_{jfn}]_{\forall f}\right) \mathbf{U}(\boldsymbol{\Omega}'_{jn}) + \mathrm{diag}\left([\sigma^2_{b,jfn}]_{\forall f\in\boldsymbol{\Omega}'_{jn}}\right), \tag{19}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{x}'_n} = \mathbf{U}(\boldsymbol{\Xi}'_n)^H \mathrm{diag}\left(\left[\sum\nolimits_{\forall j} v_{jfn}\right]_{\forall f}\right) \mathbf{U}(\boldsymbol{\Xi}'_n) + \mathrm{diag}\left([\sigma^2_{a,fn}]_{\forall f\in\boldsymbol{\Xi}'_n}\right), \tag{20}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{y}'_{jn}} = \mathbf{U}(\boldsymbol{\Xi}'_n)^H \mathrm{diag}\left([v_{jfn}]_{\forall f}\right) \mathbf{U}(\boldsymbol{\Omega}'_{jn}), \tag{21}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}'_{jn}\mathbf{s}_{jn}} = \mathbf{U}(\boldsymbol{\Omega}'_{jn})^H \mathrm{diag}\left([v_{jfn}]_{\forall f}\right), \tag{22}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}'_n\mathbf{s}_{jn}} = \mathbf{U}(\boldsymbol{\Xi}'_n)^H \mathrm{diag}\left([v_{jfn}]_{\forall f}\right), \tag{23}$$

$$\boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \mathrm{diag}\left([v_{jfn}]_{\forall f}\right), \tag{24}$$

where $\mathbf{U}(\boldsymbol{\Omega}'_{jn})$ is the $F \times |\boldsymbol{\Omega}'_{jn}|$ matrix of columns from $\mathbf{U}$ with index in $\boldsymbol{\Omega}'_{jn}$ and $L_{1,jn} \triangleq \sum_{l=1}^{j-1} |\boldsymbol{\Omega}'_{ln}|$, $L_{2,jn} \triangleq \sum_{l=j+1}^{J} |\boldsymbol{\Omega}'_{ln}|$. The term $\mathrm{diag}(\mathbf{x})$ represents a diagonal matrix with the vector $\mathbf{x}$ along the diagonal.

Finally, the posterior power spectra, $\widehat{\mathbf{P}} = [\hat{p}_{jfn}]_{\forall j,f,n}$ can be computed as

$$\hat{p}_{jfn} = \mathbb{E}\left[\,|s_{jfn}|^2 \big|\, \mathbf{o}'_n; \boldsymbol{\theta}\right] = |\hat{s}_{jfn}|^2 + \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}(f,f). \tag{25}$$

D. M-step: Updating NTF Model Parameters

Estimating NTF model $\boldsymbol{\theta}$ in the ML sense is proven [4] equivalent to minimizing the IS divergence $D_{IS}(\widehat{\mathbf{P}}\|\mathbf{V})$ as defined in (14) between the tensor of variances, $\mathbf{V}$, and the given posterior power spectra tensor, $\widehat{\mathbf{P}}$.

A common optimization approach to estimate the model parameters, $\boldsymbol{\theta}$, that minimizes (14) is using multiplicative updates (MU) as described in [4]. In our case, starting from some initial nonnegative model parameters, the model parameters that minimize (14) can be found by applying several iterations of the following updates

$$q_{jk} \leftarrow q_{jk} \left(\frac{\sum_{f,n} w_{fk} h_{nk} \hat{p}_{jfn} v^{-2}_{jfn}}{\sum_{f,n} w_{fk} h_{nk} v^{-1}_{jfn}} \right), \tag{26}$$

$$w_{fk} \leftarrow w_{fk} \left(\frac{\sum_{n,j} q_{jk} h_{nk} \hat{p}_{jfn} v^{-2}_{jfn}}{\sum_{n,j} h_{nk} q_{jk} v^{-1}_{jfn}} \right), \tag{27}$$

$$h_{nk} \leftarrow h_{nk} \left(\frac{\sum_{f,j} q_{jk} w_{fk} \hat{p}_{jfn} v^{-2}_{jfn}}{\sum_{f,j} w_{fk} q_{jk} v^{-1}_{jfn}} \right). \tag{28}$$

In the beginning of the proposed GEM algorithm, the model parameters can be initialized randomly with nonnegative values. In the following iterations however, the update of the model parameters can be always applied starting from the current model parameters (instead of randomly initializing them before MU iterations each time $\widehat{\mathbf{P}}$ is updated).

E. Applying Additional Constraints

In many practical audio inverse problems, there may be additional knowledge on the signal to be estimated apart from the observed samples. We shall consider mainly two complementary types of knowledge on the signal to be treated, which provide:

i *Constraints on NTF model parameters*, such as some characteristic spectral patterns being active, some frequency or time bins being silent, or simply the frequency response being symmetric (time domain signal being real valued);

ii *Constraints on framed-time domain samples*, such as the constraints, $\{\Gamma'_c(s')\}_{\forall c}$, that were defined earlier.

The additional constraints on the model parameters, $\boldsymbol{\theta}$, are often easy to incorporate during the MU iterations or simply initializing them in a specific way. For instance, the symmetry in frequency (hence being real valued in time) can be enforced if the matrix $\mathbf{W}$ is updated to be always symmetrical along the frequency axis. Similarly if some of the characteristic spectral patterns are known *a priori* to be present in the sources, $\mathbf{W}$ can be initialized with a specific dictionary and then may never be updated to enforce using only these patterns. Another

example is, when certain entries of the matrices $\mathbf{W}, \mathbf{H}$ and $\mathbf{Q}$ are known to be zero, they can be simply initialized to be zero and these zero values will be automatically enforced in the following MU iterations. Lastly, in certain applications, it is even possible to change the model to enforce additional structures on the matrices $\mathbf{W}, \mathbf{H}$ and $\mathbf{Q}$, such as sparsity by small modifications on the MU equations [34].

Dealing with constraints on framed-time domain samples, unlike the constrains on the model parameters, is not straightforward. When a framed-time domain sample is known to be clipped or quantized, the original value of this sample is known to be above (below) a certain threshold or to lay within a certain interval, and the resulting posterior probability distribution of the sample is no longer Gaussian. As a result, estimating the posterior power spectrum with this modified probability distribution is not as simple as described in Section III-C. To overcome this problem, we propose to estimate the posterior power spectrum by computing the posterior mean, $\hat{\mathbf{s}}_{jn}$ and the posterior covariance, $\widehat{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$, as described in Section III-C, but then projecting them so as to satisfy the time domain constraints to obtain modified statistics, $\tilde{\mathbf{s}}_{jn}$ and $\widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$ respectively. As a result, the modified posterior power spectrum (to be used as the input for the NTF model update in M-step) is obtained as

$$\tilde{p}_{jfn} = |\tilde{s}_{jfn}|^2 + \widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}(f,f). \tag{29}$$

We define several approaches to compute the aforementioned modified statistics to satisfy the constraints in framed-time domain samples:

1) **Unconstrained:** The simplest way to perform the estimation is to ignore completely the constraints, treating the problem as a more generic audio inpainting in time domain. Hence during the iterations, the "constrained" signal is taken simply as the estimated signal, *i.e.* $\tilde{\mathbf{s}}_{jn} = \hat{\mathbf{s}}_{jn}, \forall n, j$, as is the posterior covariance matrix, $\widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \widehat{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}, \forall n, j$.
2) **Ignored projection:** Another simple way to proceed is to ignore the constraint during the iterative estimation process and to enforce it at the end as a post-processing of the estimated signal. In this case, the signal is treated the same way as in the unconstrained case during the iterations.
3) **Signal projection:** A more advanced approach is to update the estimated signal at each iteration so that the magnitude obeys the constraints. As an example, let us suppose we have a constraint in the form $s'_{j_c f_c n_c} \geq \gamma'_c$ and it is not satisfied by the estimated posterior mean, *i.e.* $\hat{s}'_{j_c f_c n_c} < \gamma'_c$. We can simply set $\tilde{s}'_{j_c f_c n_c} = \gamma'_c$ and $\tilde{s}'_{jfn} = \hat{s}'_{jfn}$ for the rest of the support (and $\tilde{\mathbf{s}}_{jn} = \mathbf{U}\tilde{\mathbf{s}}'_{jn}$). Formally we can define,

$$\{\tilde{s}'_{jfn}\}_{\forall j,f,n} = \underset{\{z'_{jfn}\}_{\forall j,f,n}}{\operatorname{argmin}} \sum_{\forall j,f,n} |z'_{jfn} - \hat{s}'_{jfn}|^2 \quad \text{s.t.} \quad \{\Gamma'_c(z')\}_{\forall c} \tag{30}$$

Note that this approach does not update the posterior covariance matrix, *i.e.* $\widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \widehat{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}, \forall n, j$.
4) **Covariance projection:** In order to update the posterior mean and the posterior covariance matrix in a consistent manner, we can re-compute the posterior mean and the posterior covariance by (15) and (16) respectively, by treating the projected signal samples in (30) at the support $\mathbf{\Omega}'_{\mathbf{m},jn} \triangleq \{f | \tilde{s}'_{jfn} \neq \hat{s}'_{jfn}\}$ as observed values for the current iteration. If the resulting estimation of the sources violates the time domain constraints on additional indices, those samples are also projected to obey the constraints and treated as observed. This process is repeated until a posterior mean, $\tilde{\mathbf{s}}_{jn}$, and a posterior covariance, $\widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$, that are consistent with all the time domain constraints, are obtained. Note that in addition to updating the posterior covariance matrix, this approach also updates the entire posterior mean (or estimated signal) and not just the posterior mean at the indices of violated constraints.

