Computational Auditory Induction as a Missing-Data Model-Fitting Problem with Bregman Divergence[☆]

Jonathan Le Roux[a,1], Hirokazu Kameoka[b], Nobutaka Ono[a],
Alain de Cheveigné[c], Shigeki Sagayama[a]

[a] *Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*
[b] *NTT Communication Science Laboratories, NTT Corporation, 3-1, Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan*
[c] *Centre National de la Recherche Scientifique, Université Paris 5, and Ecole Normale Supérieure, 29 rue d'Ulm, 75230 Paris Cedex 05, France*

Abstract

The human auditory system has the ability, known as auditory induction, to estimate the missing parts of a continuous auditory stream briefly covered by noise and perceptually resynthesize them. In this article, we formulate this ability as a model-based spectrogram analysis and clustering problem with missing data, show how to solve it using an auxiliary function method, and explain how this method is generally related to the Expectation-Maximization (EM) algorithm for a certain type of divergence measures called Bregman divergences, thus enabling the use of prior distributions on the parameters. We illustrate how our method can be used to simultaneously analyze a scene and estimate missing information with two algorithms: the first, based on non-negative matrix factorization (NMF), performs analysis of polyphonic multi-instrumental musical pieces. Our method allows this algorithm to cope with gaps within the audio data, estimating the timbre of the instruments and their pitch, and reconstructing the missing parts. The second, based on a recently introduced technique for the analysis of complex acoustical scenes called Harmonic-Temporal Clustering (HTC), enables us to perform robust fundamental frequency estimation from incomplete speech data.

Keywords: Auditory induction, Acoustical Scene Analysis, Missing data,

[☆] Part of this work was presented at the 2008 ISCA Workshop on Statistical and Perceptual Audition (SAPA 2008)

[*] Corresponding author.
Email addresses: leroux@hil.t.u-tokyo.ac.jp (J. Le Roux), kameoka@cs.brl.ntt.co.jp (H. Kameoka), onono@hil.t.u-tokyo.ac.jp (N. Ono), Alain.de.Cheveigne@ens.fr (A. de Cheveigné), sagayama@hil.t.u-tokyo.ac.jp (S. Sagayama)

[1] Present address: NTT Communication Science Laboratories, NTT Corporation, 3-1, Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan. E-mail: leroux@cs.brl.ntt.co.jp, Tel.: +81-46-240-4029; fax: +81-46-240-4708

Auxiliary function, Bregman divergence, EM algorithm, Non-negative matrix factorization, Harmonic-Temporal Clustering

1. Introduction

The main goal of Computational Auditory Scene Analysis (CASA) is to enable computers to imitate human auditory segregation abilities. CASA has been an area of intensive research in recent years. Particular attention has been given to solving the so-called "cocktail party problem", the computational counterpart of the "cocktail party effect" (Cherry, 1953; von Helmholtz, 1954), i.e., the ability of the human auditory system to focus on a single talker within a mixture of conversations and background noise. This has lead to the development of methods for multi-pitch estimation, noise canceling, or source separation (Wang and Brown, 2006). Less emphasis has been put on the computational realization of another remarkable ability of the human auditory system, auditory induction. Humans are able, under certain conditions, to estimate the missing parts of a continuous acoustic stream briefly covered by noise, to perceptually resynthesize and clearly hear them (Bregman, 1990; Kashino, 2006; Warren, 1970, 1982). They are thus able to simultaneously analyze an auditory scene, as in the cocktail party effect, in the presence of gaps and to perceive the underlying acoustic events as if the information inside those gaps had not been missing (whether the incomplete stimuli are actually reconstructed at low levels in human perception or not is a different issue, which we shall not address here).

An effective computational counterpart to this ability would have many important engineering applications, from polyphonic music recording analysis and restoration to mobile communications robust to both packet-loss and background noise. Attempts to combine scene analysis on incomplete data and reconstruction of the missing parts are rare, with the notable exception of Ellis (1993, 1996). There have been few attempts to address this problem through a statistical approach.

This article aims at developing such a computational counterpart to auditory induction, by simultaneously performing a decomposition of the magnitude wavelet spectrogram of an acoustical scene with missing or corrupted samples, and filling in the gaps into that spectrogram. Various approaches have emerged recently which attempt to analyze the structure of the spectrogram of an acoustical scene (Kameoka et al., 2007; Schmidt and Mørup, 2006; Smaragdis, 2004), while on the other side gap interpolation techniques have been the subject of research for many years (Achan et al., 2005; Cemgil and Godsill, 2005; Clark and Atlas, 2008; Esquef and Biscainho, 2006; Godsill and Rayner, 1998; Lu et al., 2003; Wolfe and Godsill, 2005). However, only few models so far try to deal with both issues. One example is the framework developed by Reyes-Gomez et al. (2004) that relies on local regularities of the spectrogram. The framework that we introduce can use both local and global regularities.

We show here how statistical models that globally model acoustical scenes can be extended for the analysis of scenes with incomplete data. We first derive

the method for a general class of distortion functions that measure the goodness of fit between the model and the observed data. We then show how, for a particular class of functions called Bregman divergences (Banerjee et al., 2005; Bregman, 1967), the method can be interpreted in terms of the Expectation-Maximization (EM) algorithm, enabling the use of prior distributions on the parameters, for example to enforce local smoothness or other regularities. To illustrate the concept, we apply it to the non-negative matrix factor 2D deconvolution algorithm (NMF2D) (Schmidt and Mørup, 2006), and evaluate its performance on a polyphonic multi-instrumental musical piece: the proposed method is able to analyze the scene in spite of the presence of gaps, i.e., it can estimate the timbre of the instruments, their pitch and the time of their activation, and separate their contributions from those of other instruments, while simultaneously reconstructing the missing parts. We finally show how to apply this method to the Harmonic-Temporal Clustering framework (HTC) (Kameoka et al., 2007; Le Roux et al., 2007a), and how the obtained algorithm can be used to perform robust fundamental frequency (F_0) estimation of speech on incomplete data.

