Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain

Roland Badeau, *Senior Member, IEEE,* Mark D. Plumbley, *Senior Member, IEEE*

***Abstract*—Several probabilistic models involving latent components have been proposed for modelling time-frequency (TF) representations of audio signals such as spectrograms, notably in the nonnegative matrix factorization (NMF) literature. Among them, the recent high resolution NMF (HR-NMF) model is able to take both phases and local correlations in each frequency band into account, and its potential has been illustrated in applications such as source separation and audio inpainting. In this paper, HR-NMF is extended to multichannel signals and to convolutive mixtures. The new model can represent a variety of stationary and non-stationary signals, including autoregressive moving average (ARMA) processes and mixtures of damped sinusoids. A fast variational expectation-maximization (EM) algorithm is proposed to estimate the enhanced model. This algorithm is applied to piano signals, and proves capable of accurately modelling reverberation, restoring missing observations, and separating pure tones with close frequencies.**

***Index Terms*—Non-stationary signal modelling, Time-frequency analysis, Nonnegative matrix factorisation, Multichannel signal analysis, Variational EM algorithm.**

I. Introduction

NONNEGATIVE matrix factorisation was originally introduced as a rank-reduction technique, which approximates a non-negative matrix $\boldsymbol{V} \in \mathbb{R}^{F\times T}$ as a product $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$ of two non-negative matrices $\boldsymbol{W} \in \mathbb{R}^{F\times S}$ and $\boldsymbol{H} \in \mathbb{R}^{S\times T}$ with $S < \min(F,T)$ [1]. In audio signal processing, it is often used for decomposing a magnitude or power TF representation, such as a Fourier or a constant-Q transform (CQT) spectrogram. The columns of $\boldsymbol{W}$ are then interpreted as a dictionary of spectral templates, whose temporal activations are represented in the rows of $\boldsymbol{H}$. Several applications to audio have been addressed, such as multi-pitch estimation [2]–[4], automatic music transcription [5], [6], musical instrument recognition [7], and source separation [7]–[10].

In the literature, several probabilistic models involving latent components have been proposed to provide a probabilistic framework to NMF. Such models include NMF with additive Gaussian noise [11], probabilistic latent component analysis (PLCA) [12], NMF as a sum of Poisson components [13], and NMF as a sum of Gaussian components [14]. Although they have already proven successful in a number of audio applications such as source separation [11]–[13] and multipitch estimation [14], most of these models still lack of consistency in some respects.

Firstly, they focus on modelling a magnitude or power TF representation, and simply ignore the phase information. In an application of source separation, the source estimates are then obtained by means of Wiener-like filtering [8]–[10], which consists in applying a mask to the magnitude TF representation of the mixture, while keeping the phase field unchanged. It can be easily shown that this approach cannot properly separate sinusoidal signals lying in the same frequency band, which means that the frequency resolution is limited by that of the TF transform. In other respects, the separated TF representation is generally not consistent, which means that it does not correspond to the TF transform of a temporal signal, resulting in artefacts such as musical noise. Therefore enhanced algorithms are needed to reconstruct a consistent TF representation [15]. In the same way, in an application of model-based audio synthesis, where there is no available phase field to assign to the sources, reconstructing consistent phases requires employing ad-hoc methods [16], [17].

Secondly, these models generally focus on the spectral and temporal dynamics, and assume that all time-frequency bins are independent. This assumption is clearly not relevant in the case of sinusoidal or impulse signals for instance, and it is not consistent with the existence of spectral or temporal dynamics. Indeed, in the case of wide sense stationary (WSS) processes, spectral dynamics (described by the power spectral density) is closely related to temporal correlation (described by the autocovariance function). Reciprocally, in the case of uncorrelated processes (all samples are uncorrelated with different variances), temporal dynamics induces spectral correlation. In other respects, further dependencies in the TF domain may be induced by the TF transform, due to spectral and temporal overlap between TF bins.

In order to overcome the assumption of independent TF bins, Markov models have been introduced for taking the local dependencies between contiguous TF bins of a magnitude or power TF representation into account [18]–[20]. However, these models still ignore the phase information. Conversely, the complex NMF model [21], [22], which was explicitly designed to represent phases alongside magnitudes in a TF representation, is based on a deterministic framework that does not represent statistical correlations. More recently, two probabilistic models have been proposed, which partially take

Roland Badeau is with Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, 37-39 rue Dareau, 75014 Paris, France.

Mark D. Plumbley is with the Centre for Digital Music, Queen Mary University of London, Mile End Road, E14NS London, UK.

the phase information into account. The multichannel NMF presented in [23] is able to exploit phase relationships between different sensors via the mixing matrix, but the phases and correlations of source signals over time and frequency are not modelled. The infinite positive semidefinite tensor factorization method presented in [24] is able to exploit phase information by modelling correlations over frequency bands, but correlations over time frames are still ignored.

Alternatively, the high resolution (HR) NMF model that we introduced in [25], [26], is able to model both phases and correlations over time frames (within frequency bands) in a principled way. We showed that this model offers an improved frequency resolution, able to separate sinusoids within the same frequency band, and an improved synthesis capability, able to restore missing TF observations. It can be used with both complex-valued and real-valued TF representations, such as the short-time Fourier transform (STFT) and the modified discrete cosine transform (MDCT). It also generalizes some popular models, such as the Itakura-Saito NMF model (IS-NMF) [14], autoregressive (AR) processes [27], and the exponential sinusoidal model (ESM), commonly used in HR spectral analysis of time series [27].

In this paper, HR-NMF is extended to multichannel signals and to convolutive mixtures. Contrary to the multichannel NMF [23] where convolution was approximated, convolution is here accurately implemented in the TF domain by following the exact approach proposed in [28]. Consequently, correlations over time frames and over frequency bands are both taken into account. In order to estimate this multichannel HR-NMF model, we propose a fast variational EM algorithm. This paper further develops a previous work presented in [29], by providing a theoretical ground for the TF implementation of convolution.

The paper is structured as follows. The HR-NMF model is first introduced in the time domain, then the filter bank used to compute the TF representation is presented in Section II. We show in Section III how convolutions in the original time domain can be accurately implemented in the TF domain. The multichannel HR-NMF model in the TF domain is presented in Section IV, and the variational EM algorithm is derived in Section V. This model is applied to audio inpainting and source separation in Section VI. Finally, conclusions are drawn in Section VII.

NOTATION

The following notation will be used throughout the paper (words in italics refer to the state space representation):

- z^*: complex conjugate of $z \in \mathbb{C}$;
- m: sensor index (related to the multichannel mixture);
- s: source index (related to the latent components);
- n: time index in the original time domain;
- v_m: observed mixture;
- w_m: additive white Gaussian noise of variance σ_w^2;
- y_{ms}: source images (*output variables*);
- z_s: latent components (*state variables*);
- x_s: latent innovations (*input variables*);
- t: time frame index in the TF domain;
- f: frequency band index in the TF domain;
- τ: time shift of a TF convolution kernel;
- φ: frequency shift of a TF convolution kernel;
- $b_{ms}(f, \varphi, \tau)$: moving average parameters (*output weights*);
- $a_s(f, \tau)$: autoregressive parameters (*transition weights*).

II. FROM TIME DOMAIN TO TIME-FREQUENCY DOMAIN

Before defining HR-NMF in the TF domain in Section IV, we first provide a simple definition of this model in the time domain.

A. HR-NMF in the time domain

The HR-NMF model of a multichannel signal $v_m(n) \in \mathbb{F}$ (where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is defined for all channels $m \in [0 \dots M-1]$ and times $n \in \mathbb{Z}$, as the sum of S source images $y_{ms}(n) \in \mathbb{F}$ plus a Gaussian noise $w_m(n) \in \mathbb{F}$:

$$v_m(n) = w_m(n) + \sum_{s=0}^{S-1} y_{ms}(n). \tag{1}$$

Moreover, each source image $y_{ms}(f,t)$ for any $s \in [0 \dots S-1]$ is defined as

$$y_{ms}(n) = (g_{ms} * x_s)(n), \tag{2}$$

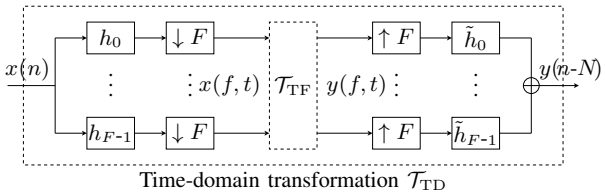
where g_{ms} is the impulse response of a causal and stable recursive filter, and $x_s(n)$ is a Gaussian process[1]. Additionally, processes x_s and w_m for all s and m are mutually independent. In order to make this model identifiable, we will further assume that the spectrum of $x_s(n)$ is flat, because the variability of source s w.r.t. frequency can be modelled within filters g_{ms} for all m. Thus filter g_{ms} represents both the transfer from source s to sensor m and the spectrum of source s.