Algorithm 1 GEM algorithm for solving Time Domain Audio Inverse Problems with NTF model

procedure RESTORE-AUDIO-WNTF
  Initialize nonnegative $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}\}$ randomly
  repeat
    E-step : Estimate $\hat{\mathbf{s}}_{jn}, \widehat{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}, \forall n, j$, given $\boldsymbol{\theta}$, $\mathbf{o}'_{jn}$ $\forall n, j$ ▷ see § III-C
    Time domain constraints : Estimate $\tilde{\mathbf{s}}_{jn}$, $\widetilde{\mathbf{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$, $\forall n, j$ and $\widetilde{\mathbf{P}}$ given $\{\Gamma'_c\}_{\forall c}$ ▷ see § III-E
    M-step : Update $\boldsymbol{\theta}$ given $\widetilde{\mathbf{P}}$ ▷ see § III-D, § III-E
  until convergence criteria met
end procedure

IV. Important Applications and Experimental Results

The proposed algorithm is adapted to solve a number of audio inverse problems, some of which are explored for the first time in this work. For each of these problems, we performed a set of experiments on various audio examples and compared the performance to that of known state of the art algorithms when applicable.

In the experiments below, all the audio signals are sampled at 16 kHz, and the STFT within the various instances of the proposed algorithm is computed using a half-overlapping sine window of 1024 samples (64 ms).

A. Time Domain Audio Inpainting and Audio Declipping

The problem of recovering audio samples that are lost or corrupted is often called audio inpainting [13]. We use the term "time domain audio inpainting" to refer to the problems with missing or corrupted time domain audio samples as opposed to the audio inpainting problems with missing samples in the STFT domain, for which NMF/NTF models are already being used prominently [10]–[12]. We still prefer to differentiate these problems from "audio interpolation" since the missing samples might sometimes arrive in large gaps instead of being distributed over time, and sometimes we might even encounter time domain samples missing in conjunction with missing STFT coefficients as can be encountered with audio editing applications. Two specific instances of the time domain audio

inpainting problem are called audio declipping and audio declicking [13], in which one recovers the time domain audio samples that are lost due to clipping and clicking effects caused by audio recording and compression processes. The declipping problem in particular provides additional challenges with respect to the general time domain audio inpainting problem, because it often includes additional constraints for the time domain signal to be estimated. In the recent years, the models based on sparse, cosparse or group-sparse representations in certain dictionaries are shown to be performing best to solve these problems [13], [25]–[28]. Clipping and interpolation from a Bayesian perspective has been also addressed earlier [35]–[38], mostly relying on autoregressive modeling. Though recent approaches [13], [25]–[28] have been shown performing better than (or on par with) them (see, *e.g.*, [13], [25]). Despite the success of modeling audio signals with low rank NMF representations, the time domain audio inpainting problem, especially with the additional constraints as in audio declipping, is not trivial to solve with an NMF/NTF model in the time-frequency domain. This is possibly the main reason why these models have not been utilized in time domain audio inpainting problems successfully. The proposed approach can overcome this limitation and provides a new perspective on time domain audio signal recovery with equivalent or better performance than the state of the art.

In our experiments with the audio declipping problem, we consider an audio signal with no known source information (as such it is modeled as a single source, $J = 1$) that is clipped to a known threshold of magnitude $\tau > 0$. Thus the signal is accurately known for a subset of the support, $\bar{\boldsymbol{\Xi}}''$, where signal magnitude is smaller than τ. For the remaining support, $\bar{\boldsymbol{\Xi}}'' = [\![1, T]\!] \setminus \bar{\boldsymbol{\Xi}}''$, the signal is unknown but obeys the time domain constraints of the form,

$$\begin{array}{ll} s''_{t_c} \geq \tau, & \text{for} \quad x''_{\mathrm{c},t_c} > 0 \\ s''_{t_c} \leq -\tau, & \text{for} \quad x''_{\mathrm{c},t_c} < 0 \end{array}, \quad \forall t_c \in \bar{\boldsymbol{\Xi}}''. \tag{31}$$

where x''_{c,t_c} is the clipped signal. We also assume that there is no observation noise, *i.e.*, $\sigma^2_{a,fn} = \sigma^2_{b,jfn} = 0$, $\forall j, f, n$, in (6).

In [28], various state of the art audio declipping algorithms are compared based on the experiments performed on music and speech examples. We have repeated these experiments using our approach with the same methodology and the datasets as reported in [28] and provided an overall comparison of our algorithm to the other approaches. The experiment procedure can be summarized as follows; 10 music and 10 speech signals, each of length of 4 seconds, are scaled to have maximum magnitude of 1 in time domain, and then artificially clipped at eight different clipping thresholds (uniformly spaced from 0.2 to 0.9). The proposed algorithm is tested with four different methods to handle the clipping constraints as described in Section III-E, namely *Unconstrained (NMF-U)*, *Ignored Projection (NMF-IP)*, *Signal Projection (NMF-SP)* and *Covariance Projection (NMF-CP)*. The music signals are declipped with 20 NMF components ($K = 20$), while 28 components are used for speech signals ($K = 28$). The proposed GEM algorithm is run for 50 iterations. The performance of the proposed algorithm is compared to five state of the art methods: iterative hard-thresholding (HT) [25], cosparsity (Cosp) [27], orthogonal matching pursuit (OMP) [13], social sparsity with empirical Wiener operator (SS-EW) and social sparsity with posterior empirical Wiener operator (SS-PEW) [28].

The performance metric that is used to compare the algorithms is the improvement of the signal to noise ratio (computed only on the clipped regions) with respect to the clipped signal, SNR_m, that is computed as [28]:

$$\mathrm{SNR}_m = 10 \log_{10} \frac{\sum_{\forall t \in \bar{\boldsymbol{\Xi}}''} |x''_{\mathrm{o},t}|^2}{\sum_{\forall t \in \bar{\boldsymbol{\Xi}}''} |x''_{\mathrm{o},t} - x''_{\mathrm{e},t}|^2}, \tag{32}$$

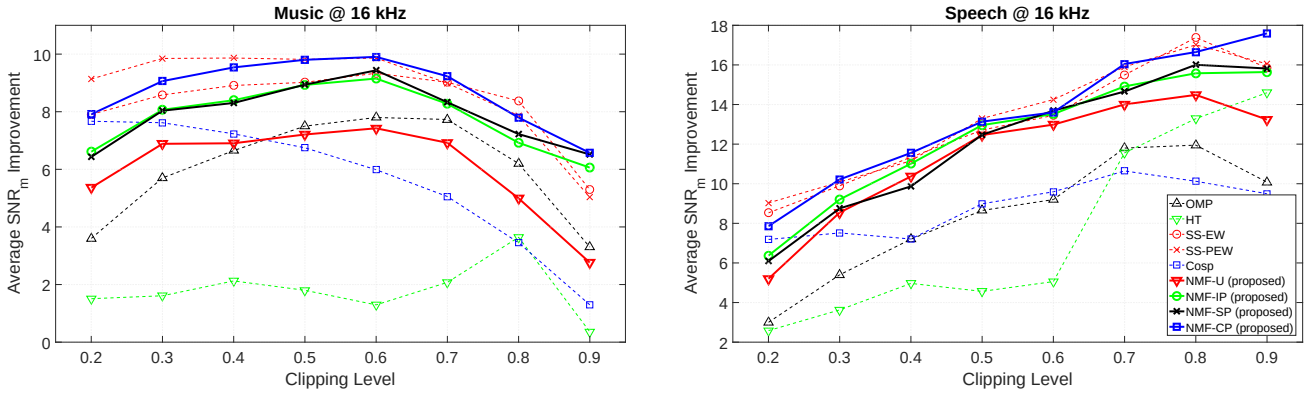
where $x''_{\mathrm{o},t}$ is the original time domain signal sample and $x''_{\mathrm{e},t}$ is the estimated signal sample. Finally, the performance is measured in terms of the SNR_m improvement, which is the difference between the SNR_m computed on the estimated signal and the SNR_m computed on the clipped signal.

The average performance of all the algorithms for declipping of music and speech signals is represented on Figure 2. It can be seen from the overall results that the proposed algorithm with the covariance projection (NMF-CP) has almost identical performance with the social sparsity based methods (SS-EW and SS-PEW) proposed in [28] while outperforming others. It can be also seen in the results that the model based algorithms (social sparsity and the NMF model) significantly outperform the methods relying on just sparsity (OMP and HT) or on just cosparsity (Cosp).

Regarding the effect of clipping constraints, the first thing to notice is that the performance of NMF-U with respect to NMF-IP (and NMF-SP) shows that simple constraints on the signal magnitude can noticeably improve the performance especially for music signals, hence they should not be ignored when possible. NMF-IP and NMF-SP are shown to have almost identical performance, even though the latter applies the constraints on the posterior mean of the signal at every iteration and the former simply applies a post processing to the final result. This observation combined with the superior performance of NMF-CP compared to the other methods demonstrates the importance of updating the posterior power spectrum more accurately for the success of the NMF-based methods.

Even though the performance is not better than the social sparsity approaches at first glance, the proposed algorithm has room for improvements in various aspects:

- NMF model can be easily extended to other, more structured NMF-like models such as source-excitation model or harmonic NMF [31]. As shown in [31] in case of source separation, having a specific model with structure that is well adapted to the considered class of signals (*e.g.*, speech, music, etc.) may improve the overall performance.
- It is shown in the results that the performance of our method depends significantly one the way the clipping constraint is handled. Therefore an alternative, more accurate computation of the posterior power spectrum might also improve the results further, whereas in dictionary based methods there is no approximation for the



(a) Average SNR_m improvement computed over 10 music signals. (b) Average SNR_m improvement computed over 10 speech signals.