2. Computational Auditory Induction

2.1. Problem setting

We consider the problem of interpolating gaps in audio signals by filling in the gaps in their magnitude spectrogram. We will not consider here the reconstruction of the phase: if the magnitude spectrogram can be accurately reconstructed, other methods could be used to obtain a phase consistent with it (Griffin and Lim, 1984; Le Roux et al., 2008b). We are interested in using local and global regularities in the spectrogram to simultaneously analyze the acoustical scene and fill in gaps that may have occurred into it, or in other words to perform "audio inpainting" by reconstructing missing regions of the spectrogram in the same spirit as what is done in image inpainting (Bertalmio et al., 2000), where diffusion-based (local) and exemplar-based (global) techniques are used to restore missing parts of an image (Criminisi et al., 2004).

This is analogous to what is performed by humans in auditory induction, when for example phonemes deleted from a speech signal and replaced by louder broadband noise can be perceptually synthesized by the brain and subjectively heard as if they were present (Bregman, 1990; Kashino, 2006; Warren, 1970, 1982). The auditory induction phenomenon is a striking illustration of the law of closure of Gestalt psychology, according to which the human perception system has a tendency to close "strong" perceptual forms which are incomplete, such as a circle partially occluded by an irregular form. More generally, it can be considered as an expression of Mach's "economy of thought" (Barlow, 2001), in that it is more rewarding in terms of simplicity of explanation to assume that some parts are occluded but actually present and need to be reconstructed (either at the primitive grouping stage or at higher levels) than to assume that the stimuli is actually composed of several disconnected parts. The fact that

auditory induction does not occur if the phonemes are replaced by silence can be linked to the point made by Bregman (1990) that our perceptual system needs to be shown that some evidence is missing: humans can indeed see figures with actual gaps in them, as with no special hint or trigger mechanism, they have no reason to believe that the "missing" parts are not missing but merely hidden. The localization of the gaps will thus be considered known in the following.

Most previous methods for gap interpolation focus on a local modeling of the signal around the gaps. An important corpus of work is for example based on auto-regressive (AR) modeling, stemming from work by Janssen et al. (1986) on an AR interpolator which alternately maximizes the likelihood w.r.t. the missing data and the model parameters, and later extended by Vaseghi and Rayner (1990) to consider samples which are a pitch period apart, by Rayner and Godsill (1991) to cope with the tendency of the AR interpolator to lead to over-smoothed interpolants whose amplitude decreases at the center of the gap, or by Rajan et al. (1997) to consider time-varying AR processes, among many others. Apart from AR, sinusoidal modeling (McAulay and Quatieri, 1986) was also used by Maher (1993) for missing-data interpolation by performing the interpolation directly on the parameters of the sinusoidal model. More details and references can be found in Veldhuis (1990) and Godsill and Rayner (1998).

While previous works on gap interpolation mainly focus on local regularities and do not attempt to explicitly model and exploit the underlying structure of the scene, our approach is more global and conceptually closer to missing-feature approaches to automatic speech recognition (Barker et al., 2005; Cooke et al., 2001; Raj et al., 2004). Trying to understand speech in the presence of noise can be considered as a particular type of missing-data scene analysis, in which time-frequency regions which are dominated by noise are assumed missing. There as well, the goal is to analyze a scene (recognize its speech content) in spite of the presence of unreliable data, and high-order knowledge given by the acoustic and language models can be exploited to estimate the unreliable parts. We shall refer to the very good reviews by Raj and Stern (2005) and Barker (2006) for more details on these methods.

Assuming we have at hand a statistical framework which globally models an acoustical scene, we explain in this article how to use it on scenes with incomplete data and reconstruct to some extent the missing information. We present two examples of statistical frameworks which can be used in this context, Schmidt and Mørup's NMF2D algorithm (Mørup and Schmidt, 2006; Schmidt and Mørup, 2006), and the Harmonic-Temporal Clustering framework introduced in Kameoka et al. (2007) and Le Roux et al. (2007a).

2.2. General method and applicability

The general idea is simple: given a statistical model that can be matched to observed data, we show how it can be used on incomplete data by iterating between analysis steps and reconstruction steps. Furthermore, if the model is sufficiently specified so as to describe the underlying data everywhere, it can be used to reconstruct the missing parts as well. The procedure goes as follows: during a reconstruction step, the missing data are estimated based on

the current value of the model; during an analysis step, the model is updated based on the data completed during the reconstruction step. We show in the following subsection how this iterative algorithm can be interpreted as using an auxiliary function method to optimize the fitting of the statistical model on the regions where data were actually observed.

Situations in which such an incomplete data framework needs to be used are quite varied. One can cite for example situations where a portion of the power spectrogram

1. is lost, for example after a packet loss during a network communication,
2. has been discarded, for example by a binary mask designed to suppress noise or select a particular speaker inside an acoustical scene,
3. or is simply not observed, for example because it lies outside the observed frequency band or the time interval of analysis.