The purpose of the next sections is to transpose this definition of HR-NMF into the TF domain. The advantages of switching to the TF domain are well-known: in this domain audio signals generally admit a sparse representation, and the overlap of different sound sources is reduced. In Section II-B, we introduce the filter bank notation that will be used in the following developments. Then the accurate implementation of filtering in the TF domain will be addressed in Section III.

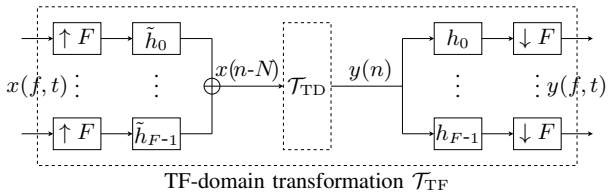
B. Time-frequency analysis: filter bank notation

To perform the time-frequency analysis of a signal, we propose to use the general and flexible framework of perfect reconstruction (PR) filter banks [30], which include both the STFT and MDCT. In the literature, the STFT is often preferred over other existing TF transforms, because under some smoothness assumptions it allows the approximation of linear filtering by multiplying each column of the STFT by the frequency response of the filter. However we will show in Section III that such an approximation is not necessary, and that any PR filter bank will allow us to *accurately* implement convolutions in the TF domain.

[1]The probability distributions of processes $w_m(n)$ and $x_s(n)$ will be defined in the TF domain in Section IV.



(a) Applying a TF transformation to a TD signal



(b) Applying a TD transformation to TF data

Fig. 1. Time-frequency vs. time domain transformations

We thus consider a filter bank [30], which transforms an input signal $x(n) \in l^\infty(\mathbb{F})$ in the original time domain $n \in \mathbb{Z}$ (where $l^\infty(\mathbb{F})$ denotes the space of bounded sequences on $\mathbb{F}$) into a 2D-array $x(f,t) \in l^\infty(\mathbb{F})$ $\forall f \in [0 \dots F-1]$ in the TF domain $(f,t) \in [0 \dots F-1] \times \mathbb{Z}$. More precisely, $x(f,t)$ is defined as

$$x(f,t) = (h_f * x)(Dt), \tag{3}$$

where D is the decimation factor, $*$ denotes standard convolution, and $h_f(n)$ is an analysis filter of support $[0 \dots N-1]$ with $N = LD$ and $L \in \mathbb{N}$. The synthesis filters $\tilde{h}_f(n)$ of same support $[0 \dots N-1]$ are designed so as to guarantee PR. This means that the output, defined as

$$x'(n) = \sum_{f=0}^{F-1} \sum_{t \in \mathbb{Z}} \tilde{h}_f(n - Dt) x(f,t), \tag{4}$$

satisfies $x'(n) = x(n-N)$, which corresponds to an overall delay of N samples. Let

$$H_f(\nu) = \sum_{n \in \mathbb{Z}} h_f(n) e^{-2i\pi\nu n} \tag{5}$$

(with an upper case letter) denote the discrete time Fourier transform (DTFT) of $h_f(n)$ over $\nu \in \mathbb{R}$. Considering that the time supports of $h_f(Dt_1 - n)$ and $h_f(Dt_2 - n)$ do not overlap provided that $|t_1 - t_2| \geq L$, we similarly define a whole number K, such that the overlap between the frequency supports of $H_{f_1}(\nu)$ and $H_{f_2}(\nu)$ can be neglected provided that $|f_1 - f_2| \geq K$, due to high rejection in the stopband.

III. TF IMPLEMENTATION OF CONVOLUTION

In this section, we consider a stable filter of impulse response $g(n) \in l^1(\mathbb{F})$ (where $l^1(\mathbb{F})$ denotes the space of sequences on $\mathbb{F}$ whose series is absolutely convergent) and two signals $x(n) \in l^\infty(\mathbb{F})$ and $y(n) \in l^\infty(\mathbb{F})$, such that $y(n) = (g * x)(n)$. Our purpose is to directly express the TF representation $y(f,t)$ of $y(n)$ as a function of $x(f,t)$, *i.e.* to find a TF transformation $\mathcal{T}_{\mathrm{TF}}$ in Figure 1(a) such that if the input of the filter bank is $x(n)$, then the output is $y(n-N)$ (y is delayed by N samples in order to take the overall delay of the filter bank into account). The following developments further investigate and generalize the study presented in [28], which focused on the particular case of critically sampled PR cosine modulated filter banks. The general case of stable linear filters is first addressed in Section III-A, then the particular case of stable recursive filters is addressed in Section III-B.

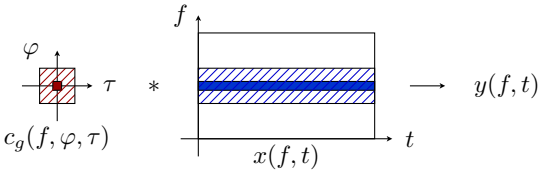


Fig. 2. TF implementation of convolution

A. Stable linear filters

The PR property of the filter bank implies that the relationship between $y(f,t)$ and $x(f,t)$ is given by the transformation $\mathcal{T}_{\mathrm{TF}}$ described in the larger frame in Figure 1(b), where the input is $x(f,t)$, the output is $y(f,t)$, and transformation $\mathcal{T}_{\mathrm{TD}}$ is defined as the time-domain convolution by $g(n+N)$. The resulting mathematical expression is given in Proposition 1.

Proposition 1. *Let $g(n) \in l^1(\mathbb{F})$ be the impulse response of a stable linear filter, and $x(n) \in l^\infty(\mathbb{F})$ and $y(n) \in l^\infty(\mathbb{F})$ two signals such that*

$$y(n) = (g * x)(n). \tag{6}$$

Let $y(f,t)$ and $x(f,t)$ be the TF representations of these signals as defined in Section II-B. Then

$$y(f,t) = \sum_{\varphi \in \mathbb{Z}} \sum_{\tau \in \mathbb{Z}} c_g(f,\varphi,\tau)\ x(f-\varphi, t-\tau) \tag{7}$$

where $\forall f \in [0 \dots F-1]$, $\forall \varphi \in \mathbb{Z}$, $\forall \tau \in \mathbb{Z}$,

$$c_g(f,\varphi,\tau) = (h_f * \tilde{h}_{f-\varphi} * g)(D(\tau + L)), \tag{8}$$

with the convention $\forall f \notin [0 \dots F-1]$, $h_f = 0$.

Proof. Firstly, applying equation (3) to signal y yields

$$y(f,t) = (h_f * y)(Dt). \tag{9}$$

Secondly, equation (4) yields

$$x(n) = \sum_{f=0}^{F-1} \sum_{t \in \mathbb{Z}} \tilde{h}_f(n - D(t-L)) x(f,t). \tag{10}$$

Lastly, equations (7) and (8) are obtained by successively substituting equations (6) and (10) into equation (9). □

Remark 1. As mentioned in Section II-B, if $|\varphi| \geq K$, then frequency bands f and $f - \varphi$ do not overlap, thus $c_g(f,\varphi,\tau)$ can be neglected.

Equation (7) shows that a convolution in the original time domain is equivalent to a 2D-convolution in the TF domain, which is stationary w.r.t. time, and non-stationary w.r.t. frequency, as illustrated in Figure 2.

B. Stable recursive filters

In this section, we introduce a parametric family of TF filters based on a state space representation, and we show a relationship between these TF filters and equation (7).

Definition 1. *Stable recursive filtering in TF domain is defined by the following state space representation:* $\forall f \in [0 \dots F-1], \forall t \in \mathbb{Z}$,

$$\begin{aligned} z(f,t) &= x(f,t) - \sum_{\tau=1}^{Q_a} a_g(f,\tau) z(f,t-\tau) \\ y(f,t) &= \sum_{\varphi=-P_b}^{P_b} \sum_{\tau \in \mathbb{Z}} b_g(f,\varphi,\tau)\ z(f-\varphi,t-\tau) \end{aligned} \quad (11)$$

where $Q_a \in \mathbb{N}$, $P_b \in \mathbb{N}$, *and* $\forall f \in [0 \dots F-1]$, $x(f,t) \in l^\infty(\mathbb{F})$ *is the sequence of input variables,* $z(f,t) \in l^\infty(\mathbb{F})$ *is the sequence of state variables, and* $y(f,t) \in l^\infty(\mathbb{F})$ *is the sequence of output variables. The autoregressive parameters* $a_g(f,\tau) \in \mathbb{F}$ *define a causal sequence of support* $[0 \dots Q_a]$ *w.r.t.* τ *(with* $a_g(f,0) = 1$*), having only simple poles lying inside the unit circle. The moving average parameters* $b_g(f,\varphi,\tau) \in \mathbb{F}$ *define a sequence of finite support w.r.t.* τ*, and* $\forall f \in [0 \dots F-1]$, $\forall \varphi \in [-P_b \dots P_b]$, $b_g(f,\varphi,\tau) = 0$ *provided that* $f - \varphi \notin [0 \dots F-1]$.