Fig. 2: The average performance of all the audio declipping algorithms as a function of the clipping threshold (lower threshold corresponds to more severe clipping).

clipping constraints, hence performance improvement in this regard is not possible.

It should be noted that dealing with time domain constraints while enforcing a model on the STFT domain comes at a computational cost in the Wiener filtering stage of the proposed algorithm. Luckily, this step is independent for each frame of the signal and hence can be easily parallelized, *e.g.*, using graphical processing units (GPUs), to get a significant speed-up. On the other hand, estimating the signal independently within each window comes with the disadvantage that the estimation is not possible when there are no observed samples within a window. In practice, however, the loss of an entire window due to clipping is not probable for natural audio signals when the window size is chosen properly and the clipping threshold is not extremely low.

B. Joint Audio Inpainting and Source Separation

The audio source separation is a well known problem for which the NMF/NTF modeling in the time-frequency domain is shown to be quite successful [3]–[5]. However in all source separation problems, the audio mixture is assumed to be known perfectly whereas in practice the mixture can also have missing or corrupted (due to noise or quantization) samples in time domain. This joint problem naturally arises when one would like to perform source separation on a mixture that is degraded due to clipping effects or other degradations. This problem can also often arise in audio editing applications where some part of the audio is intentionally removed to suppress unwanted artefacts. Additionally one can also consider the case when source separation is not really needed, but a multi-source model is still employed to improve the performance of audio inpainting when dealing with mixtures of different sources.

A source separation problem with an incomplete and/or corrupted mixture is in fact a new problem that we introduce and address in this work, which, to our best knowledge, has not been properly solved by any of the existing source separation approaches in the literature, except a naive way: sequentially performing audio inpainting followed by source separation on the reconstructed mixture. The latter sequential approach can be quite suboptimal since neither of these two tasks use all of the information efficiently. The problem of jointly performing the two tasks is for the first time addressed by our proposed approach, which can recover the signal in a way that is more consistent with the multiple source nature of the corrupted mixture while simultaneously estimating the individual sources.

The global setup of our modeling to handle joint audio declipping and source separation is the same as the one for declipping in Section IV-A, except that $J > 1$ sources are considered instead of just one. In order to assess the performance of declipping and source separation using the proposed algorithm, 5 different music mixtures[5], each composed of 3 sources (bass, drums and vocals), are considered under 3 different clipping conditions. For each mixture with a maximum magnitude of 1 in time domain, 3 clipping levels at the thresholds of 0.2 (heavy clipping), 0.5 (moderate clipping) and 0.8 (light clipping) are considered, resulting in a total of 15 mixtures with different clipping levels. Each mixture is reconstructed by joint declipping and source separation, sequential declipping and source separation and only source separation ignoring the clipping artefacts. The proposed GEM algorithm (run for 100 iterations) has been used for all the reconstructions[6] with $K = 15$ components. Inline with [33] and so as to inject some information about the sources to be separated, the sources in the mixtures are artificially silenced during a percentage of the total time, and the corresponding indices in $\mathbf{H}$ are set to zero so as to inject this information into the modeling. An example of the activation periods of the sources and corresponding indices set to zero in $\mathbf{H}$ during NTF model estimation are shown in Figure 3. Similarly $\mathbf{Q}$ is simply chosen as a $J \times K$ matrix with a single 1 on each column and zeros everywhere else, to describe the assignment

[5]The mixtures are taken from the "professionally produced music recordings" task dataset of SiSEC 2015 source separation evaluation campaign (https://sisec.inria.fr/sisec-2015/).

[6]For declipping only, the algorithm is used with a single source (as in Sec. IV-A), and for source separation only, the algorithm is used with the observed support set being the entire time axis.

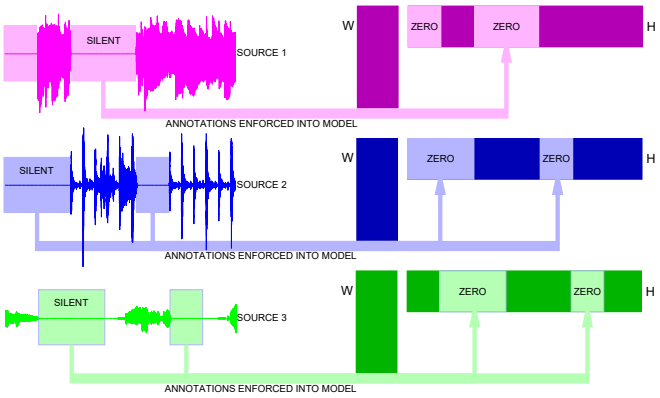


Fig. 3: Depiction of the experiment set-up for the injection of information on different source characteristics. The information on the known silent time durations of each source (represented in different colors) is directly utilized by setting the corresponding coefficients in $\mathbf{H}$ to 0. Note that the matrices $\mathbf{H}$ and $\mathbf{W}$ are formed by concatenating the Hs for each source and Ws for each source (depicted above) along the component dimension respectively.

of the components to the sources.

It should be noted that, the sequential reconstruction as well as performing only the source separation in this experiment could also have been performed with other existing methods from the literature. However, we have opted for using the same algorithm for each recovery scenario so as to clearly observe the difference due to jointly treating the two problems, rather than other differences in the reconstruction algorithms. Furthermore, as it is demonstrated that the performance of our algorithm for declipping is on par with the state of the art algorithms in Section IV-A, we find this comparison still very relevant.

The results of the simulations can be seen in Figure 4. Signal to noise ratio on the clipped support (SNR_m) computed as in (32) for the declipped mixture is shown to demonstrate the declipping performance while signal to distortion ratio (SDR) as described in [39] is shown to demonstrate the source separation performance.

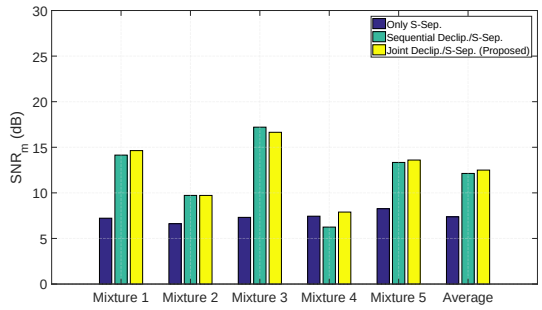
The results in Figure 4 show that when the clipping is severe, joint approach is almost always preferable since it provides improvement on both the quality of the mixture and the quality of the separated sources with respect to source separation without declipping. This is as opposed to the sequential approach which provides comparable quality improvement in the mixture at the expense of the performance in source separation. In fact, for heavy clipping the declipping in sequential approach often reduced the performance of source separation noticeably with respect to separation without declipping. As the clipping gets lighter, the performance of sequential method approaches to that of joint method, and finally performs slightly better for light clipping. The joint optimization, however, still has few drawbacks which could be improved upon. The declipping in the sequential approach is performed with $K = 15$ components without any restrictions whereas the joint optimization is performed with the additional limitation that each source uses 5 components *independently*. Hence it is not possible that two sources share a common component in the joint optimization. This can be overcome by devising better methods to inject the prior information regarding the sources, see, *e.g.*, [23]. It should be also noted that the sequential optimization is approximately twice as fast as joint optimization due to handling much less complicated problems in either steps of the sequential processing. The fact that the Wiener filtering stage is independent for each window and can be parallelized to provide significant speed improvements, can be helpful to overcome this problem in the future.

C. Compressive Sampling Recovery

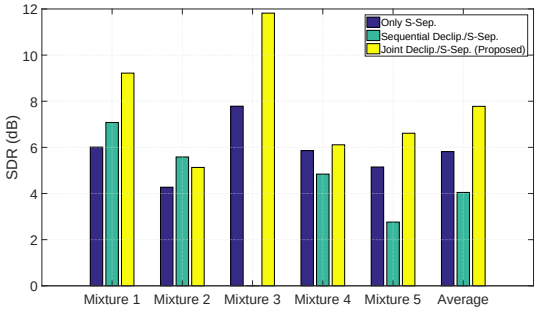
Compressive sampling [29] is the theory and application of (often) randomly subsampling a signal that is known to be compressible (*e.g.*, with sparse or low rank representations) in an *incoherent* domain and making sense of the random samples by using the prior information of compressibility. As our algorithm is well fitted for time domain audio inverse problems, the reconstruction of the randomly sampled audio signals is another field of application for which it can be useful. Even though all the model-based signal estimations rely on compressibility of signals, the differentiating factor of compressive sampling comes from the fact that the compact representation of the signal is in an incoherent (in layman terms, very different or opposite) domain to the sampling domain. As an example, frequency domain and time domain are two domains which are maximally incoherent, *i.e.*, an impulse (maximally compact) signal in one is a uniform energy (maximally distributed) signal in the other.

Looking from the compressive sampling perspective, the compressible characteristics of the audio signals exploited by our algorithm are two fold: i) the significant reduction of the probability space of the possible solutions given the known samples, through the maximum likelihood estimate ii) the further reduction of the possible solutions through the low rank modeling of the NMF/NTF representation in the STFT domain. This application can in fact be seen as another instance of audio inpainting, however we have investigated it separately as the random subsampling changes the characteristic of the problem with respect to the other more typical audio inpainting problems such as audio declipping. It must be also noted that, this application is more than mere interpolation from irregular samples, as the reconstruction model enforces dimensionality reduction in an incoherent domain, fitting well into the compressive sensing paradigm.