The method we introduce can be used in general to match a statistical model to incomplete data based on the fitting on observed regions even if the original optimization algorithm was designed to be effective only on complete data, such as Gaussian fitting for example. For reconstruction purposes, however, an important point to ensure is that the statistical model has sufficient "prediction power" to interpolate the missing parts. The capacity to reconstruct the missing-data regions will indeed depend on the design of the model, and especially on the constraints introduced: the only guaranty is to obtain, on the whole domain, a complete model which fits the data were they were observed. If the model is designed in such a way that it inherently encompasses the same regularities as the data that it is supposed to fit, then we can expect that what can be inferred on the missing data, based on these regularities, from the observed parts of the data will naturally be reconstructed by the model in the course of the optimization.

For example, models which enforce continuity constraints would ensure a reconstruction with smooth transitions over the missing-data regions. Decomposition models, such as NMF2D, which use information from the whole domain to build a lower-dimensional representation of the acoustical scene, will ensure a reconstruction that conforms to the underlying representation. We will show in particular in Section 4 how incomplete polyphonic music scenes can be analyzed with NMF2D on the basis of information on the spectro-temporal envelopes of the notes of each instrument gathered from the non-missing portions of the music. Similarly, a model such as HTC, through the use of relevant prior distributions, will lead to reconstructions that are inherently guaranteed to respect Bregman's grouping cues (Bregman, 1990; Kameoka, 2007). We will show in Section 6 in particular that the continuity constraint on the pitch contour of the HTC model enables us to perform robust F_0 estimation on incomplete speech data.

2.3. Auxiliary function method

Suppose one wants to fit a parametric distribution to an observed contour which is incomplete, in the sense that its values are only known on a subset

$I \subset D \subseteq \mathbb{R}^n$, where D is the domain of definition of the problem of interest. Suppose also that if the data were complete, the fitting could be performed (e.g., Gaussian distribution fitting, etc.). Then we show that using an iterative procedure based on the auxiliary function method, the fitting to the incomplete data can also be performed.

Let f be the observed contour, and $g(\cdot; \boldsymbol{\Theta})$ a model parameterized by $\boldsymbol{\Theta}$ such that the fitting of this model to an observed contour defined on the whole domain D can be performed.

We consider a distortion function $d : \mathcal{S} \times \mathcal{S} \to [0, +\infty)$ where $\mathcal{S} \subseteq \mathbb{R}^n$, such that $d(x, y) \geq 0, \forall x, y \in \mathcal{S}$ and equality holds if and only if $x = y$. As this function d is not required to respect the triangle inequality, it is not necessarily a metric. For such a distortion function, we can introduce a measure of the distance between the observed data and the model by integrating d between f and $g(\cdot; \boldsymbol{\Theta})$ on the subset I:

$$\mathcal{L}(\boldsymbol{\Theta}) = \int_I d(f(x), g(x; \boldsymbol{\Theta}))\, dx. \tag{1}$$

In this kind of situation, it is often preferable, instead of defining an "incomplete model" whose estimation may be cumbersome, to try to fall back on a complete data estimation problem. This is what we do here by introducing an auxiliary function. For any function h taking values in $\mathcal{S}$ and defined on $D \setminus I$, let us define

$$\mathcal{L}^+(\boldsymbol{\Theta}, h) = \mathcal{L}(\boldsymbol{\Theta}) + \int_{D \setminus I} d(h(x), g(x; \boldsymbol{\Theta}))\, dx. \tag{2}$$

As the second term on the right-hand side is itself derived from the distortion measure, it is non-negative, and thus

$$\mathcal{L}(\boldsymbol{\Theta}) \leq \mathcal{L}^+(\boldsymbol{\Theta}, h),\ \forall h. \tag{3}$$

Moreover, there is equality in the inequality for $h = g(\cdot; \boldsymbol{\Theta})$.

The minimization procedure can now be described as follows. After initializing $\boldsymbol{\Theta}$ for example by performing the distribution fitting on the observed data completed by 0 on $D \setminus I$, one then iteratively performs the following updates:

Step 1 Estimate h such that $\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}^+(\boldsymbol{\Theta}, h)$:

$$\hat{h} = g(\cdot; \boldsymbol{\Theta}). \tag{4}$$

Step 2 Update $\boldsymbol{\Theta}$ with $\hat{h}$ fixed:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, \mathcal{L}^+(\boldsymbol{\Theta}, \hat{h}). \tag{5}$$

The optimization process is illustrated in Fig. 1.

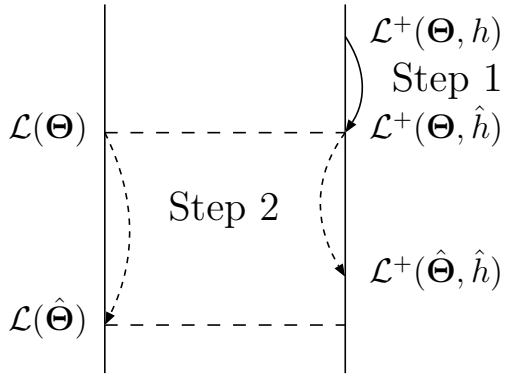


Figure 1: *Optimization through the iterative procedure. During the step 1, the auxiliary parameter h is updated to $\hat{h}$ so that $\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}^+(\boldsymbol{\Theta}, \hat{h})$. Then, during the step 2, $\mathcal{L}^+(\boldsymbol{\Theta}, \hat{h})$ is optimized w.r.t. $\boldsymbol{\Theta}$, ensuring that $\mathcal{L}(\hat{\boldsymbol{\Theta}}) \leq \mathcal{L}^+(\hat{\boldsymbol{\Theta}}, \hat{h}) < \mathcal{L}^+(\boldsymbol{\Theta}, \hat{h}) = \mathcal{L}(\boldsymbol{\Theta})$. The minimization of $\mathcal{L}(\boldsymbol{\Theta})$ can thus be performed through the minimization of the auxiliary function $\mathcal{L}^+(\boldsymbol{\Theta}, h)$ alternately w.r.t. h and $\boldsymbol{\Theta}$.*

3. Probabilistic interpretation for Bregman divergences

We investigate in this section the probabilistic interpretation of the auxiliary function framework introduced above in the particular case where the distortion function is a Bregman divergence.