Proposition 2. *If* $g(n) \in l^1(\mathbb{F})$ *is the impulse response of a causal and stable recursive filter, then the TF input/output system defined in Proposition 1 admits the state space representation* (11)*, where* $P_b = K-1$ *and* $\forall f \in [0 \dots F-1]$, $\forall \varphi \in [-P_b, P_b]$, $b_g(f,\varphi,\tau)$ *is a sequence of support* $[-L+1 \dots -L+1+Q_b]$ *w.r.t.* τ*, where* $Q_b \geq 2L + Q_a - 1$.

Proposition 2 is proved in Appendix A.

Proposition 3. *In Definition 1, equation* (11) *can be rewritten in the form of equation* (7)*, where* $\forall f \in [0 \dots F-1]$, $\forall \tau \in \mathbb{Z}$, $c_g(f,\varphi,\tau) = 0$ *if* $|\varphi| > P_b$*, and* $\forall f \in [0 \dots F-1]$, $\forall \varphi \in [-P_b \dots P_b]$*, filter* $c_g(f,\varphi,\tau)$ *is defined as the only stable (bounded-input, bounded-output) solution of the following recursion:*

$$\forall \tau \in \mathbb{Z},\ \sum_{t=0}^{Q_a} a_g(f-\varphi,t) c_g(f,\varphi,\tau-t) = b_g(f,\varphi,\tau). \quad (12)$$

Proposition 3 is proved in Appendix A.

Remark 2. In Definition 1, $a_g(f,\tau)$ and $b_g(f,\varphi,\tau)$ are over-parametrised compared to $g(n)$ in Proposition 1. Consequently, if the values of $a_g(f,\tau)$ and $b_g(f,\varphi,\tau)$ are arbitrary, then it is possible that no filter $g(n)$ exists such that equation (8) holds, which means that this state space representation does no longer correspond to a convolution in the original time domain. In this case, we will say that the TF transformation defined in equation (11) is *inconsistent*[2].

[2]In the TF domain HR-NMF model introduced in Section IV, as well as in the variational EM algorithm presented in Section V, the consistency of the filter parameters is not explicitly enforced. In practice, the consistency of the estimated parameters will depend on the observed data itself. If the data is clean and informative enough, then the estimated parameters should be consistent. If the data is noisy and poorly informative (for instance in a frequency band where there is no harmonic partial but only noise), then the estimated parameters may not be consistent. However the impact of this discrepancy on the performance might be rather limited in applications.

IV. Multichannel HR-NMF in TF domain

In this section we present the multichannel HR-NMF model in the TF domain, as initially introduced in [29]. Here this model will be derived from the definition of HR-NMF provided in the time domain in Section II-A.

Following the definition in equation (1), the multichannel HR-NMF model of TF data $v_m(f,t) \in \mathbb{F}$ is defined for all channels $m \in [0 \dots M-1]$, discrete frequencies $f \in [0 \dots F-1]$, and times $t \in [0 \dots T-1]$, as the sum of S source images $y_{ms}(f,t) \in \mathbb{F}$ plus a 2D-white noise

$$w_m(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, \sigma_w^2), \quad (13)$$

where $\mathcal{N}_{\mathbb{F}}(0,\sigma_w^2)$ denotes a real (if $\mathbb{F} = \mathbb{R}$) or circular complex (if $\mathbb{F} = \mathbb{C}$) normal distribution of mean 0 and variance σ_w^2:

$$v_m(f,t) = w_m(f,t) + \sum_{s=0}^{S-1} y_{ms}(f,t). \quad (14)$$

Then Proposition 2 shows how the convolution in equation (2) can be rewritten in the TF domain: the recursive filters g_{ms} can be accurately implemented via equations (15) and (17), which come from Definition 1[3]. Each source image $y_{ms}(f,t)$ for $s \in [0 \dots S-1]$ is thus defined as

$$y_{ms}(f,t) = \sum_{\varphi=-P_b}^{P_b} \sum_{\tau=0}^{Q_b} b_{ms}(f,\varphi,\tau)\ z_s(f-\varphi,t-\tau) \quad (15)$$

where $P_b, Q_b \in \mathbb{N}$, $b_{ms}(f,\varphi,\tau) = 0$ if $f - \varphi \notin [0 \dots F-1]$, and the latent components $z_s(f,t) \in \mathbb{F}$ are defined as follows:

- $\forall t \in [-Q_z \dots -1]$ where $Q_z = \max(Q_b, Q_a)$,

$$z_s(f,t) \sim \mathcal{N}(\mu_s(f,t), 1/\rho_s(f,t)), \quad (16)$$

- $\forall t \in [0 \dots T-1]$,

$$z_s(f,t) = x_s(f,t) - \sum_{\tau=1}^{Q_a} a_s(f,\tau) z_s(f,t-\tau) \quad (17)$$

  where $x_s(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, \sigma_{x_s}^2(t))$, $Q_a \in \mathbb{N}$ and $a_s(f,\tau)$ defines a stable autoregressive filter.

Note that the variance $\sigma_{x_s}^2(t)$ of $x_s(f,t)$ does not depend on frequency f. This particular choice allows us to make the model identifiable, as suggested in Section II-A (the variability w.r.t. frequency is already modelled via the the filters g_{ms}).

The random variables $w_m(f_1,t_1)$ and $x_s(f_2,t_2)$ for all s, m, f_1, f_2, t_1, t_2 are assumed mutually independent. Additionally, $\forall m \in [0 \dots M-1]$, $\forall f \in [0 \dots F-1]$, $\forall t \in [-Q_z \dots -1]$, $v_m(f,t)$ is unobserved, and $\forall s \in [0 \dots S-1]$, the prior mean $\mu_s(f,t) \in \mathbb{F}$ and the prior precision (inverse variance) $\rho_s(f,t) > 0$ of the latent variable $z_s(f,t)$ are considered to be fixed parameters.

The set θ of parameters to be estimated consists of:

- the **autoregressive parameters** $a_s(f,\tau) \in \mathbb{F}$ for $s \in [0 \dots S-1]$, $f \in [0 \dots F-1]$, $\tau \in [1 \dots Q_a]$ (we further define $a_s(f,0) = 1$),

[3]More precisely, compared to the result of Proposition 2, processes $z_s(f,t)$ and $x_s(f,t)$ as defined in Section IV are shifted $L-1$ samples backward, in order to write $b_{ms}(f,\varphi,\tau)$ in a causal form. This does not alter the definition of HR-NMF, since equation (17) is unaltered by this time shift, and $y_{ms}(f,t)$ is unchanged in equation (15).

- the **moving average parameters** $b_{ms}(f,\varphi,\tau) \in \mathbb{F}$ for $m \in [0\dots M-1]$, $s \in [0\dots S-1]$, $f \in [0\dots F-1]$, $\varphi \in [-P_b\dots P_b]$, and $\tau \in [0\dots Q_b]$,
- the **variance parameters** $\sigma_w^2 > 0$ and $\sigma_{x_s}^2(t) > 0$ for $s \in [0\dots S-1]$ and $t \in [0\dots T-1]$.

We thus have $\theta = \{\sigma_w^2, \sigma_{x_s}^2, a_s, b_{ms}\}_{s\in[0\dots S-1], m\in[0\dots M-1]}$. This model encompasses the following special cases:

- If $M=1$, $\sigma_w^2 = 0$ and $P_b = Q_b = Q_a = 0$, then equation (14) reduces to $v_0(f,t) = \sum_{s=0}^{S-1} b_{0s}(f,0,0) x_s(f,t)$, thus $v_0(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, \widehat{V}_{ft})$, where matrix $\widehat{\boldsymbol{V}}$ of coefficients $\widehat{V}_{ft}$ is defined by the NMF $\widehat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$ with $W_{fs} = |b_{0s}(f,0,0)|^2$ and $H_{st} = \sigma_{x_s}^2(t)$. The maximum likelihood estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$ is then equivalent to the minimization of the Itakura-Saito (IS) divergence between matrix $\widehat{\boldsymbol{V}}$ and spectrogram $\boldsymbol{V}$ (where $V_{ft} = |v_0(f,t)|^2$), hence this model is referred to as **IS-NMF** [14].
- If $M=1$ and $P_b = Q_b = 0$, then $v_0(f,t)$ follows the monochannel **HR-NMF** model [25], [26], [31] involving variance σ_w^2, autoregressive parameters $a_s(f,\tau)$ for all $s \in [0\dots S-1]$, $f \in [0\dots F-1]$ and $\tau \in [1\dots Q_a]$, and the NMF $\widehat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$.
- If $S=1$, $\sigma_w^2 = 0$, $P_b = 0$, $\sigma_{x_0}^2(t) = 1\ \forall t \in [0\dots T-1]$, and $\mu_s(f,t) = 0$ and $\rho_s(f,t) = 1\ \forall t \in [-Q_z\dots -1]$, then $\forall m \in [0\dots M-1]$, $\forall f \in [0\dots F-1]$, $v_m(f,t)$ is an autoregressive moving average **(ARMA) process** [27, Section 3.6].
- If $S=1$, $\sigma_w^2 = 0$, $P_b = 0$, $Q_a > 0$, $Q_b = Q_a - 1$, $\forall t \in [-Q_z\dots -1]$, $\mu_0(f,t) = 0$, $\rho_0(f,t) \gg 1$, and $\sigma_{x_0}^2(t) = \mathbb{1}_{\{t=0\}}$ (where $\mathbb{1}_{\mathcal{S}}$ denotes the indicator function of a set $\mathcal{S}$), then $\forall m \in [0\dots M-1]$, $\forall f \in [0\dots F-1]$, $v_m(f,t)$ can be written in the form $v_m(f,t) = \sum_{\tau=1}^{Q_a} \alpha_{m\tau}\, z_\tau(f)^t$, where $z_1(f)\dots z_{Q_a}(f)$ are the roots of the polynomial $z^{Q_a} + \sum_{\tau=1}^{Q_a} a_0(f,\tau) z^{Q_a-\tau}$. This corresponds to the **Exponential Sinusoidal Model (ESM)** commonly used in HR spectral analysis of time series [27], [32].