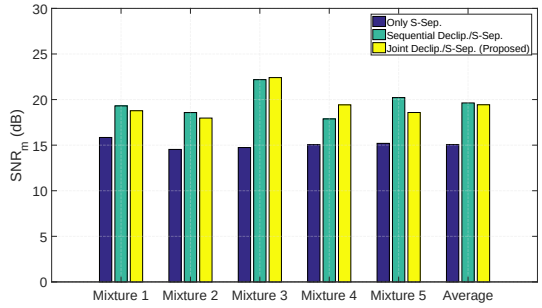
In order to demonstrate the ability of the proposed approach to reconstruct randomly subsampled signals, we have randomly subsampled a typical music signal of 4 seconds at different average rates (percentage of retained samples at $2, 4, 8, 16, 32$), and then reconstructed with our algorithm in a similar fashion to the experiments in Section IV-A, but without any clipping constraints (hence $J = 1$ and $\sigma^2_{a,jfn} = \sigma^2_{b,jfn} = 0$, $\forall j, f, n$). The reconstruction is performed with different number of components ($K = 2, 8, 24, 32, 48, 72$) in order to observe the sensitivity of the results to the parameter K. In



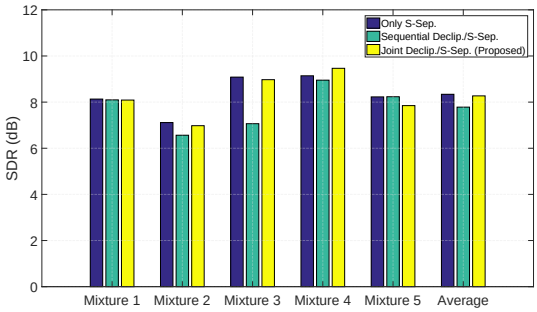
(a) Declipping performance for clipping level 0.2.



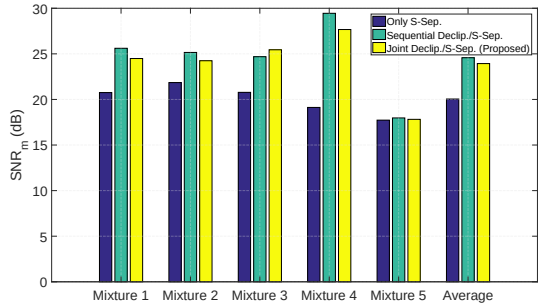
(b) Source separation performance for clipping level 0.2.



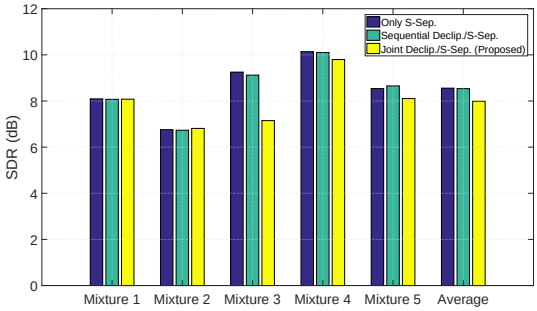
(c) Declipping performance for clipping level 0.5.



(d) Source separation performance for clipping level 0.5.



(e) Declipping performance for clipping level 0.8.



(f) Source separation performance for clipping level 0.8.

Fig. 4: The declipping and source separation performance of joint optimization compared to sequential.

order to provide a reference for the reconstruction capability of our algorithm, the results with shape preserving piecewise cubic interpolation are also provided[7].

The reconstruction results can be observed in Figure 5. The first thing to notice is that the reconstruction results with the proposed algorithm (solid lines) are significantly better than the results with simple interpolation (dashed lines) as expected. Another noticeable behaviour in the results is that once the number of components, K, is sufficiently large, the reconstruction performance does not seem to suffer. This behaviour is unlike what we have observed for other problems such as declipping, for which the choice of number of components is an important factor for obtaining best performance. Looking more closely to the estimated NMF components, we have seen that the maximum likelihood estimate combined with random sampling already provided a strong prior for signal estimation and the benefit from low rank model was minimal in this case. Hence, as long as the number of components are chosen sufficiently large, the accuracy of estimated variances, $\mathbf{V}$, are effectively independent of K.

[7]For the interpolation, the *interp1()* function of Matlab 2016a is used with *phcip* method, which gave the best results among the available interpolation methods.

D. Compressive sampling-based informed source separation

Informed source separation (ISS) [7], [30] is a variant of source separation that is in fact a source compression problem assuming that the mixture is known. The ISS problem can be defined as the problem of encoding multiple audio sources to create a bitstream (also called a *side-information*) so that the audio from the sources can be recovered given the bitstream and the mixture of the sources. The main difference of ISS from joint compression of multiple audio signals is the assumption that the mixture is available at both encoding and decoding stages. Several ISS methods were proposed [7], [30], [40] including those based on the NTF modeling [7], [30]. In all these approaches the encoding stage is usually more complex and computationally expensive than the decoding stage. The framework proposed in this work can be used

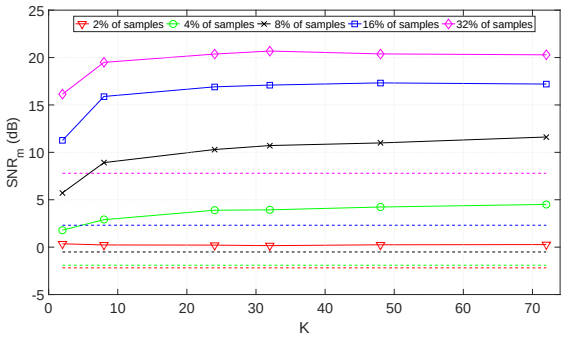
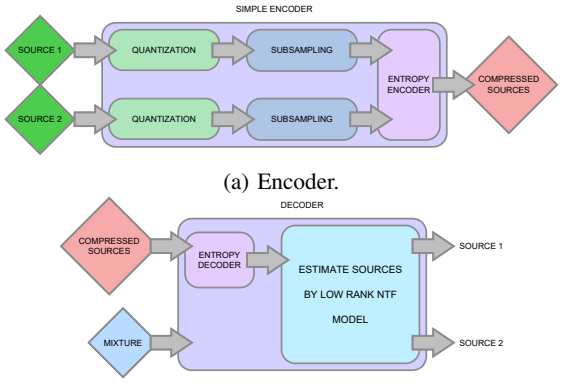


Fig. 5: The reconstruction performance measured in terms of SNR_m of a 4s long music signal from its random samples. The reconstruction results with our proposed algorithm (solid lines) are shown for different percentage of samples and different number of components, K, used in our approach. The results with shape preserving piecewise cubic interpolation are also shown for comparison (dashed lines), with the colors indicating corresponding percentage of samples.



(a) Encoder.

(b) Decoder.

Fig. 6: The encoding and decoding processes for the compressive sensing-based informed source separation.

to realize a new variant of ISS, where the computational complexity is moved from the encoder to the decoder side. To our best knowledge, this is another application that is realized for the first time with our proposed algorithm. This feat is accomplished by reducing the encoder to simply subsampling the sources in a random and independent fashion and quantizing the samples. The proposed algorithm can then be used to recover the sources at the decoder side given the encoded samples and the mixture, similar to the case of compressive sampling recovery (in fact this can be seen as more practical use of compressive sampling recovery in audio). This new approach, which we call *compressive sampling-based ISS (CS-ISS)*, is inline with both the compressive sampling [29] paradigm, since the sampling is random and in a sufficiently incoherent domain, and with the distributed source/video coding [41], [42], since the posterior source dependencies (the sources are highly correlated *a posteriori* given the mixture) and the source structure are exploited only at the decoding stage, thus allowing the complexity shift. The CS-ISS also allows independent structures between the encoder and the decoder, *i.e.*, the decoder algorithm can be modified without the need to change the encoder and the encoded bitstream. More precisely, by that we mean that given a bitstream a totally different source recovery algorithm (*e.g.*, based on social sparsity) may be developed and applied for decoding.

A summary of our CS-ISS scheme is shown in Figure 6. In order to assess the performance of our approach, three ($J = 3$) 11-second long sources of a music recording are encoded and then decoded using the proposed CS-ISS with different levels of quantization (16 bits, 11 bits, 6 bits and 1 bit) and different raw sampling bitrates[8] per source (0.64, 1.28, 2.56, 5.12 and 10.24 kbps/source). Since uniform quantization is used, the noise variance in time domain is $\sigma^2 = \Delta^2/12$ where Δ is the quantization step size. Hence $\sigma^2_{b,jfn} = \omega_f^2\Delta^2/12$, where ω_f^2 are the framing (or STFT) window coefficients. The mixture is available in entirety at the decoder, therefore the noise variance of the mixture is zero ($\sigma^2_{a,fn} = 0$). It is assumed that the random sampling pattern is *pre-defined* and known during both encoding and decoding. The quantized samples are truncated and compressed using an arithmetic encoder with a zero mean Gaussian distribution assumption. At the decoder side, following the arithmetic decoder, the sources are decoded from the quantized samples using 50 iterations of the GEM algorithm with the number of components fixed at $K = 18$, *i.e.* in average 6 components per source. The quality of the reconstructed samples is measured with SDR as described in [39]. The resulting encoded bitrates and SDR of decoded signals are presented in Table I along with the percentage of the encoded samples in parentheses. Note that the compressed rates in Table I differ from the corresponding raw bitrates due to the variable performance of the entropy coding stage, which is expected.