3.1. Relation between Bregman divergence-based optimization and Maximum-Likelihood estimation

We follow Banerjee et al. (2005) and Grünwald (2007) to give a brief overview of the concepts of exponential family and Bregman divergence and to present the relation between them. As a complete presentation would take us too far from the purpose of the present discussion, we shall refer to them for more details and for rigorous derivations. We tried however to keep this article as self-contained as possible.

Exponential families form a group of probability distributions which comprise many common families of probability distributions such as the normal, gamma, Dirichlet, binomial and Poisson distributions, among others. They are defined as follows.

Definition 1. *Let Λ be an open convex subset of $\mathbb{R}^d$ and let $\mathcal{M} = \{P_\beta \mid \beta \in \Lambda\}$ be a family of probability distributions on a sample space $\mathcal{X}$. $\mathcal{M}$ is an exponential family if there exist a function $\zeta = (\zeta_1, \ldots, \zeta_d) : \mathcal{X} \to \mathbb{R}^d$ and a non-negative function $r : \mathcal{X} \to [0, +\infty)$ such that, for all $\beta \in \Lambda$,*

$$P_{\psi,\beta}(X) \triangleq e^{\langle \beta; \zeta(X) \rangle - \psi(\beta)} r(X), \tag{6}$$

where $\langle \beta; \zeta(X) \rangle$ *is the inner product between* β *and* $\zeta(X)$*, and*

$$\psi(\beta) = \log \int_{\mathcal{X}} \exp(\langle \beta; \zeta(x) \rangle) r(x)\, dx < +\infty.$$

An exponential family defined in terms of a function $\zeta = (\zeta_1, \ldots, \zeta_d)$ *is called a regular exponential family if the representation (6) is minimal, i.e., there exists no* $\alpha_0, \alpha_1, \ldots, \alpha_d \in \mathbb{R}^{d+1} \setminus \{0\}$ *such that for all* x *with* $r(x) > 0$, $\sum_{j=1}^{d} \alpha_j \zeta_j(x) = \alpha_0$.

As an example, we consider the family of Poisson distributions $\{P_\theta \mid \theta \in (0, +\infty)\}$ on the sample space $\mathcal{X} = \mathbb{N}$ defined as $P_\theta(x) = \frac{1}{x!} e^{-\theta} \theta^x$. We see that it is an exponential family, with $\beta = \log \theta$, $\zeta(X) = X$, $\psi(\beta) = e^\beta$, and $r(x) = 1/x!$. The function ζ is not always the identity function as in the Poisson case, as can be seen with the family of normal distributions $\{f_{\mu,\sigma^2} \mid (\mu, \sigma^2) \in \mathbb{R} \times [0, +\infty)\}$ with $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, which can be seen to be an exponential family by setting $\beta = (\mu/\sigma^2, -1/(2\sigma^2))$, $\zeta(X) = (X, X^2)$ and $r(x) = 1/\sqrt{2\pi}$.

It can be shown (Banerjee et al., 2005; Grünwald, 2007) that exponential families can actually be parameterized by the mean value $\mu(\beta) = \mathbb{E}(\zeta(X))$ of $\zeta(X)$. If we let $\Lambda_{\text{mean}} \triangleq \{\mu \mid \exists \beta \in \Lambda \text{ such that } \mu(\beta) = \mu\}$, then $\mu(\cdot)$ is a 1-to-1 mapping from Λ to Λ_{mean}. Moreover, there exists a function $\phi : \Lambda_{\text{mean}} \to \mathbb{R}$ such that for all $\beta \in \Lambda$ and for all $\mu \in \Lambda_{\text{mean}}$ such that $\mu = \mu(\beta)$,

$$\phi(\mu) + \psi(\beta) = \langle \beta; \mu \rangle, \tag{7}$$

from which one can deduce in particular that $\beta(\mu) = \nabla\phi(\mu)$. Altogether, by noticing that

$$\begin{aligned} \langle \beta; \zeta(X) \rangle - \psi(\beta) &= \langle \beta; \mu \rangle - \psi(\beta) + \langle \beta; \zeta(X) - \mu \rangle \\ &= \phi(\mu) + \langle \nabla\phi(\mu); \zeta(X) - \mu \rangle, \end{aligned} \tag{8}$$

$P_{\psi,\beta}$ can be rewritten parameterized by $\mu = \mu(\beta)$, leading to the so-called *mean-value parameterization* of the exponential family:

$$P_{\phi,\mu}(X) \triangleq P_{\psi,\beta(\mu)}(X) = e^{\phi(\mu) - \langle \nabla\phi(\mu); \zeta(X) - \mu \rangle} r(X). \tag{9}$$

We will call μ the expectation parameter of the exponential family, which will be denoted by $\mathcal{F}_\phi$.

We are now ready to introduce the concept of Bregman divergence and to derive its relation with the exponential families.