Because it generalizes both IS-NMF and ESM models to multichannel data, the model defined in equation (14) is called multichannel HR-NMF.

V. Variational EM algorithm

In early works that focused on monochannel HR-NMF [25], [26], in order to estimate the model parameters we proposed to resort to an expectation-maximization (EM) algorithm based on a Kalman filter/smoother. The approach proved to be appropriate for modelling audio signals in applications such as source separation and audio inpainting. However, its computational cost was high, dominated by the Kalman filter/smoother, and prohibitive when dealing with high-dimensional signals.

In order to make the estimation of HR-NMF faster, we then proposed two different strategies. The first approach aimed to improve the convergence rate, by replacing the M-step of the EM algorithm by multiplicative update rules [33]. However we observed that the resulting algorithm presented some numerical stability issues[4]. The second approach aimed to reduce the computational cost, by using a variational EM algorithm, where we introduced two different variational approximations [31]. We observed that the mean field approximation led to both improved performance and maximal decrease of computational complexity.

In this section, we thus generalize the variational EM algorithm based on mean field approximation to the multichannel HR-NMF model introduced in Section IV, as proposed in [29]. Compared to [31], novelties also include a reduced computational complexity and a parallel implementation.

A. Review of variational EM algorithm

Variational inference [34] is now a classical approach for estimating a probabilistic model involving both observed variables v and latent variables z, determined by a set θ of parameters. Let $\mathcal{F}$ be a set of probability density functions (PDFs) over the latent variables z. For any PDF $q \in \mathcal{F}$ and any function $\phi(z)$, we note $\langle\phi\rangle_q = \int \phi(z) q(z) dz$. Then for any set of parameters θ, the *variational free energy* is defined as

$$\mathcal{L}(q;\theta) = \left\langle \ln\left(\frac{p(v,z;\theta)}{q(z)}\right)\right\rangle_q. \tag{18}$$

The variational EM algorithm is a recursive algorithm for estimating θ. It consists of the two following steps at each iteration i:

- Expectation (E)-step (update q):

$$q^\star = \operatorname*{argmax}_{q\in\mathcal{F}} \mathcal{L}(q;\theta_{i-1}) \tag{19}$$

- Maximization (E)-step (update θ):

$$\theta_i = \operatorname*{argmax}_{\theta} \mathcal{L}(q^\star;\theta). \tag{20}$$

In the case of multichannel HR-NMF, θ has been specified in Section IV. We further define $\delta_m(f,t) = 1$ if $v_m(f,t)$ is observed, otherwise $\delta_m(f,t) = 0$, in particular $\delta_m(f,t) = 0$ $\forall (f,t) \notin [0\dots F-1] \times [0\dots T-1]$. The complete set of variables consists of:

- the set v of **observed variables** $v_m(f,t)$ for $m \in [0\dots M-1]$ and for all f and t such that $\delta_m(f,t) = 1$,
- the set z of **latent variables** $z_s(f,t)$ for $s \in [0\dots S-1]$, $f \in [0\dots F-1]$, and $t \in [-Q_z\dots T-1]$.

We use a *mean field approximation* [34]: $\mathcal{F}$ is defined as the set of PDFs which can be factorized in the form

$$q(z) = \prod_{s=0}^{S-1}\prod_{f=0}^{F-1}\prod_{t=-Q_z}^{T-1} q_{sft}(z_s(f,t)). \tag{21}$$

[4]Indeed, the convergence of multiplicative update rules was not proved in [33] (more specifically, there is no theoretical guarantee that the log-likelihood is non-decreasing), whereas the convergence of EM strategies is well established. Besides, as stated in [33], we observed that multiplicative update rules may exhibit some numerical instabilities for small values of the tuning parameter ϵ (the variation of the log-likelihood oscillates instead of monotonically increasing), which was the reason for introducing a more stable tempering approach, that consists in making ε vary from 1 to a lower value over iterations. In this paper, we therefore preferred to use a slower method with guaranteed convergence. It is possible that the convergence rate could be improved in future using multiplicative update rules.

With this particular factorization of $q(z)$, the solution of (19) is such that each PDF q_{sft} is Gaussian:

$$z_s(f,t) \sim \mathcal{N}_{\mathbb{F}}(\overline{z_s}(f,t), \gamma_{z_s}(f,t)).$$

In the following sections, we will use notation $\overline{\phi} = \langle \phi \rangle_q$ and $\gamma_\phi = \langle |\phi - \overline{\phi}|^2 \rangle_q$, for any function ϕ of the latent variables.

B. Variational free energy

Let $\alpha = 1$ if $\mathbb{F} = \mathbb{C}$, and $\alpha = 2$ if $\mathbb{F} = \mathbb{R}$. Let $D_v = \sum_{m=0}^{M-1}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} \delta_m(f,t)$ be the number of observations, and

$$\begin{aligned} I(f,t) &= \mathbb{1}_{\{0\le f<F,\, 0\le t<T\}}, \\ e_{v_m}(f,t) &= \delta_m(f,t)\left(v_m(f,t) - \sum_{s=0}^{S-1} y_{ms}(f,t)\right), \\ x_s(f,t) &= I(f,t)\Big(\sum_{\tau=0}^{Q_a} a_s(f,\tau) z_s(f,t-\tau)\Big). \end{aligned}$$

Then using equations (13) to (16), the joint log-probability distribution $L = \log(p(v,z;\theta))$ of the complete set of variables satisfies

$$\begin{aligned} -\alpha L = & -\alpha\left(\ln(p(v|z;\theta)) + \ln(p(z;\theta))\right) \\ = & (D_v + SF(T+Q_z))\ln(\alpha\pi) \\ & + D_v \ln(\sigma_w^2) + \tfrac{1}{\sigma_w^2}\sum_{m=0}^{M-1}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} |e_{v_m}(f,t)|^2 \\ & + \sum_{s=0}^{S-1}\sum_{f=0}^{F-1}\sum_{t=-Q_z}^{-1} \ln(\tfrac{1}{\rho_s(f,t)}) \\ & + \sum_{s=0}^{S-1}\sum_{f=0}^{F-1}\sum_{t=-Q_z}^{-1} \rho_s(f,t)|z_s(f,t) - \mu_s(f,t)|^2 \\ & + \sum_{s=0}^{S-1}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} \ln(\sigma_{x_s}^2(t)) + \tfrac{1}{\sigma_{x_s}^2(t)}\left|x_s(f,t)\right|^2. \end{aligned}$$

Thus the variational free energy defined in (18) satisfies

$$\begin{aligned} -\alpha\mathcal{L} = & D_v \ln(\alpha\pi) - SF(T+Q_z) \\ & + D_v\ln(\sigma_w^2) + \sum_{m=0}^{M-1}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} \tfrac{\gamma_{e_{v_m}}(f,t) + |\overline{e}_{v_m}(f,t)|^2}{\sigma_w^2} \\ & + \sum_{s=0}^{S-1}\sum_{f=0}^{F-1}\sum_{t=-Q_z}^{-1} -\ln(\rho_s(f,t)\gamma_{z_s}(f,t)) \\ & + \rho_s(f,t)\left(\gamma_{z_s}(f,t) + |\overline{z}_s(f,t) - \mu_s(f,t)|^2\right) \\ & + \sum_{s=0}^{S-1}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} \ln\left(\tfrac{\sigma_{x_s}^2(t)}{\gamma_{z_s}(f,t)}\right) + \tfrac{\gamma_{x_s}(f,t) + |\overline{x}_s(f,t)|^2}{\sigma_{x_s}^2(t)} \end{aligned} \tag{22}$$

where $\forall f \in [0 \dots F-1]$, $\forall t \in [0 \dots T-1]$,

$$\begin{aligned} \gamma_{e_{v_m}}(f,t) &= \delta_m(f,t)\sum_{s=0}^{S-1}\gamma_{y_{ms}}(f,t), \\ \gamma_{y_{ms}}(f,t) &= \sum_{\varphi=-P_b}^{P_b}\sum_{\tau=0}^{Q_b} |b_{ms}(f,\varphi,\tau)|^2 \gamma_{z_s}(f-\varphi,t-\tau), \\ \overline{e}_{v_m}(f,t) &= \delta_m(f,t)\left(v_m(f,t) - \sum_{s=0}^{S-1}\overline{y}_{ms}(f,t)\right), \\ \overline{y}_{ms}(f,t) &= \sum_{\varphi=-P_b}^{P_b}\sum_{\tau=0}^{Q_b} b_{ms}(f,\varphi,\tau)\ \overline{z}_s(f-\varphi,t-\tau), \\ \gamma_{x_s}(f,t) &= I(f,t)\Big(\sum_{\tau=0}^{Q_a} |a_s(f,\tau)|^2\gamma_{z_s}(f,t-\tau)\Big), \\ \overline{x}_s(f,t) &= I(f,t)\Big(\sum_{\tau=0}^{Q_a} a_s(f,\tau)\overline{z}_s(f,t-\tau)\Big). \end{aligned}$$