The performance of CS-ISS is compared to a classical ISS approach with a more complicated encoder and a simpler decoder presented in [30], as well as much better performing coding-based approach proposed in [7]. Both the classical ISS and coding-based ISS algorithms are used with NTF model quantization and encoding in a similar fashion as in the experiments described by [7], *i.e.*, NTF coefficients are uniformly quantized in logarithmic domain, quantization step sizes of different NTF matrices are computed using equations (31)-(33) from [7] and the indices are encoded using an arithmetic coder based on a two-state Gaussian mixture model (GMM) (see Fig. 5 of [7]). The approach is evaluated for different quantization step sizes and different numbers of NTF components, *i.e.*, $\Delta = 2^{-2}, 2^{-1.5}, 2^{-1}, \ldots, 2^4$ and $K = 4, 6, \ldots, 30$. The results are generated with 250 iterations of model update. The performance of CS-ISS and the earlier approaches are shown in Figure 7 in which CS-ISS clearly outperforms the classical ISS approach and is on par with coding-based ISS approach, even though both of these approaches can use an optimized number of components as opposed to our decoder which uses a fixed number of components (the encoder is very simple and does not compute or transmit this value). The performance

[8]The raw sampling bitrate is defined as the bitrate before the entropy encoding step.

Bits per Sample	Raw rate (kbps / source) 0.64	1.28	2.56	5.12	10.24
	Compressed Rate / SDR (% of Samples Kept)				
16 bits	0.50 / -1.64 dB (0.25%)	1.00 / 4.28 dB (0.50%)	2.00 / 9.54 dB (1.00%)	4.01 / 16.17 dB (2.00%)	8.00 / 21.87 dB (4.00%)
11 bits	0.43 / 1.30 dB (0.36%)	0.87 / 6.54 dB (0.73%)	1.75 / 13.30 dB (1.45%)	**3.50 / 19.47 dB** (2.91%)	**7.00 / 24.66 dB** (5.82%)
6 bits	**0.27 / 4.17 dB** (0.67%)	**0.54 / 7.62 dB** (1.33%)	**1.08 / 12.09 dB** (2.67%)	2.18 / 14.55 dB (5.33%)	4.37 / 16.55 dB (10.67%)
1 bit	0.64 / -5.06 dB (4.00%)	1.28 / -2.57 dB (8.00%)	2.56 / 1.08 dB (16.00%)	5.12 / 1.59 dB (32.00%)	10.24 / 1.56 dB (64.00%)

TABLE I: The final bitrates (in kbps per source) after the entropy coding stage of CS-ISS with corresponding SDR (in dBs) for different (uniform) quantization levels and different raw bitrates before entropy coding. The percentage of the samples kept is also provided for each case in parentheses. Results corresponding to the best rate-distortion compromise are in bold.

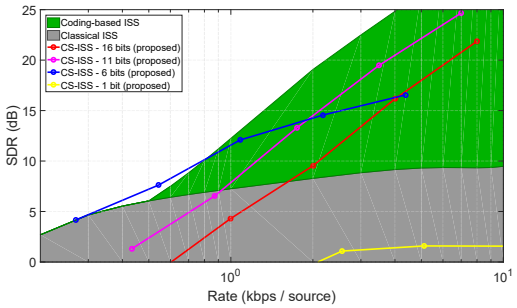


Fig. 7: The rate-distortion performance of CS-ISS using different quantization levels of the encoded samples. The performance of the ISS algorithm from [30] and the coding-based ISS algorithm from [7] are also shown for comparison.

difference with classical ISS is due to the high efficiency achieved by the CS-ISS decoder thanks to the incoherency of random sampled time domain and of maximum likelihood estimation along with low rank NTF model. Also, the classical ISS approach [30] is unable to perform beyond an SDR of 10 dBs due to the lack of additional information about STFT phase as explained in [7]. The results indicate that the rate distortion performance exhibits a similar behaviour as to the coding-based ISS algorithm. It should be reminded that the proposed approach distinguishes itself by its low complexity encoder and hence can still be advantageous against other ISS approaches with better or seemingly equivalent rate distortion performance.

The performance of CS-ISS in Table I and Figure 7 indicates that different levels of quantization may be preferable in different rates. Even though neither 16 bits nor 1 bit quantization seem well performing, the performance indicates that 16 bits quantization may be superior to other schemes when a much higher bitrate is available. Coarser quantization such as 1 bit, on the other hand, had very poor performance in the experiments. The choice of quantization can be performed in the encoder with a simple look up table as a reference. One must also note that even though the encoder in CS-ISS is very simple, the proposed decoder is significantly high complexity, typically higher than the encoders of traditional ISS methods. However, this can also be overcome by exploiting the independence of Wiener filtering among the frames in the proposed decoder with parallel processing, *e.g.*, using GPUs.

V. Conclusions

In this paper, we have presented a novel approach for time domain signal estimation in the maximum likelihood manner. It relies on the low rank NTF modeling of the power spectrum of the signal and can be applied to many types of problems that were not previously solved using the NMF/NTF model. The proposed algorithm is demonstrated to be very effective for several audio inverse problems while providing multiple advantages compared to other existing methods. For the audio declipping problem, clipped sections of music and speech signals are restored using the proposed approach as well as state of the art methods, and the proposed algorithm is shown to be highly competitive while providing complementary advantages such as naturally handling noise and quantization artefacts and easily incorporating various types of constraints. For audio source separation and mixture declipping, the proposed algorithm is shown to be capable of jointly solving these two separate problems which was not possible with any other method in the literature. Joint handling of these problems is also demonstrated to be more effective than sequentially approaching each problem in case of severe distortions. The proposed algorithm is also shown to be highly effective for the reconstruction of randomly subsampled signals such as in the case of compressive sampling approaches. This advantage of our algorithm is further utilised for the problem of informed source separation, to create a compression scheme which uses the principles of compressive sampling and distributed coding. For this application, the proposed algorithm is not only shown to achieve compression performance equivalent to that of the state of the art, but also shown to have unique advantages, specifically having a very simple encoder as well as the decoding stage being independent of the encoding stage.

The NMF and NTF representations are gaining a lot of popularity in signal modelling community and we see the algorithm presented in this paper to be a step towards the application of these models to a wider class of signal estimation problems. Even though the provided examples in this paper are all audio inverse problems, the proposed algorithm is by no means limited to audio applications. It could be used in any application for which a low rank NMF/NTF model is an accurate representation for the power spectrum.

We consider several improvements and extensions to the proposed algorithm as future work. An extension to multichannel audio is an interesting step for dealing with real world audio problems. Furthermore, adapting the proposed algorithm

for imaging problems with multiple additive components, such as imaging through transparent and reflective surfaces, is another intriguing direction.

VI. Acknowledgment

The authors would like to thank Kai Siedenburg and Matthieu Kowalski for kindly sharing numerical results from [28], and Srđan Kitić for providing the results of the algorithm from [27] on the corresponding dataset.

References

[1] D. Lee and H. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] A. Cichocki, R. Zdunek, and S. Amari, "Nonnegative matrix and tensor factorization," *IEEE Signal Processing Magazine*, pp. 142–145, 2008.

[3] 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, 2007.

[4] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[6] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *128th Audio Engineering Society Convention (AES 2010)*, London, UK, May 2010.

[7] 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*, vol. 21, no. 8, pp. 1699–1712, Aug. 2013.

[8] 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 (WASPAA)*, Oct. 2003.

[9] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[10] J. L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting problem with Bregman divergence," *Speech Communication*, vol. 53, no. 5, pp. 658–676, May-June 2011.

[11] P. Smaragdis, R. Bhiksha, and S. Madhusudana, "Missing data imputation for time-frequency representations of audio signals," *Journal of signal processing systems*, vol. 65, no. 3, pp. 361–370, 2011.

[12] U. Simsekli, A. T. Cemgil, and Y. K. Yilmaz, "Score guided audio restoration via generalised coupled tensor factorisation," in *International Conference on Acoustics Speech and Signal Processing (ICASSP'12)*, 2012, pp. 5369 – 5372.

[13] A. Adler, V. Emiya, M. 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, 2012.

[14] D. Griffin and J. S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Transactions of Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 11, pp. 1670–1680, 2014.

[16] R. Badeau and A. Ozerov, "Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain," in *Proc. 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[17] A. Dempster, N. Laird, and D. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[18] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[19] Ç. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[20] ——, "Joint audio inpainting and source separation," in *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, August 2015.

[21] ——, "Compressive sampling-based informed source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[22] ——, "Audio inpainting, source separation, audio compression. all with a unified framework based on ntf model," in *MissData 2015*, 2015.

[23] ——, "Automatic allocation of NTF components for user-guided audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 484–488.

[24] A. Ozerov, Ç. Bilen, and P. Pérez, "Multichannel audio declipping," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 659–663.

[25] S. Kitić, L. Jacques, N. Madhu, M. Hopwood, A. Spriet, and C. D. Vleeschauwer, "Consistent iterative hard thresholding for signal declipping," in *ICASSP - The 38th International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013.

[26] S. Kitić, N. Bertin, and R. Gribonval, "Audio declipping by cosparse hard thresholding," in *iTwist - 2nd international - Traveling Workshop on Interactions betweenSparse models and Technology*, Namur, Belgium, August 2014.

[27] ——, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, August, 2015.

[28] K. Siedenburg, M. Kowalski, and M. Dörfler, "Audio declipping with social sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1577–1581.

[29] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, 2008.

[30] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[31] 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, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[32] R. Bro, "Parafac. tutorial and applications," *Chemometrics and intelligent laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.

[33] 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 (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[34] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF? half-baked or well done?" Mitsubishi Electric Research Laboratories, Tech. Rep., 2015.