Definition 2. *Let* $\phi : \mathcal{S} \to \mathbb{R}$ *be a strictly convex function defined on an open convex set* $\mathcal{S} \subseteq \mathbb{R}^d$ *such that* ϕ *is differentiable on* $\mathcal{S}$*. The Bregman divergence* $d_\phi : \mathcal{S} \times \mathcal{S} \to [0, +\infty)$ *is defined as*

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y; \nabla\phi(y) \rangle,$$

where $\nabla\phi(y)$ *is the gradient vector of* ϕ *evaluated at* y*.*

Bregman divergences include a large number of useful loss functions such as squared loss, KL-divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, and the $\mathcal{I}$-divergence. They verify a non-negativity property: $d_\phi(x, y) \geq 0, \forall x, y \in \mathcal{S}$, and equality holds if and only if $x = y$.

Banerjee et al. (2005) showed that the following informal derivation can be rigorously justified for a wide subclass of Bregman divergences, which includes in particular all the loss functions cited above.

If $P_{\phi,\mu}$ is the probability density function of the regular exponential family $\mathcal{F}_\phi$ (in its mean-value parameterization) associated to the function ϕ defining the Bregman divergence d_ϕ, from (9) we have,

$$\begin{aligned} P_{\phi,\mu}(x) &= e^{\phi(\mu)+\langle\nabla\phi(\mu);\zeta(x)-\mu\rangle} r(x) \\ &= e^{-d_\phi(\zeta(x),\mu)+\phi(\zeta(x))} r(x) \end{aligned}$$

and eventually

$$P_{\phi,\mu}(x) = e^{-d_\phi(\zeta(x),\mu)} b_\phi(x) \tag{10}$$

where $b_\phi(x) = e^{\phi(\zeta(x))} r(x)$. This relation holds for all $x \in \mathrm{dom}(\phi)$, which can be shown (Banerjee et al., 2005) to include the set of all the instances that can be drawn from the distribution $P_{\phi,\mu}$. However, one must be careful when using this relation in certain cases where the inclusion is strict, in particular when the support of the carrier $r(x)$ is strictly smaller than $\mathrm{dom}(\phi)$. Indeed, for all x outside that support, Eq. (10) is verified as both members are equal to zero, but it is not informative on the relation between $P_{\phi,\mu}(x)$ and $d_\phi(\zeta(x), \mu)$ as the right-hand side member is zero only because $b_\phi(x)$ is. This is what happens for example for the $\mathcal{I}$-divergence (with $\phi(\mu) = \mu \log \mu - \mu$) for which $\mathrm{dom}(\phi) = \mathbb{R}^+$ (extending the definition of ϕ for $\mu = 0$). The corresponding exponential family is the Poisson family, for which the set of instances and the support of the carrier are only $\mathbb{N}$.

The relation (10) builds a bridge between optimization based on Bregman divergences and Maximum-Likelihood (ML) estimation with exponential families. As distribution-fitting problems usually involve only a first-order statistic, we will focus on the case $\zeta(X) = X$. Trying to fit a model $g(\cdot; \boldsymbol{\Theta})$, defined on a domain D with parameter $\boldsymbol{\Theta}$, to an observed distribution f with a measure of distance between the two based on a Bregman divergence d_ϕ then amounts to looking for $\boldsymbol{\Theta}$ minimizing $\int_D d_\phi(f(x), g(x; \boldsymbol{\Theta}))\, dx$. But according to (10), this is equivalent (up to some precautions which may have to be taken because of the misfit between the domains of definition of the Bregman divergence and the exponential family evoked above) to maximizing w.r.t. $\boldsymbol{\Theta}$ the log-likelihood $\int_D P_{\phi,g(x;\boldsymbol{\Theta})}(f(x))\, dx$ where the observed data points $f(x)$ at point x are assumed to have been independently generated from $P_{\phi,g(x;\boldsymbol{\Theta})}$.

3.2. Relation to the EM algorithm

As we showed above, optimization based on a Bregman divergence corresponds to an ML problem in which the data are supposed to have been generated independently from probability distributions of an associated exponential

family with expectation parameters $g(x, \mathbf{\Theta})$. We investigate here the relation between the application of the EM algorithm to this ML problem and the auxiliary function framework of Section 2.3, in the particular case where the distortion function is a Bregman divergence d_ϕ such that the associated exponential family $\mathcal{F}_\phi$ verifies $\zeta(X) = X$.

The EM algorithm is based on the derivation of a so-called Q-function, which is classically obtained by considering the expectation of the log-likelihood $\log P(f|\mathbf{\Theta})$ of the observed data f against the conditional probability of the unobserved data h with respect to the observed data and the model with parameter $\bar{\mathbf{\Theta}}$:

$$\begin{aligned} \log P(f|\mathbf{\Theta}) &= \mathbb{E}(\log P(f|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\ &= \mathbb{E}(\log P(f,h|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} - \mathbb{E}(\log P(h|f,\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\ &= Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) - H(\mathbf{\Theta},\bar{\mathbf{\Theta}}), \end{aligned} \tag{11}$$

where the functions Q and H were defined in the obvious way from the previous line. One notices through Jensen's inequality that

$$\forall \mathbf{\Theta}, H(\mathbf{\Theta},\bar{\mathbf{\Theta}}) \leq H(\bar{\mathbf{\Theta}},\bar{\mathbf{\Theta}}),$$

such that if one can update $\mathbf{\Theta}$ such that $Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) > Q(\bar{\mathbf{\Theta}},\bar{\mathbf{\Theta}})$, then $\log P(f|\mathbf{\Theta}) > \log P(f|\bar{\mathbf{\Theta}})$.