C. Variational EM algorithm for multichannel HR-NMF

According to the mean field approximation, the maximizations in equations (19) and (20) are performed for each scalar parameter in turn [34]. The dominant complexity of each iteration of the resulting variational EM algorithm is $4MFST\Delta f\Delta t$, where $\Delta f = 1 + 2P_b$ and $\Delta t = 1 + Q_z$ (by updating the model parameters in turn rather than jointly, the complexity of the M-step has been divided by a factor $(\Delta t)^2$ compared to [31]). However we highlight a possible parallel implementation, by making a difference between **parfor** loops which can be implemented in parallel, and **for** loops which have to be implemented sequentially.

1) E-step: For all $s \in [0 \dots S-1]$, $f \in [0 \dots F-1]$, $t \notin [-Q_z, -1]$, let $\rho_s(f,t) = 0$. Considering the mean field approximation (21), the E-step defined in equation (19) leads to the updates described in Table I (where $*$ denotes complex conjugation). Note that $\overline{z}_s(f,t)$ has to be updated after $\gamma_{z_s}(f,t)$.

2) M-step: The M-step defined in (20) leads to the updates described in Table II. The updates of the four parameters can be processed in parallel.

---

parfor $s \in [0 \dots S-1]$, $f \in [0 \dots F-1]$, $t \in [-Q_z \dots T-1]$ **do**

$$\begin{aligned} \gamma_{z_s}(f,t)^{-1} = & \ \rho_s(f,t) + \sum_{\tau=0}^{Q_a} \tfrac{I(f,t+\tau)|a_s(f,\tau)|^2}{\sigma_{x_s}^2(t+\tau)} \\ & + \sum_{m=0}^{M-1}\sum_{\varphi=-P_b}^{P_b}\sum_{\tau=0}^{Q_b} \tfrac{\delta_m(f+\varphi,t+\tau)|b_{ms}(f+\varphi,\varphi,\tau)|^2}{\sigma_w^2} \end{aligned}$$

end parfor

for $s \in [0 \dots S\text{-}1]$, $f_0 \in [0 \dots \Delta f\text{-}1]$, $t_0 \in [-Q_z \dots -Q_z{+}\Delta t\text{-}1]$ **do**

parfor $\frac{f-f_0}{\Delta f} \in [0 \dots \lfloor \frac{F-1-f_0}{\Delta_f} \rfloor]$, $\frac{t-t_0}{\Delta t} \in [0 \dots \lfloor \frac{T-1-t_0}{\Delta t} \rfloor]$ **do**

$$\begin{aligned} \overline{z}_s(f,t) = & \ \overline{z}_s(f,t) - \gamma_{z_s}(f,t)\Big(\rho_s(f,t)(\overline{z}_s(f,t) - \mu_s(f,t)) \\ & + \sum_{\tau=0}^{Q_a} \tfrac{a_s(f,\tau)^*\,\overline{x}_s(f,t+\tau)}{\sigma_{x_s}^2(t+\tau)} \\ & - \sum_{m=0}^{M-1}\sum_{\varphi=-P_b}^{P_b}\sum_{\tau=0}^{Q_b} \tfrac{b_{ms}(f+\varphi,\varphi,\tau)^*\,\overline{e}_{v_m}(f+\varphi,t+\tau)}{\sigma_w^2}\Big) \end{aligned}$$

end parfor

end for

TABLE I
E-STEP OF THE VARIATIONAL EM ALGORITHM

VI. Simulation results

In this section, we present a basic proof of concept of the multichannel HR-NMF model. The ability to accurately model reverberation and restore missing observations is illustrated in Section VI-A, and the ability to separate pure tones with close frequencies is illustrated in Section VI-B.

A. Audio inpainting

The following experiments deal with a single source ($S = 1$) formed of a real piano sound sampled at 11025 Hz. A 1.25ms-short stereophonic signal ($M = 2$) has been synthesized by filtering the monophonic recording of a loud C3 piano note from the MUMS database [35] with two room impulse responses

$\sigma_w^2 = \frac{1}{D_v} \sum_{m=0}^{M-1} \sum_{f=0}^{F-1} \sum_{t=0}^{T-1} \gamma_{e_{v_m}}(f,t) + |\overline{e}_{v_m}(f,t)|^2$

parfor $s \in [0 \dots S-1]$, $t \in [0 \dots T-1]$ **do**

$\sigma_{x_s}^2(t) = \frac{1}{F} \sum_{f=0}^{F-1} \gamma_{x_s}(f,t) + |\overline{x}_s(f,t)|^2$

end parfor

for $\tau \in [1 \dots Q_a]$ **do**

parfor $s \in [0 \dots S-1]$, $f \in [0 \dots F-1]$ **do**

$$a_s(f,\tau) = \frac{\sum_{t=0}^{T-1} \frac{1}{\sigma_{x_s}^2(t)} \left(\overline{z}_s(f,t-\tau)^* (a_s(f,\tau)\overline{z}_s(f,t-\tau) - \overline{x}_s(f,t)) \right)}{\sum_{t=0}^{T-1} \frac{1}{\sigma_{x_s}^2(t)} \left(\gamma_{z_s}(f,t-\tau) + |\overline{z}_s(f,t-\tau)|^2 \right)}$$

end parfor

end for

for $s \in [0 \dots S-1]$, $\varphi \in [-P_b \dots P_b]$, $\tau \in [0 \dots Q_b]$ **do**

parfor $m \in [0 \dots M\text{-}1]$, $f \in [\max(0,\varphi) \dots F\text{-}1{+}\min(0,\varphi)]$ **do**

$$b_{ms}(f,\varphi,\tau) = \frac{\sum_{t=0}^{T-1} \overline{z}_s(f\text{-}\varphi,t\text{-}\tau)^* \left(\delta_m(f,t) b_{ms}(f,\varphi,\tau) \overline{z}_s(f\text{-}\varphi,t\text{-}\tau) + \overline{e}_{v_m}(f,t) \right)}{\sum_{t=0}^{T-1} \delta_m(f,t) \left(\gamma_{z_s}(f\text{-}\varphi,t\text{-}\tau) + |\overline{z}_s(f\text{-}\varphi,t\text{-}\tau)|^2 \right)}$$

end parfor

end for

TABLE II
M-STEP OF THE VARIATIONAL EM ALGORITHM

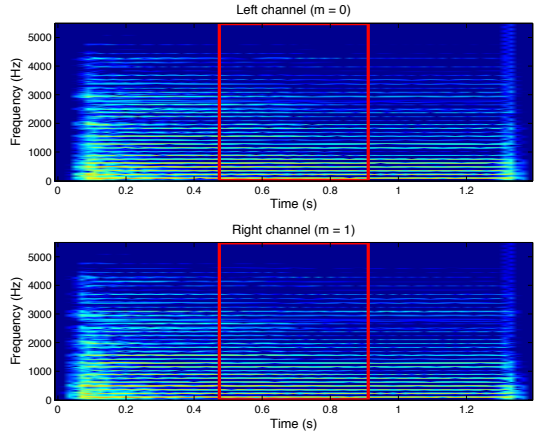


Fig. 3. Input stereo signal $v_m(f,t)$.

simulated using the Matlab code presented in [36][5]. The TF representation $v_m(f,t)$ of this signal has then been computed by applying a critically sampled PR cosine modulated filter bank ($\mathbb{F} = \mathbb{R}$) with $F = 201$ frequency bands, involving filters of length $8F = 1608$ samples. The resulting TF representation, of dimension $F \times T$ with $T = 77$, is displayed in Figure 3. In particular, it can be noticed that the two channels are not synchronous (the starting time in the left channel is ≈ 0.04s, whereas it is ≈ 0.02s in the right channel), which suggests that the order Q_b of filters $b_{ms}(f,\varphi,\tau)$ should be chosen greater than zero.