[35] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 317–330, 1986.

[36] S. J. Godsill and P. J. Rayner, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 267–278, 1995.

[37] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1124–1135, 1996.

[38] A. Dahimene, M. Noureddine, and A. Azrar, "A simple algorithm for the restoration of clipped speech signal," *Informatica*, vol. 32, no. 2, 2008.

[39] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[40] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[41] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, September 2004.

[42] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71 – 83, January 2005.

Curriculum Vitae

ALEXEY OZEROV

Senior Scientist in InterDigital - PhD in Signal Processing

InterDigital,
975, avenue des Champs Blancs, 35576 Cesson Sévigné, France
Tel: (office) +33 (0)2 99 27 30 34
E-mail: alexey.ozerov@interdigital.com
web: https://www.interdigital.com/talent/?id=131

Russian and French citizen,
driving licence B

EDUCATION

PhD in Signal Processing, December 2006
University of Rennes 1, France

MSc in applied mathematics: DESS "Scientific Calculation and Applications", September 2003
University of Bordeaux 1, France

MSc in mathematics, department of Ordinary Differential Equations, June 1999
Mathematics and Mechanics faculty, St. Petersburg State University, Russia

PROFESSIONAL EXPERIENCE

Senior Scientist in InterDigital, Cesson Sévigné, France June 2019 - present

Senior Scientist in Technicolor December 2014 - May 2019

Researcher Technicolor, Cesson Sevigné, France November 2011 - December 2014

Research Engineer in METISS team September 2009 - October 2011
of IRISA / INRIA, Rennes, France

Post-Doc in the Signal and Image Processing Department February 2008 - July 2009
of TELECOM ParisTech / CNRS LTCI, Paris, France

Post-Doc in Sound and Image Processing (SIP) Lab January - December 2007
of KTH (Royal Institute of Technology), Stockholm, Sweden

Preparation of PhD in Signal Processing, November 2003 - December 2006
France Télécom R&D, in collaboration with the METISS team of IRISA, Rennes, France

Period of trainee, March - September 2003
METISS team of IRISA, Rennes, France

R&D Software Engineer, Terayon Communicational Systems (USA) November 1999 - July 2002
St. Petersburg, Russia and then Prague, Czech Republic

AWARD

IEEE Signal Processing Society 2014 Best Paper Award for the paper "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation" published in IEEE TASLP journal in 2010 and co-authored with Cédric Févotte.

TUTORIALS

1. **Tutorial** at 2014 IEEE International Conference on Multimedia & Expo (ICME), Chengdu, China: "Tutorial on Nonnegative Matrix Factorisation with Applications to Audiovisual Content Analysis" (with S. Essid).
2. **Tutorial** at 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy: "Informed Audio Source Separation: Trends, Approaches and Algorithms" (with A. Liutkus and G. Richard).

RECENT TALKS

1. **Invited talk** at RWTH, Aachen, Germany on January 2018: "Probabilistic nonnegative matrix / tensor factorization for multichannel audio source separation and other audio inverse problems".
2. **Invited talk** at TELECOM ParisTech on November 2017: "Reconstructing missing samples in audio sources and mixtures / Audio style transfer".
3. **Invited talk** at a meeting of "Groupement de Recherche Information, Signal, Images et ViSion" (GdR ISIS) on February 2011: "Probabilistic nonnegative latent matrix factorization. Application to multichannel audio analysis and source separation".
4. **Invited talk** at TELECOM ParisTech on November 2010: "Flexible Approaches for Audio Coding and Source Separation".

MEMBERSHIP

1. **Senior member of IEEE and IEEE Signal Processing Society (SPS)** since 2017 (Member of IEEE and IEEE SPS partly since 2005).
2. **Distinguished member of the Technicolor Fellowship Network** 2016 - 2019 (Associated member 2014-16).
3. **Member of IEEE SPS Audio and Acoustic Signal Processing Technical Committee** (IEEE AASP TC) (2015.1 - 2017.12, 2018.1 - 2020.12).

EDITORSHIP

1. **Guest editor** of the special issue on "Reconstruction of audio from incomplete or highly degraded observations" in IEEE Journal of Selected Topics in Signal Processing (with P. Rajmic, V. Emiya, N. Holighaus and N. Bertin) (*launched, in progress*).
2. **Associate editor** of IEEE/ACM Transactions on Audio, Speech, and Language Processing (2017.2 - 2020.1).
3. **Guest editor** of the special issue on "Informed Acoustic Source Separation" in EURASIP Journal on Advances in Signal Processing (with G. Richard, T. Cemgil, T. Virtanen and D. Fitzgerald).

PUBLIC RESPONSIBILITIES

1. **Area co-chair for ICASSP 2020:** special sessions, and Audio for Multimedia and Audio Processing Systems.
2. **Member of evaluation committee of the ANR** CE23: "Artificial Intelligence" (2018-2019).
3. **Area co-chair for ICASSP 2019:** Audio for Multimedia and Audio Processing Systems.
4. **Area co-chair for ICASSP 2018:** Audio for Multimedia.
5. **Co-organizing a special session** on "New Extensions and Applications of Non-Negative Audio Modeling" at IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (with H. Kameoka, C. Févotte and P. Smaragdis) - 2017.

6. **Area chair and co-chair for ICASSP 2017:** Audio for Multimedia (chair) and Audio and Speech Source Separation (co-chair).
7. **Area co-chair for ICASSP 2016:** Audio and Speech Source Separation.
8. **Co-organizing a Technicolor Fellowship Network Workshop** on "Multimedia Inpainting" (with P. Pérez) - 2016.
9. **Co-organizing a Technicolor Fellowship Network Workshop** on "Source Separation" (with L. Chevallier) - 2013.
10. **Member of scientific seminars organization committee** in Technicolor (since Oct. 2012).
11. **Member of the organizing committee of the international signal separation evaluation campaign** SiSEC 2010 (`sisec2010.wiki.irisa.fr`).
12. **Member of the local organization committee of the international conference** "Latent Variable Analysis and Signal Separation" (LVA/ICA'10) (`lva2010.inria.fr`), Sep. 2010 in St. Malo.
13. **Coordination of preparation of a European project proposal** (STREP) for FP7 (ICT Call 1: FP7-ICT-2007-1, Challenge 4: "Digital libraries and Content", Objective 1: "Digital libraries and technology-enhanced learning") with 5 European research labs (from Sweden, France, Denmark, Finland and Portugal).
14. **Member of the local organization committee of the international conference** "Signal Processing with Adaptative Sparse Structured Representations" (SPARS'05) (`spars05.irisa.fr`), 16 to 18 November 2005 in IRISA/INRIA - Rennes, France.

SUPERVISION

PhD students:

1. **Sanjeel Parekh** (2016 - 2019) "Learning Representations for Robust Audio-Visual Scene Analysis" (co-advised with N. Duong, P. Pérez, S. Essid and G. Richard).

Master students:

1. **Charles Blandin**, Inria (2010) "Audio source separation by hierarchical clustering from partial data" (co-advised with E. Vincent).
2. **Luc Le Magoarou**, Technicolor (2013) "Text-informed audio source separation" (co-advised with N. Duong) **Technicolor Best Internship Award** (annual competition among 30 to 40 interns).
3. **Dalia El Badawy**, Technicolor (2014) "On-the-fly audio source separation" (co-advised with N. Duong) **Technicolor Best Internship Award**.
4. **Pierre Prablanc**, Technicolor (2015) "Voice conversion for speech inpainting" (co-advised with N. Duong and P. Pérez) **running for Technicolor Best Internship Award (top 3)**.
5. **Gustavo Sena Mafra**, Technicolor (2015) "Acoustic scene classification with deep neural network" (co-advised with N. Duong and P. Pérez).
6. **Swann Leduc**, Technicolor (2016) "Cross-lingual voice-conversion" (co-advised with N. Duong and P. Pérez).
7. **Pei-I Chen**, Technicolor (2017) "Video matting for high-end VFX" (co-advised with T. Crivelli) **running for Technicolor Best Internship Award (top 3)**.
8. **Eric Grinstein**, Technicolor (2017) "Audio manipulation with deep representations" (co-advised with N. Duong and P. Pérez).
9. **Le HaQuang**, Technicolor (2018) "Audio attributes modification with deep representations" (co-advised with N. Duong and G. Puy).
10. **Antoine Caillon**, Technicolor/InterDigital (2019) "Speech transformations using deep generative models" (co-advised with N. Duong and G. Puy).

11. **Valentin Bilot**, Technicolor/InterDigital (2019) "Audio event classification via multiple instance learning" (co-advised with N. Duong).

Student projects:

1. **5 students from ENSAI school** (2013) "Multimodal identification of charachters in videos" (co-advised with J.-R. Vigouroux).

PARTICIPATION TO PHD JURIES AND FOLLOWINGS

1. **Joonas Nikunen** (2015) "Object-based modeling of audio for coding and source separation", Tampere University of Technology (Pre-examiner).
2. **Clément Gaultier** (2015 - 2018) Inria, Centre Inria Rennes (mid-term following in 2017, 2018).

TEACHING ACTIVITIES

1. Seminars on "**Information theory and source coding**" (6 hours) in the Royal Institute of Technology (KTH), Stockholm, Sweden (February 2007).
2. Seminars on "**Theory of functions of a complex variable**" (8 hours) in the Institute of Intellectual Systems and Technology (IIST), Saint-Petersburg, Russia (March 1999).