In the problem we consider, we can show that there is actually a correspondence between the Q-function and the auxiliary function $\mathcal{L}^+$ that we introduced in Section 2.3. The computations of Appendix Appendix A indeed lead to the following relation:

$$Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) = -\mathcal{L}^+(\mathbf{\Theta}, g(x;\bar{\mathbf{\Theta}})) + C(f,\bar{\mathbf{\Theta}}), \tag{12}$$

where $C(f,\bar{\mathbf{\Theta}})$ does not depend on $\mathbf{\Theta}$. Computing the Q-function, i.e., the E-step of the EM algorithm, corresponds to computing the auxiliary function, which is done by replacing the unknown data by the model at the current step. Maximizing the Q-function w.r.t. $\mathbf{\Theta}$, i.e., the M-step of the EM algorithm, corresponds to minimizing the auxiliary function w.r.t. $\mathbf{\Theta}$. This shows how to derive the auxiliary function in an EM point of view, and enables us for example to consider prior distributions on the parameters and perform a MAP estimation.

3.3. Remark on the limitations of this interpretation

We showed that the auxiliary function method in Section 2.3 could be derived through the EM algorithm in the special case of the function d being a Bregman divergence d_ϕ such that the associated exponential family verifies $\zeta(X) = X$. We shall note however that one has to pay attention to the support of the probability distributions of the exponential family. Indeed, as noted earlier, it may happen that these distributions have a smaller support than the original set on which the Bregman divergence is defined. This is for example the case

for the $\mathcal{I}$-divergence, which is defined on $\mathbb{R}^+$ but is associated to the Poisson distribution, whose support is $\mathbb{N}$. The formulation presented in Section 2.3 is thus more general than its EM counterpart, although it does not justify the use of penalty functions as prior distributions on the parameters. In the particular case of the $\mathcal{I}$-divergence, it is actually possible to justify the use of the ML interpretation with real data (Le Roux et al., 2008a).

4. Missing-data non-negative matrix factor 2D deconvolution

4.1. Overview of the original algorithm

The NMF2D algorithm is an extension of Smaragdis's non-negative matrix factor deconvolution (NMFD) (Smaragdis, 2004), itself an extension of the original non-negative matrix factorization (NMF) (Lee and Seung, 1999). NMF is a general tool which attempts to decompose a non-negative matrix $V \in \mathbb{R}^{\geq 0, M \times N}$ in the product of two usually lower-rank non-negative matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$,

$$V \approx WH. \tag{13}$$

In applications to audio, the matrix V to decompose is usually the magnitude or power spectrogram of the observed signal. The horizontal and vertical dimensions of the matrices then respectively represent time and frequency (or log-frequency), and the non-negative factorization of V is expected to lead to a decomposition of the spectrogram in spectral templates W and their activations H. The assumption behind this decomposition is that the spectrogram of an acoustical scene can be modeled as the repetition through time of characteristic spectral templates with varying amplitudes, the shape of these templates being time-invariant, i.e., invariant with the time at which they appear. NMFD extends NMF by introducing a convolution in the time direction, and looks for a decomposition of V as

$$V \approx \Lambda = \sum_{\tau} W^{\tau} \overrightarrow{H}^{\tau} \tag{14}$$

where each W^{τ} is a set of bases and $\rightarrow\tau$ denotes the right-shift operator which moves each element in a matrix τ columns to the right, e.g.,

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \overrightarrow{A}^{1} = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \\ 0 & 7 & 8 \end{pmatrix}.$$

NMFD thus also enables the representation of time structure in the extracted templates W, as, for each k, the set $W_k = (W^{\tau}_{m,k})_{m,\tau}$ of k-th columns of all W^{τ} can be considered as a spectro-temporal template whose activation is determined by the k-th row of H. NMF2D generalizes this approach to the frequency direction through a 2D convolution. Using a log-frequency spectrogram and assuming that the spectral patterns to be modeled are roughly pitch-invariant, i.e., that the spectral patterns of similar sounds differing only by their pitch are approximately equal up to a shift on the frequency axis, NMF2D can account

for the repetition of a similar spectro-temporal structure at various instants and with various frequency shifts as the convolution of a single spectro-temporal template with the information on the time and height of its activations. Concretely, the NMF2D model is

$$V \approx \Lambda = \sum_{\tau} \sum_{\phi} \overset{\downarrow\phi}{W^{\tau}} \overset{\rightarrow\tau}{H^{\phi}} \tag{15}$$

where H^{ϕ} is a set of activations such that the k-th row of H^{ϕ} corresponds to the activations of the k-th spectro-temporal template W_k pitch-shifted by ϕ frequency bins down, and $\downarrow \phi$ denotes the down-shift operator which moves each element in a matrix ϕ lines down. Up-shift and left-shift operators can be introduced in the same way. Applying NMF, NMFD or NMF2D to audio signals implies making a sparseness assumption on the signal, as the additivity of magnitudes in the spectral domain is only true if the underlying components of the signal are sparse enough to minimize overlaps.

Lee and Seung (1999) introduced efficient algorithms for computing the NMF of a matrix V based on both the least-squares error and the $\mathcal{I}$-divergence, which have been extended by Smaragdis for NMFD (Smaragdis, 2004) and Schmidt and Mørup for NMF2D (Mørup and Schmidt, 2006; Schmidt and Mørup, 2006). These algorithms are based on multiplicative updates. If Λ is defined as in (15), we define the objective function as $\mathcal{J}(W, H|V) = \frac{1}{2}\|V - \Lambda\|_F^2$ for the least-squares error, where $\|\cdot\|_F$ denotes the Frobenius norm (sum of the squares of all the elements), or $\mathcal{J}(W, H|V) = \sum_{m,n} V_{m,n} \log\left(\frac{V_{m,n}}{\Lambda_{m,n}}\right) - (V_{m,n} - \Lambda_{m,n})$ for the $\mathcal{I}$-divergence. For the least-squares error, the updates can be written as