In the following experiments, we have set $\mu_s(f,t) = 0$ and $\rho_s(f,t) = 10^5$. These values force $\overline{z}_s(f,t)$ to be close to zero $\forall t \in [-Q_z \dots -1]$ (since the prior mean and variance of $\overline{z}_s(f,t)$ are $\mu_s(f,t) = 0$ and $1/\rho_s(f,t) = 10^{-5}$), which is relevant if the observed sound is preceded by silence. The variational EM algorithm is initialized with the neutral values $\overline{z}_s(f,t) = 0$, $\gamma_{z_s}(f,t) = \sigma_w^2 = \sigma_{x_s}^2(t) = 1$, $a_s(f,\tau) = \mathbb{1}_{\{\tau=0\}}$, and $b_{ms}(f,\varphi,\tau) = \mathbb{1}_{\{\varphi=0,\tau=0\}}$. In order to illustrate the capability of the multichannel HR-NMF model to synthesize realistic audio data, we address the case of missing observations. We suppose that all TF points within the red frame in Figure 3 are unobserved: $\delta_m(f,t) = 0$ $\forall t \in [26 \dots 50]$ (which corresponds to the time range 0.47s-0.91s), and $\delta_m(f,t) = 1$ for all other t in $[0 \dots T-1]$. In each experiment, 100 iterations of the algorithm are performed, and the restored signal is returned as $\overline{y}_{ms}(f,t)$.

In the first experiment, a multichannel HR-NMF with $Q_a = Q_b = P_b = 0$ is estimated. Similarly to the example provided in Section IV, this is equivalent to modelling the two channels by two rank-1 IS-NMF models [14] having distinct spectral atoms $\boldsymbol{W}$ and sharing the same temporal activation $\boldsymbol{H}$, or by a rank 1 multichannel NMF [23]. The resulting TF representation $\overline{y}_{ms}(f,t)$ is displayed in Figure 4. It can be noticed that wherever $v_m(f,t)$ is observed ($\delta_m(f,t) = 1$), $\overline{y}_{ms}(f,t)$ does not accurately fit $v_m(f,t)$ (this is particularly visible in high frequencies), because the length Q_b of filters $b_{ms}(f,\varphi,\tau)$ has been underestimated: the source to distortion ratio (SDR)[6] in the observed area is 11.7dB. In other respects, the missing observations ($\delta_m(f,t) = 0$) could not be restored ($\overline{y}_{ms}(f,t)$ is zero inside the frame, resulting in an SDR of 0dB in this area), because the correlations between contiguous TF coefficients in $v_m(f,t)$ have not been taken into account.

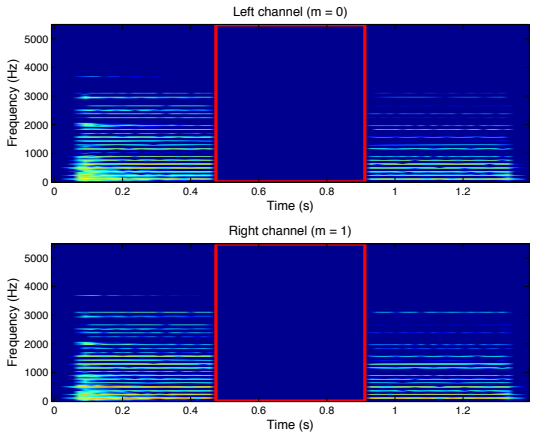


Fig. 4. Stereo signal $\overline{y}_{ms}(f,t)$ estimated with filters of length 1.

In the second experiment, a multichannel HR-NMF model with $Q_a = 2$, $Q_b = 3$, and $P_b = 1$ is estimated. The resulting TF representation $\overline{y}_{ms}(f,t)$ is displayed in Figure 5. It can be noticed that wherever $v_m(f,t)$ is observed, $\overline{y}_{ms}(f,t)$ better fits $v_m(f,t)$: the SDR is 36.8dB in the observed area. Besides, the missing observations have been better estimated: the SDR is 4.8dB inside the frame. Actually, choosing $P_b > 0$ was

[5]Those impulse responses were simulated using 15625 virtual sources. The dimensions of the room were [20m, 19m, 21m], the coordinates of the two microphones were [19m, 18m 1.6m] and [15m, 11m, 10m], and those of the source were [5m, 2m, 1m]. The reflection coefficient of the walls was 0.3.

[6]The SDR between a data vector v and an estimate $\hat{v}$ is defined as $20 \log_{10} \left(\frac{\|v\|_2}{\|v - \hat{v}\|_2} \right)$, where $\|.\|_2$ denotes the Euclidean norm.

necessary to obtain this result, which means that the spectral overlap between frequency bands cannot be neglected in this multichannel setting.

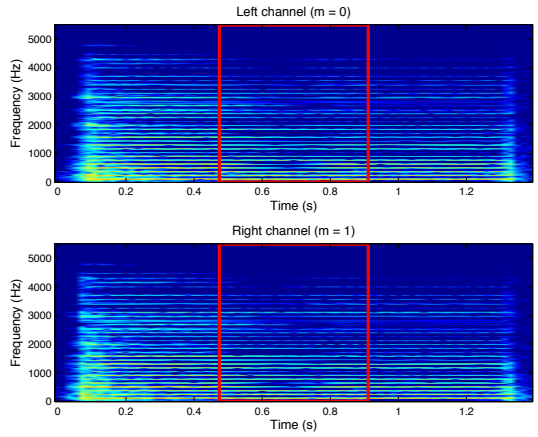


Fig. 5. Stereo signal $\overline{y}_{ms}(f,t)$ estimated with longer filters.

B. Source separation

In this section, we aim to illustrate the ability of HR-NMF to separate pure tones with close frequencies, based on the autoregressive parameters $a_s(f,\tau)$, in a difficult underdetermined setting ($M < S$)[7]. For simplicity, we have chosen to deal with a 2s-long monophonic mixture ($M = 1$), composed of a chord of $S = 2$ piano notes, one semitone apart (A4 and Ab4 from the MAPS database [37][8], whose fundamental frequencies are 440 Hz and 415.30 Hz), resampled at 8600 Hz. The TF representation $v_0(f,t)$ of this mixture signal was computed via an STFT ($\mathbb{F} = \mathbb{C}$), involving 90 ms-long Hann windows with 75% overlap, $F = 400$ frequency bands and $T = 87$ time frames. Here the full TF representation displayed in Figure 6 is observed ($\delta_0(f,t) = 1$). In this experiment, we compare the signals separated by means of the HR-NMF model in two configurations. In the first configuration, $Q_a = Q_b = P_b = 0$ and $\sigma_w^2 = 0$, which means that each source follows a rank-1 IS-NMF model. In the second configuration, $Q_a = 1$ and $Q_b = P_b = 0$, which permits us to accurately model pure tones by means of the autoregressive parameters $a_s(f,\tau)$.

Contrary to the monophonic case ($S = 1$) addressed in Section VI-A, applying the variational EM to multiple sources ($S > 1$) in a fully unsupervised way is difficult: except in some simple settings such as $Q_a = Q_b = P_b = 0$, the algorithm hardly converges to a relevant solution, possibly because of a higher number of local maxima in the variational free energy. Nevertheless, separation of multiple sources is still feasible in a semi-supervised situation, where some parameters are learned beforehand. Here the spectral parameters $a_s(f,\tau)$

[7]In a similar experiment involving a determined multichannel setting ($M = S = 2$), the spatial information proved to be sufficient to accurately separate the two tones, without even using autoregressive parameters ($Q_a = 0$).

[8]MAPS database, ISOL set, ENSTDkCl instrument, mezzo-forte loudness, with the sustain pedal.

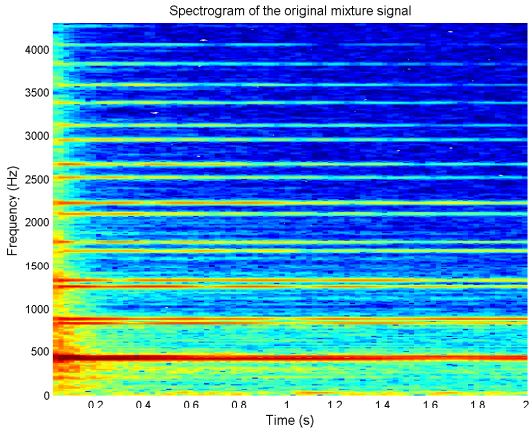


Fig. 6. Spectrogram of the mixture of the A4 and Ab4 piano notes.

and $b_{ms}(f,\varphi,\tau)$ are thus estimated in a first stage from the original source signals. In the first configuration, the values of all NMF parameters are initialized to 1, and 30 iterations of multiplicative update rules [14] are performed. In the second configuration, the variational EM algorithm is initialized with $\mu_s(f,t) = 0$, $\rho_s(f,t) = 10^5$, $\overline{z}_s(f,t) = 0$, $\gamma_{z_s}(f,t) = 1$, $a_s(f,\tau) = \mathbb{1}_{\{\tau=0\}}$, $b_{ms}(f,\varphi,\tau) = \mathbb{1}_{\{\varphi=0,\tau=0\}}$, $\sigma_w^2 = \sigma_{x_s}^2(t) = 1$, and 100 iterations are performed.