COLLABORATIVE PROJECTS

1. **MIP-Frontiers** "New Frontiers in Music Information Processing" is a European Training Network funded by the European Commission (*in progress*).
2. **MAD** "Missing Audio Data Inpainting" is a French ANR project on developing new audio inpainting concepts and algorithms relying on signal processing and machine learning (*completed*).
3. **AXES** is a european FP7 project on conception and development of innovative search and indexing tools facilitating the access to audiovisual digital libraries (*completed*).
4. **Quaero** is a collaborative research and development program promoting research and industrial innovation on technologies for automatic analysis of multimedia and multilingual documents (*completed*).
5. **SARAH** "StAndardisation du Remastering Audio Haute-Définition" is a French ANR project on high-quality HD remastering of music recordings (*completed*).
6. **FlexCode** is a european FP6 project on new practical flexible, parameterized and generic coding system for speech and audio coding (*completed*).

OTHER EXTERNAL COLLABORATIONS

1. While in TELECOM ParisTech (2008-2009), collaborating with S. Arberet, F. Bimbot and R. Gribonval from IRISA / INRIA on audio source separation.
2. While in IRISA / INRIA (2009-2011), collaborating with A. Liutkus, R. Badeau and G. Richard from TELECOM ParisTech on informed audio source separation.
3. While in IRISA / INRIA (2009-2011), collaborating with M. Lagrange from IRCAM on uncertainty-based learning of acoustic models.
4. While in Technicolor (2011-2019), collaborating with S. Essid and G. Richard from TELECOM ParisTech to co-supervise PhD of Sanjeel Parekh.

PARTICIPATION IN INTERNATIONAL EVALUATION CAMPAIGNS

1. Detection and Classification of Acoustic Scenes and Events challenge (DCASE 2019).
2. Detection and Classification of Acoustic Scenes and Events challenge (DCASE 2016).
3. Fourth community-based Signal Separation Evaluation Campaign (SiSEC 2013).
4. Third community-based Signal Separation Evaluation Campaign (SiSEC 2011).
5. The PASCAL 'CHiME' Speech Separation and Recognition Challenge (CHiME 2011).
6. Second community-based Signal Separation Evaluation Campaign (SiSEC 2010).
7. First community-based Signal Separation Evaluation Campaign (SiSEC 2008).

LONG TERM STAYS ABROAD

Russia (originated from): 24 years
Czech Republic: 1.5 years
Sweden : 1 year

LANGUAGES

Russian : native
French, English : fluent
Czech : speaking
Swedish, Spanish : beginner

COMPUTER SCIENCE ABILITIES

Languages: C / C++, Python, Matlab, Perl, FORTRAN
Platforms: UNIX / Linux, Windows (95, 98, NT, 2000, XP, etc.)

REVIEWING FOR JOURNALS AND CONFERENCES

[J]	IEEE Signal Processing Magazine (SPM):	2013
[J]	IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP):	2005,07-16
[J]	IEEE Transactions on Signal Processing (TSP):	2013-15
[J]	IEEE Journal of Selected Topics in Signal Processing (J-STSP):	2010,15
[J]	IEEE Signal Processing Letters (SPL):	2011,13,14
[J]	IEEE Transactions on Neural Networks and Learning Systems (TNNLS):	2014,15
[J]	IEEE Transactions on Image Processing (TIP):	2014
[J]	EURASIP Journal on Advances in Signal Processing (Springer):	2013,14
[J]	EURASIP Journal on Audio, Speech, and Music Processing (Springer):	2016
[J]	International Journal of Computer Vision (Springer):	2014,18
[J]	Signal, Image and Video Processing (Springer):	2011-13
[J]	Neural Processing Letters (Springer):	2014
[J]	EURASIP Signal Processing (Elsevier):	2007,09,11,12,14
[J]	Neurocomputing (Elsevier):	2007,13
[J]	Computer Speech and Language (Elsevier):	2012
[J]	Speech Communication (Elsevier):	2013
[J]	Digital Signal Processing (Elsevier):	2013,14
[J]	International Journal of Information and Communication Technology (IJICT):	2014
[C]	Advances In Neural Information Processing Systems (NIPS):	2016
[C]	IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP):	2008,09,10,13-19
[C]	IEEE International Workshop on Machine Learning for Signal Processing (MLSP):	2017,18
[C]	IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA):	2011,13,15
[C]	IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP):	2013
[C]	IEEE International Conference on Electronics, Circuits, and Systems (ICECS):	2012
[C]	IEEE International Conference on Multimedia & Expo (ICME):	2014
[C]	IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM):	2014
[C]	Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA):	2005
[C]	International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA):	2010,12,15
[C]	European Signal Processing Conference (EUSIPCO):	2009-16
[C]	International Conference on Music Information Retrieval (ISMIR):	2007,09,14
[C]	International Workshop on Machine Listening in Multisource Environments (CHiME):	2011,13
[C]	Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA):	2014
[C]	International Conference on Digital Audio Effects (DAFx):	2012
[C]	International Symposium on Chinese Spoken Language Processing (ISCSLP):	2012
[C]	Journées d'Etude sur la Parole (JEP):	2006
[C]	Colloque Gretsi:	2015

PUBLICATIONS

Under revision

1. S. Parekh, S. Essid, A. Ozerov, N. Duong, P. Pérez and G. Richard, "Weakly supervised representation learning for audio-visual scene analysis" *IEEE/ACM Transactions on Audio, Speech and Language Processing* (*accepted with minor revisions*).

Book chapters

1. A. Ozerov, and H. Kameoka, "Gaussian model based multichannel separation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, S. Gannot (Eds.), Wiley, Aug. 2018.
2. A. Ozerov, C. Févotte, and E. Vincent, "An introduction to multichannel NMF for audio source separation," in *Audio Source Separation*, S. Makino (Eds.), Springer, 2018.
3. C. Févotte, A. Ozerov, and E. Vincent, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*, S. Makino (Eds.), Springer, 2018.
4. S. Essid, S. Parekh, N. Q. K. Duong, A. Ozerov, R. Serizel, F. Antonacci, and A. Sarti, "Multiview approaches to event detection and scene analysis," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, D. Ellis, M. Plumbley (Eds.), Springer, 2017.

Journal Articles

1. C. Bilen, A. Ozerov, and P. Pérez, "Solving time domain audio inverse problems using nonnegative tensor factorization," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5604-5617, 2018.
2. S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, pp. 692-730, 2017.
3. D. El Badawy, N. Q. K. Duong and A. Ozerov, "On-the-fly audio source separation - a novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 261-272, 2017.
4. L. Le Magoarou, A. Ozerov and N. Q. K. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, May, 2015.
5. 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*, vol. 21, no. 8, pp. 1699-1712, 2013.
6. A. Ozerov, M. Lagrange and E. Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech and Language*, vol. 27, no. 3, pp. 874?894, 2013.
7. S. Arberet, A. Ozerov, F. Bimbot and R. Gribonval, "A tractable framework for estimating and combining spectral source models for audio source separation," *Signal Processing*, vol. 92, no. 8, pp. 1886-1901, 2012.
8. E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H Sawada, A Ozerov, V. Gowreesunker, D. Lutter, N.Q.K. Duong, "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928-1936, 2012.
9. 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.
10. A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 20, no. 4, pp. 1118-1133, 2012.
11. A. Ozerov and W. B. Kleijn, "Asymptotically optimal model estimation for quantization," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 1031-1042 , April 2011.

12. A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 18, no. 3, pp. 550-563, March 2010. **IEEE SPS 2014 Best Paper Award**
13. 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 Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 5, pp. 1564–1578, July 2007.
14. A. Ozerov, R. Gribonval, P. Philippe and F. Bimbot, "Choix et adaptation de modèles statistiques pour la séparation de voix chantée à partir d'un seul microphone," *Traitement du signal*, vol. 24, no. 3, pp. 211–224, 2007.

Conferences

1. S. Parekh, A. Ozerov, S. Essid, N. Duong, P. Pérez and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'19)*, Mohonk, NY, Oct. 2019.
2. S. Parekh, S. Essid, A. Ozerov, N. Duong, P. Pérez and G. Richard, "Weakly Supervised Representation Learning for Unsynchronized Audio-Visual Events" *Conference on Computer Vision and Pattern Recognition (CVPR'18)* "Sight and Sound" Workshop, Salt Lake City, Utah, USA, June 2018.
3. E. Grinstein, N. Q. K. Duong, A. Ozerov and P Pérez, "Audio style transfer" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18)*, Calgary, Canada, Apr. 2018.
4. A. Ozerov, S. Kitić and P. Pérez, "A comparative study of example-guided audio source separation approaches based on nonnegative matrix factorization" *27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, Japan, Sept. 2017.
5. S. Parekh, S. Essid, A. Ozerov, N. Duong, P. Pérez and G. Richard, "Guiding audio source separation by video object information" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'17)*, Mohonk, NY, Oct. 2017.
6. G. Puy, A. Ozerov, N. Q. K. Duong and P Pérez, "Informed source separation via compressive graph signal sampling" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*, New Orleans, USA, Mar. 2017.
7. S. Parekh, S. Essid, A. Ozerov, N. Duong, P. Pérez and G. Richard, "Motion informed audio source separation" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*, New Orleans, USA, Mar. 2017.
8. N. Q. K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov and L. Chevallier, "Audio zoom for smartphones based on multiple adaptive beamformers," *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)*, Grenoble, France, Feb. 2017.
9. G. Mafra, N. Q. K. Duong, A. Ozerov and P. Pérez, "Acoustic scene classification: An evaluation of an extremely compact feature representation," *Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*, Budapest, Hungary, Sep. 2016.
10. P. Prablanc, A. Ozerov, N. Q. K. Duong and P. Pérez, "Text-informed speech inpainting via voice conversion," *24th European Signal Processing Conference (EUSIPCO 2016)*, Budapest, Hungary, Aug. 2016.
11. A. Ozerov, C. Bilen, and P. Pérez, "Multichannel audio declipping" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*, Shanghai, China, Mar. 2016.
12. C. Bilen, A. Ozerov, and P. Pérez, "Automatic allocation of NTF components for user-guided audio source separation" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*, Shanghai, China, Mar. 2016.
13. C. Bilen, A. Ozerov, and P. Pérez, "Compressive sampling-based informed source separation" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15)*, Mohonk, NY, Oct. 2015.