$$W^{\tau} \leftarrow W^{\tau} \odot \frac{\sum_{\phi} \overset{\uparrow\phi}{V} \overset{\rightarrow\tau\top}{H^{\phi}}}{\sum_{\phi} \overset{\uparrow\phi}{\Lambda} \overset{\rightarrow\tau\top}{H^{\phi}}}, \quad H^{\phi} \leftarrow H^{\phi} \odot \frac{\sum_{\tau} \overset{\downarrow\phi\top}{W^{\tau}} \overset{\leftarrow\tau}{V}}{\sum_{\tau} \overset{\downarrow\phi\top}{W^{\tau}} \overset{\leftarrow\tau}{\Lambda}}, \tag{16}$$

while for the $\mathcal{I}$-divergence they become

$$W^{\tau} \leftarrow W^{\tau} \odot \frac{\sum_{\phi} \overset{\uparrow\phi}{\left(\frac{V}{\Lambda}\right)} \overset{\rightarrow\tau\top}{H^{\phi}}}{\sum_{\phi} \mathbf{1} \overset{\rightarrow\tau\top}{H^{\phi}}}, \quad H^{\phi} \leftarrow H^{\phi} \odot \frac{\sum_{\tau} \overset{\downarrow\phi\top}{W^{\tau}} \overset{\leftarrow\tau}{\left(\frac{V}{\Lambda}\right)}}{\sum_{\tau} \overset{\downarrow\phi\top}{W^{\tau}} \mathbf{1}}, \tag{17}$$

where $\odot$ denotes the Hadamard product, i.e., element-wise matrix multiplication, matrix divisions are also performed element-wise, $\top$ denotes the matrix transposition and $\mathbf{1}$ denotes a $M \times N$ matrix with all its elements set to 1.

4.2. Sparseness as a key to global structure extraction

As pointed out by Mørup and Schmidt (2006), there is an intrinsic ambiguity in the decomposition (15): the structure of a factor in H can to some extent be put into the signature of the same factor in W and vice versa. One way to alleviate this ambiguity is to impose sparseness on H, thus forcing the structure to go into W. In the case of spectrograms with missing regions, this becomes

even more critical if one expects to retrieve a "meaningful" reconstruction of the missing parts, and sparseness becomes compulsory, as can be clearly seen in the particular case where some time frames are completely missing: indeed, without a sparseness term on the activations, assuming that the spectral envelopes were time- and pitch-invariant (which is only approximately true), a perfect reconstruction of the spectrogram with gaps could be obtained with a single frame representing the instantaneous spectral envelope template in W modulated by the power envelope in the time direction (gaps included) in H. If sparseness of H is enforced, then typical spectro-temporal templates would be learnt in W, and seeing only a part of those templates in the incomplete spectrogram would give us information on their activations' time, pitch and strength, which in turn would enable us to reconstruct the unseen parts. The role of sparseness is thus to ensure that global and recurring structures are extracted and used throughout the spectrogram, and it will be the key that will enable us to fill in the gaps in the underlying acoustical scene, assuming the scene is characterized by the same kind of regularity.

A sparseness-promoting penalty function can be added to the NMF2D objective function, in the form of the L^p (quasi-)norm, for $0 < p < 2$, of the matrix H, or of this quantity raised to the power p (Kameoka et al., 2009; Mørup and Schmidt, 2006). Concretely, we define here a new objective function as

$$\mathcal{J}_s(W, H|V) = \mathcal{J}(W, H|V) + \lambda \sum_{\phi,m,n} |H^{\phi}_{m,n}|^p. \tag{18}$$

The update equations for the minimization of this objective function can be obtained similarly to the ones without the sparseness term through an auxiliary function approach (Lee and Seung, 2001), the sparseness term being dealt with in the same way as in Kameoka et al. (2009) for sparseness-based NMF and complex NMF. Although we shall skip here the derivation, the updates for H become

$$H^{\phi} \leftarrow H^{\phi} \odot \frac{\sum_{\tau} \overset{\downarrow\phi}{W^{\tau}}^{\top} \overset{\leftarrow\tau}{V}}{\sum_{\tau} \overset{\downarrow\phi}{W^{\tau}}^{\top} \overset{\leftarrow\tau}{\Lambda} + \lambda p H^{\phi \bullet (p-1)}} \qquad \text{(least-squares error)} \tag{19}$$

$$H^{\phi} \leftarrow H^{\phi} \odot \frac{\sum_{\tau} \overset{\downarrow\phi}{W^{\tau}}^{\top} \overset{\leftarrow\tau}{\left(\frac{V}{\Lambda}\right)}}{\sum_{\tau} \overset{\downarrow\phi}{W^{\tau}}^{\top} \mathbf{1} + \lambda p H^{\phi \bullet (p-1)}} \qquad (\mathcal{I}\text{-divergence}) \tag{20}$$

where $H^{\phi \bullet (p-1)}$ denotes the matrix H^{ϕ} with all elements raised to the power $(p-1)$. To take care of the fact that the sparseness term could be made arbitrarily small by scaling down H and correspondingly scaling up W, leaving $\mathcal{J}(W, H|V)$ unchanged, a unit-norm constraint is introduced on W. This constraint could be introduced in the update equations using Lagrange multipliers: in the $\mathcal{I}$-divergence case and using L^1-norm normalization, this would simply amount to rescaling W and H appropriately afterwards, while for L^2-norm normalization as well as in the least-squares case, as an analytical solution cannot be obtained, one would need to resort to numerical computations. Here,