In a second stage, the variance parameters $\sigma_{x_s}^2(t)$ and σ_w^2 are estimated from the observed mixture, and the separated signals are obtained as $\overline{y}_{0s}(f,t)$ for $s \in \{0,1\}$. In the first configuration, the spectral parameters learned in the first stage are kept unchanged, the values of the time activations $\sigma_{x_s}^2(t)$ are initialized to 1, and 30 iterations of multiplicative update rules are performed. In the second configuration, the spectral parameters learned in the first stage are kept unchanged, the variational EM algorithm is initialized with the time activations $\sigma_{x_s}^2(t)$ estimated in the first configuration, the value $\sigma_w^2 = 10^{-2}$, and the same initial values of the other parameters as in the first stage. Then 100 iterations of the E-step are performed in order to let $\overline{z}_s(f,t)$ and $\gamma_{z_s}(f,t)$ converge to relevant values based on the learned parameters, and finally 100 iterations of the full variational EM algorithm are performed.

In order to assess the separation performance, we have evaluated the SDR obtained in the two configurations. In the first configuration, the SDR of A4 is 17.67 dB and that of Ab4 is 23.08 dB. In the second configuration, the SDR of A4 is increased to 22.37 dB and that of Ab4 is increased to 27.78 dB. Figure 7 focuses on the results obtained in the frequency band $f = 40$, where the first partials of A4 and Ab4 overlap, resulting in a challenging separation problem. The real parts of the original sources are represented as red solid lines. As expected, the two sources are not properly separated in the first configuration (IS-NMF), because the estimated signals (represented as black dashed lines) are obtained by multiplying the mixture signal by a nonnegative mask, and interferences cannot be cancelled. As a comparison, the signals estimated in the second configuration (represented as blue dots) accurately

fit the partials of the original sources. Note however that this remarkable result was obtained by guiding the variational EM algorithm with relevant initial values. In future work, we will need to develop robust estimation methods, less sensitive to initialisation, in order to perform source separation in a fully unsupervised way.

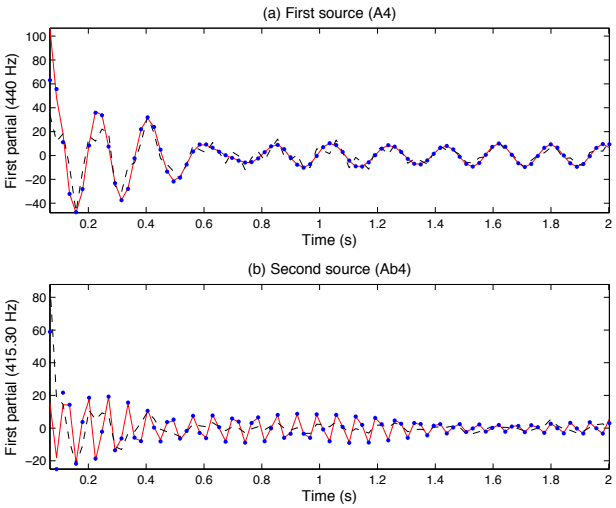


Fig. 7. Separation of two sinusoidal components. The real parts of the two components $\overline{y}_{0s}(f,t)$ are plotted as red solid lines, their IS-NMF estimates are plotted as black dashed lines, and their HR-NMF estimates are plotted as blue dots.

VII. Conclusions

In this paper, we have shown that convolution can be accurately implemented in the TF domain, by applying 2D-filters to a TF representation obtained as the output of a PR filter bank. In the particular case of recursive filters, we have also shown that filtering can be implemented by means of a state space representation in the TF domain. These results have then been used to extend the monochannel HR-NMF model initially proposed in [25], [26] to multichannel signals and convolutive mixtures. The resulting multichannel HR-NMF model can accurately represent the transfer from each source to each sensor, as well as the spectrum of each source. It also takes the correlations over frequencies into account. In order to estimate this model from real audio data, a variational EM algorithm has been proposed, which has a reduced computational complexity and a parallel implementation compared to [31]. This algorithm has been successfully applied to piano signals, and has been capable of accurately modelling reverberation due to room impulse response, restoring missing observations, and separating pure tones with close frequencies.

In order to deal with more realistic music signals, the estimation of the HR-NMF model should be performed in a more informed way, *e.g.* by means of semi-supervised learning, or by using any kind of prior information about the mixture or about the sources. For instance, harmonicity and temporal or spectral smoothness could be enforced by re-parametrising the model, or by introducing some prior distributions of the model parameters. Because audio signals are sparse in the time-frequency domain, we observed that the multichannel HR-NMF model involves a small number of non-zero parameters in practice. In future work, we will investigate enforcing this property, by introducing a prior distribution of the filter parameters such as that proposed in [38], or a prior distribution of the variances of the innovation process $x_s(f,t)$ (modelling variances with a prior distribution is an idea that has been successfully investigated in earlier works [39]–[41]). In other respects, the model could also be extended in several ways, for instance by taking the correlations over latent components into account, or by using other types of TF transforms, *e.g.* wavelet transforms.

Regarding the estimation of the HR-NMF model, the mean field approximation involved in our variational EM algorithm is known to induce a slow convergence rate. The convergence could thus be accelerated by replacing the mean field approximation by a structured mean field approximation, like in [42]. Such an approximation was already proposed to estimate the monochannel HR-NMF model [31], at the expense of a higher computational complexity per iteration. Some alternative Bayesian estimation techniques such as Markov chain Monte Carlo (MCMC) methods and message passing algorithms [34] could also be applied to the HR-NMF model. In other respects, we observed that the variational EM algorithm is hardly able to separate multiple concurrent sources in a fully unsupervised framework, because of its high sensitivity to initialisation. More robust estimation methods are thus needed, which could for instance take advantage of the algebra principles exploited in high resolution methods [32].

Lastly, the multichannel HR-NMF model could be used in a variety of applications, such as source coding, audio inpainting, automatic music transcription, and source separation.

Appendix
TF implementation of stable recursive filtering

Proof of Proposition 2. We consider the TF implementation of convolution given in Proposition 1, and we define $g(n)$ as the impulse response of a causal and stable recursive filter, having only simple poles. Then the partial fraction expansion of its transfer function [43] shows that it can be written in the form $g(n) = g_0(n) + \sum_{k=1}^{Q} g_k(n)$, where $Q \in \mathbb{N}$, $g_0(n)$ is a causal sequence of support $[0 \dots N_0 - 1]$ (with $N_0 \in \mathbb{N}$), and $\forall k \in [1 \dots Q]$,

$$g_k(n) = \beta_k e^{\delta_k n} \cos(2\pi\nu_k n + \psi_k)\mathbb{1}_{n\geq 0}$$

where $\beta_k > 0$, $\delta_k < 0$, $\nu_k \in [0, \frac{1}{2}]$, $\psi_k \in \mathbb{R}$.

Then $\forall f \in [0 \dots F-1]$, equation (8) yields $c_g(f,\varphi,\tau) = \sum_{k=0}^{Q} c_{g_k}(f,\varphi,\tau)$ with

$$c_{g_0}(f,\varphi,\tau) = (h_f * \tilde{h}_{f-\varphi} * g_0)(D(\tau+L))$$

and $\forall k \in [1 \dots Q]$,

$$\begin{aligned} c_{g_k}(f,\varphi,\tau) \quad = \quad & e^{\delta_k D\tau}(A_k(f,\varphi,\tau)\cos(2\pi\nu_k D\tau) \\ & + B_k(f,\varphi,\tau)\sin(2\pi\nu_k D\tau)) \end{aligned}$$

where we have defined

$$\begin{aligned} A_k(f,\varphi,\tau) &= \beta_k \sum_{n=-N+1}^{N-1} (h_f * \tilde{h}_{f-\varphi})(n+N) \\ &\quad \times e^{-\delta_k n} \cos(2\pi\nu_k n - \psi_k) \mathbb{1}_{n\leq D\tau}, \\ B_k(f,\varphi,\tau) &= \beta_k \textstyle\sum_{n=-N+1}^{N-1} (h_f * \tilde{h}_{f-\varphi})(n+N) \\ &\quad \times e^{-\delta_k n} \sin(2\pi\nu_k n - \psi_k) \mathbb{1}_{n\leq D\tau}. \end{aligned}$$

It can be easily proved that $\forall f \in [0 \dots F-1]$, $\forall \varphi \in \mathbb{Z}$,

- the support of $c_{g_0}(f,\varphi,\tau)$ is $[-L+1 \dots L+\lceil \frac{N_0-2}{D} \rceil]$,
- if $\tau \leq -L$, then $c_{g_0}(f,\varphi,\tau)$, $A_k(f,\varphi,\tau)$ and $B_k(f,\varphi,\tau)$ are zero, thus $c_g(f,\varphi,\tau)=0$,
- if $\tau \geq L$, then $A_k(f,\varphi,\tau)$ and $B_k(f,\varphi,\tau)$ do not depend on τ.