14. C. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15)*, Mohonk, NY, Oct. 2015.
15. V. Srinivasan, F. Lefebvre, and A. Ozerov, "Shot aggregating strategy for near-duplicate video retrieval," *23rd European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, Sept., 2015.
16. C. Bilen, A. Ozerov, and P. Pérez, "Joint audio inpainting and source separation" *12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Liberec, Czech Republic, Aug. 2015.
17. C. Bilen, A. Ozerov, and P. Pérez, "Audio inpainting, source separation, audio compression. All with a unified framework based on NTF model" *MissData 2015*, Rennes, France, June 2015.
18. D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'15)*, Brisbane, Australia, Apr., 2015.
19. D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," *24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Reims, France, Sept. 2014.
20. A. Ozerov, N. Q. K. Duong, and L. Chevallier, "On monotonicity of multiplicative update rules for weighted nonnegative tensor factorization," *Int. Symposium on Nonlinear Theory and its Applications (NOLTA)*, Luzern, Switzerland, Sept. 2014.
21. N. Q. K. Duong, A. Ozerov and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," *IEEE International Conference on Consumer Electronics (ICCE-Berlin)*, Berlin, Germany, Sept. 2014.
22. S. Kirbiz, A. Ozerov, A. Liutkus and L. Girin, "Perceptual coding-based informed source separation," *22nd European Signal Processing Conference (EUSIPCO 2014)*, Lisbon, Portugal, Sept. 2014.
23. N. Q. K. Duong, A. Ozerov, L. Chevallier and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*, Florence, Italy, May, 2014.
24. L. Le Magoarou, A. Ozerov and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, Southampton, UK, Sept. 2013.
25. A. Ozerov, J.-R. Vigouroux, L. Chevallier and P. Pérez, "On evaluating face tracks in movies," *IEEE International Conference on Image Processing (ICIP 2013)*, Melbourne, Australia, Sept. 2013.
26. A. Bagri, F. Thudor, A. Ozerov and P. Hellier, "A scalable framework for joint clustering and synchronizing multi-camera videos," *European Signal Processing Conference (EUSIPCO 2013)*, Marrakech, Marocco, Sept. 2013.
27. R. Badeau and A. Ozerov, "Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain," *European Signal Processing Conference (EUSIPCO 2013)*, Marrakech, Marocco, Sept. 2013.
28. L. Chevallier, J.-R. Vigouroux , A. Goguey and A. Ozerov, "Facial landmarks localization estimation by cascaded boosted regression," *International Conference on Computer Vision Theory and Application (VISSAP)*, 2013.
29. M. Lagrange, A. Ozerov and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," *13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, Oct. 2012.
30. A. Liutkus, A. Ozerov, R. Badeau and G. Richard, "Spatial coding-based informed source separation", 20th European Signal Processing Conference, Bucharest, Romania, Aug. 2012.

31. M. Li, J. Klejsa, A. Ozerov and W. B. Kleijn, "Audio coding with power spectral density preserving quantization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*, Kyoto, Japan, March, 2012.

32. 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, Oct. 16-19, 2011.

33. A. Ozerov, M. Lagrange and E. Vincent, "GMM-based classification from noisy features," *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, pages 30-35, Florence, Italy, September, 2011.

34. A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, pages 86-87, Florence, Italy, September, 2011.

35. A. Ozerov, C. Févotte, R. Blouet and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, pages 257-260, Prague, May, 2011.

36. C. Blandin, E. Vincent and A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, pages 2616 - 2619, Prague, May, 2011.

37. 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 (LVA/ICA'10)*, pages 33 - 40, Saint-Malo, France, Sep. 27-30, 2010.

38. S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter and N.Q.K. Duong, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010): - Audio source separation -", In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, pages 114 - 122, Saint-Malo, France, Sep. 27-30, 2010.

39. S. Araki, F. Theis, G. Nolte, D. Lutter, A. Ozerov, V. Gowreesunker, H. Sawada and N.Q.K. Duong, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010): - Biomedical source separation -", In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, pages 123 - 130, Saint-Malo, France, Sep. 27-30, 2010.

40. 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 (CMMR 2010)*, 2010.

41. S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation", In *10th International Conference on Information Sciences, Signal Processing and their applications (ISSPA 2010)*, 2010.

42. A. Ozerov, C. Févotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation", In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, pages 121-124, Mohonk, NY, Oct. 18-21, 2009.

43. 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 EUSIPCO, 17th European Signal Processing Conference, Glasgow, Scotland, August 24-28, 2009..

44. 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 (ICASSP'09)*, pages 3137-3140, Taipei, Taiwan, April 19-24, 2009.

45. A. Ozerov and W. B. Kleijn, "Optimal parameter estimation for model-based quantization," In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, pages 2497-2500, Taipei, Taiwan, April 19-24, 2009.

46. S. Arberet, A. Ozerov, R. Gribonval and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation", In *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'09)*, pages 751-758, Paraty, Brazil, March 15-18, 2009.
47. I. Potamitis and A. Ozerov, "Single channel source separation using static and dynamic features in the power domain," In *EUSIPCO, 16th European Signal Processing Conference*, Laussane, Switzerland, August 25-29, 2008.
48. A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," In *IEEE Asilomar Conference on Signals, Systems, and Computers (Asilomar CSSC'07)*, pages 535-539, Pacific Grove, CA, Nov. 4-7, 2007.
49. R. Heusdens, W. B. Kleijn and A. Ozerov, "Entropy-constrained high-resolution lattice vector quantization using a perceptually relevant distortion measure," In *IEEE Asilomar Conference on Signals, Systems, and Computers (Asilomar CSSC'07)*, pages 2075-2079, Pacific Grove, CA, Nov. 4-7, 2007.
50. W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," In *IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'07)*, pages 243 - 246, Mohonk, NY, Oct. 2007.
51. A. Ozerov, P. Philippe, R. Gribonval and F. Bimbot, "One microphone singing voice separation using source-adapted models", In *IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'05)*, pages 90 - 93, Mohonk, NY, Oct. 2005.
52. A. Ozerov, R. Gribonval, P. Philippe and F. Bimbot, "Séparation voix / musique à partir d'enregistrements mono quelques remarques sur le choix et l'adaptation des modèles", In *GRETSI'05 Symposium on Signal and Image Processing*, Louvain-la-Neuve, Belgique, Sept. 2005.
53. G. Gravier, L. Benaroya, A. Ozerov, R. Gribonval and F. Bimbot, "Séparation de sources à partir d'un seul capteur pour la reconnaissance robuste de la parole", In *Journées d'Etude sur la Parole (JEP'04)*, April 2004.

Demos (Show & Tell)

1. N. Q. K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov, and L. Chevallier, "Audio zoom for smartphones based on multiple adaptive beamformers" *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*, Shanghai, China, Mar. 2016.
2. K. McGuinness, R. Aly, K. Chatfield, O. M. Parkhi, R. Arandjelovic, M. Douze, M. Kemman, M. Kleppe, P. van der Kreeft, K. Macquarrie, A. Ozerov, N. E. O'Connor, F. De Jong, A. Zisserman, C. Schmid, P. Perez, "The AXES Research video search system", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*, Florence, Italy, May 4-9, 2014.
3. K. McGuinness, R. Aly, K. Chatfield, O. M. Parkhi, R. Arandjelovic, M. Douze, A. Ozerov, N. E. O'Connor, F. De Jong, A. Zisserman, C. Schmid, P. Perez, "The AXES PRO video search system", *14th International Workshop on Image and Audio Analysis for Multimedia Interactive services (WIAMIS'13)*, Paris, France, July 3-5, 2013.
4. A. Ozerov, C. Févotte and R. Blouet, "The SARAH project: Standardization of High-Definition Audio Remastering", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Mohonk, NY, Oct. 18-21, 2009.

Technical reports

1. A. Ozerov, N. Q. K. 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.
2. A. Ozerov, S. Essid and M. Charbit, "Reconnaissance des instruments dans la musique polyphonique par décomposition NMF et classification SVM", Technical Report TELECOM ParisTech 2009D014, July 2009.

Theses

1. A. Ozerov. "Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix / musique dans les chansons." PhD thesis, University of Rennes 1, 2006.
2. A. Ozerov. "Représentations robustes pour la reconnaissance automatique de la parole". MSc thesis, DESS "Scientific Calculation and Applications", University of Bordeaux 1, 2003.
3. A. Ozerov. "A criterion of nondisappearance of invariant sets satisfying Krasovsky property under C0 perturbations of right part of the system". MSc thesis, department of Ordinary Differential Equations, Mathematics and Mechanics faculty, St. Petersburg State University, 1999.

Patents

- **9 granted** patents (according to Google Patents on 4/10/2019)
  `https://patents.google.com/?q=ozerov&inventor=alexey&status=GRANT`
- about **32 pending** applications (according to Google Patents on 4/10/2019)
  `https://patents.google.com/?q=ozerov&inventor=alexey&status=APPLICATION`