Therefore $\forall f \in [0 \dots F-1]$, $\forall \varphi \in \mathbb{Z}$, $c_g(f,\varphi,\tau-L+1)$ is the impulse response of a causal and stable recursive filter, whose transfer function has a denominator of order $2Q$ and a numerator of order $2L+2Q-1+\lceil \frac{N_0-2}{D} \rceil]$.

As a particular case, suppose that $\forall k \in [1 \dots Q]$, $|\delta_k| \ll 1$. If $\tau \geq L$, then $A_k(f,\varphi,\tau)$ and $B_k(f,\varphi,\tau)$ can be neglected as soon as ν_k does not lie in the supports of both $H_f(\nu)$ and $H_{f-\varphi}(\nu)$, where H_f was defined in equation (5). Thus for each f and φ, there is a limited number $Q(f,\varphi) \leq Q$ (possibly 0) of $c_{g_k}(f,\varphi,\tau)$ which contribute to $c_g(f,\varphi,\tau)$. In the general case, we can still consider without loss of generality that $\forall f \in [0 \dots F-1]$, $\forall \varphi \in \mathbb{Z}$, there is a limited number $Q(f,\varphi) \leq Q$ of $c_{g_k}(f,\varphi,\tau)$ which contribute to $c_g(f,\varphi,\tau)$. We then define $Q_a = 2\max_{f,\varphi} Q(f,\varphi)$ and $Q_b = 2L+Q_a-1+\lceil \frac{N_0-2}{D} \rceil$. Then $\forall f \in [0 \dots F-1]$, $\forall \varphi \in \mathbb{Z}$, $c_g(f,\varphi,\tau-L+1)$ is the impulse response of a causal and stable recursive filter, whose transfer function has a denominator of order Q_a and a numerator of order Q_b. Considering Remark 1, we conclude that the input/output system described in equation (7) is equivalent to the state space representation (11), where $P_b = K-1$.

□

Proof of Proposition 3. We consider the state space representation in Definition 1, and we first assume that $\forall f \in [0 \dots F-1]$, sequences $x(f,t)$, $y(f,t)$, and $z(f,t)$ belong to $l^1(\mathbb{Z})$. Then the following DTFTs are well-defined:

$$\begin{aligned} Y(f,\nu) &= \textstyle\sum_{t\in\mathbb{Z}} y(f,t) e^{-2i\pi\nu t}, \\ X(f,\nu) &= \textstyle\sum_{t\in\mathbb{Z}} x(f,t) e^{-2i\pi\nu t}, \\ B_g(f,\varphi,\nu) &= \textstyle\sum_{\tau\in\mathbb{Z}} b_g(f,\varphi,\tau) e^{-2i\pi\nu\tau}, \\ A_g(f,\nu) &= \textstyle\sum_{\tau=0}^{Q_a} a_g(f,\tau) e^{-2i\pi\nu\tau}. \end{aligned}$$

Then applying the DTFT to equation (11) yields $Z(f,\nu) = \frac{1}{A_g(f,\nu)} X(f,\nu)$ and $Y(f,\nu) = \sum_{\varphi=-P_b}^{P_b} B_g(f,\varphi,\nu) Z(f-\varphi,\nu)$. Therefore

$$Y(f,\nu) = \sum_{\varphi=-P_b}^{P_b} C_g(f,\varphi,\nu) X(f-\varphi,\nu), \qquad (23)$$

where

$$C_g(f,\varphi,\nu) = \frac{B_g(f,\varphi,\nu)}{A_g(f-\varphi,\nu)} \qquad (24)$$

is the frequency response of a recursive filter. Since this frequency response is twice continuously differentiable, then this filter is stable, which means that its impulse response $c_g(f,\varphi,\tau) = \int_0^1 C_g(f,\varphi,\nu) e^{+2i\pi\nu\tau} d\nu$ belongs to $l^1(\mathbb{F})$. Equations (7) and (12) are then obtained by applying an inverse DTFT to (23) and (24). Finally, even if $x(f,t)$, $y(f,t)$, and $z(f,t)$ belong to $l^\infty(\mathbb{Z})$ but not to $l^1(\mathbb{Z})$, equations (7) and (11) are still well-defined, and the same filter $c_g(f,\varphi,\tau) \in l^1(\mathbb{F})$ is still the only stable solution of equation (12). □

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their very helpful suggestions. This work was undertaken while Roland Badeau was visiting the Centre for Digital Music, partly funded by EPSRC Platform Grant EP/K009559/1. Mark D. Plumbley is funded by EPSRC Leadership Fellowship EP/G007144/1.

REFERENCES

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

[2] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. 8th International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, Sep. 2007, 6 pages.

[3] P. Smaragdis, "Relative pitch tracking of multiple arbitrary sounds," *Journal of the Acoustical Society of America (JASA)*, vol. 125, no. 5, pp. 3406–3413, May 2009.

[4] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[5] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct. 2003, pp. 177–180.

[6] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.

[7] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, Nov. 2009.

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

[9] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, pp. 1–15, May 2008, article ID 872425.

[10] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Saint Malo, France, Sep. 2010, pp. 498–505.

[11] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with Gaussian process priors," *Computational Intelligence and Neuroscience*, 2008, Article ID 361705, 10 pages.

[12] P. Smaragdis, "Probabilistic decompositions of spectra for sound separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007, pp. 365–386.

[13] T. Virtanen, A. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, Apr. 2008, pp. 1825–1828.

[14] 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.

[15] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[16] D. Griffin and J. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 4, pp. 986–998, Aug. 1983.

[17] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, Sep. 2010, pp. 397–403.

[18] 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)*, New Paltz, New York, USA, Oct. 2009, pp. 121–124.

[19] O. Dikmen and A. T. Cemgil, "Gamma Markov random fields for audio source modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 589–601, Mar. 2010.

[20] G. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proc. 9th international Conference on Latent Variable Analysis and Signal Separation (LCA/ICA)*, St. Malo, France, Sep. 2010, 8 pages.

[21] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3437–3440.

[22] J. Le Roux, H. Kameoka, E. Vincent, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF under spectrogram consistency constraints," in *Proc. Acoustical Society of Japan Autumn Meeting*, no. 2-4-5, Sep. 2009, 2 pages.

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

[24] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Proc. 30th International Conference on Machine Learning (ICML)*, Atlanta, USA, Jun. 2013, pp. 576–584.

[25] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, USA, Oct. 2011, pp. 253–256.

[26] ——, "High resolution NMF for modeling mixtures of non-stationary signals in the time-frequency domain," Telecom ParisTech, Paris, France, Tech. Rep. 2012D004, Jul. 2012.

[27] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, Aug. 2009.

[28] R. Badeau and M. D. Plumbley, "Probabilistic time-frequency source-filter decomposition of non-stationary signals," in *Proc. 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, 5 pages.

[29] ——, "Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, USA, Oct. 2013, 4 pages.

[30] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[31] R. Badeau and A. Drémeau, "Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 6171–6175.

[32] Y. Hua, A. Gershman, and Q. Cheng, Eds., *High resolution and robust signal processing*, ser. Signal Processing and Communications. CRC Press, 2003.

[33] 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, 5 pages.

[34] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge Univ. Press, 2003.

[35] F. Opolko and J. Wapnick, "McGill University Master Samples," McGill University, Montreal, Canada, Tech. Rep., 1987.

[36] S. G. McGovern, "A model for room acoustics," http://www.sgm-audio.com/research/rir/rir.html.

[37] V. Emiya, N. Bertin, B. David, and R. Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Telecom ParisTech, Paris, France, Tech. Rep. 2010D017, Jul. 2010.

[38] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2154, Aug. 2004.

[39] A. T. Cemgil, S. J. Godsill, P. H. Peeling, and N. Whiteley, *The Oxford Handbook of Applied Bayesian Analysis*. Oxford, UK: Oxford University Press, 2010, ch. Bayesian Statistical Methods for Audio and Music Processing.

[40] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Philadephia, PA, USA, Mar. 2005, pp. 517–520.

[41] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[42] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *Proc. 13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, Sep. 2005, 4 pages.

[43] D. Cheng, *Analysis of Linear Systems*. Reading, MA, USA: Addison-Wesley, 1959.

Roland Badeau (M'02–SM'10) 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. 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 over 20 journal papers, over 60 international conference papers, and 2 patents. 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.

Mark Plumbley (S'88–M'90–SM'12) received the B.A. (honors) degree in electrical sciences and the Ph.D. degree in neural networks from the University of Cambridge, United Kingdom, in 1984 and 1991, respectively. From 1991 to 2001, he was a lecturer at Kings College London. He moved to Queen Mary University of London in 2002, where he is Director of the Centre for Digital Music. His research focuses on the automatic analysis of music and other sounds, including automatic music transcription, beat tracking, and acoustic scene analysis, using methods such as source separation and sparse representations. He is a past chair of the ICA Steering Committee and is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. He is a Senior Member of the IEEE.