we shall use instead update equations for W which are derived in Mørup and Schmidt (2006) by replacing each W_k by $W_k/\|W_k\|_F$ in the definition of the objective function, for both the least-squares and $\mathcal{I}$-divergence cases. This idea was first investigated by Eggert and Körner (2004) for sparse NMF based on least-squares minimization. Although the convergence of these updates is not proven, it is conjectured and has been observed on extensive tests (Eggert and Körner, 2004; Mørup and Schmidt, 2006). We reproduce here the updates for the sake of completeness:

$$W^\tau \leftarrow W^\tau \odot \frac{\sum_\phi \left(\overset{\uparrow\phi}{V} \overset{\rightarrow\tau\top}{H^\phi} + W^\tau \operatorname{diag}(\sum_{\tau'} \mathbf{1}((\overset{\uparrow\phi}{\lambda} \overset{\rightarrow\tau'\top}{H^\phi}) \odot W^{\tau'})) \right)}{\sum_\phi \left(\overset{\uparrow\phi}{\Lambda} \overset{\rightarrow\tau\top}{H^\phi} + W^\tau \operatorname{diag}(\sum_{\tau'} \mathbf{1}((\overset{\uparrow\phi}{V} \overset{\rightarrow\tau'\top}{H^\phi}) \odot W^{\tau'})) \right)}, \quad \text{(least-sq.)} \tag{21}$$

$$W^\tau \leftarrow W^\tau \odot \frac{\sum_\phi \left((\frac{\overset{\uparrow\phi}{V}}{\Lambda}) \overset{\rightarrow\tau\top}{H^\phi} + W^\tau \operatorname{diag}(\sum_{\tau'} \mathbf{1}((\mathbf{1} \overset{\rightarrow\tau'\top}{H^\phi}) \odot W^{\tau'})) \right)}{\sum_\phi \left(\mathbf{1} \overset{\rightarrow\tau\top}{H^\phi} + W^\tau \operatorname{diag}(\sum_{\tau'} \mathbf{1}((\frac{\overset{\uparrow\phi}{V}}{\Lambda}) \overset{\rightarrow\tau'\top}{H^\phi}) \odot W^{\tau'})) \right)}, \quad (\mathcal{I}\text{-div.}) \tag{22}$$

where diag denotes a diagonal matrix whose elements are given by the argument. W is normalized at the beginning of each step, before performing the updates for H and W.

4.3. Use of prior distributions

The NMF framework can be considered in a Bayesian way based on the correspondence between Bregman divergence-based optimization and ML estimation either for the least-squares error or the $\mathcal{I}$-divergence. Indeed, the NMF objective function can be converted into a log-likelihood (Lee and Seung, 1999; Sajda et al., 2003), to which prior constraints on the parameters can further be added (Cemgil, 2008; Févotte et al., 2009).

Sparseness terms evoked above involving L^p (quasi-)norms of H can be considered as such, corresponding in general to generalized Gaussian process priors, and the Laplace distribution for $p = 1$. But one can also introduce Markovian constraints on the parameters to ensure smooth solutions. Using Gamma chains on the coefficients of W and H in the time direction, one can show that analytical update equations can still be obtained and the objective function can be optimized based on the Expectation-Constrained Maximization (ECM) algorithm (Meng and Rubin, 1993).

As explained in Section 3, the auxiliary function method we propose can be interpreted in an EM point of view in the special case of Bregman divergences. The use of prior distributions on the parameters will thus be justified as well for the missing-data version of the NMF2D algorithm, which we describe in more details in the next section, and local convergence will be guaranteed.

4.4. NMF2D on incomplete spectrograms

We consider the wavelet magnitude spectrogram of an acoustical scene represented as a non-negative matrix $V_{m,n}$, defined on a domain of definition $D = [\![1, M]\!] \times [\![1, N]\!]$ (corresponding for example to the time-frequency region $\{x, t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$, sampled in time and frequency). We assume in general that the spectro-temporal patterns to be modeled are roughly pitch-invariant, and that the signals are sparse enough such that the additivity assumption on the magnitude spectrograms holds.

We assume that some regions of the magnitude spectrogram are degraded or missing and are interested in performing simultaneously an analysis of this acoustical scene with the NMF2D algorithm despite the presence of gaps, and a reconstruction of the missing parts.

Even if the data matrix V is incomplete, i.e., if the values $V_{m,n}$ are missing or considered not reliable for some indices $(m, n) \in J \subset D$, due to the fact that the NMF2D update equations (as well as the NMF and NMFD update equations) are in fact multiplicative versions of a gradient update, it would actually be possible to still perform the minimization of the distance taken over the observed data by computing the gradient of this restricted objective function, in the same way as was done in Virtanen et al. (2008) for NMF. However, the formulation of the update equations would then become more intricate and less obvious to interpret, and, although the updates could be originally computed simply and efficiently using FFT thanks to their convolutive nature, their missing-data version would require an additional trick in order to compute them in the same way (concretely, setting to zero the values of the term against which H or W are convolved in the denominators of (16) and (17) where data is actually missing before computing their FFT). In any case, using the method introduced in Section 2 is cleaner and easier to interpret, more systematic and general. Finally, the simplicity and ease of interpretation of NMF2D make it a good example to illustrate the general principle we presented.

Applying the method introduced in Section 2.3 to NMF2D leads to the following algorithm, which can be used to analyze incomplete spectrograms, with both objective functions:

Step 1 $V_{m,n}^{(p+1)} = \begin{cases} \Lambda_{m,n}^{(p)} & \text{if } (m,n) \in J \\ V_{m,n} & \text{if } (m,n) \notin J \end{cases}$

Step 2 Update W through (16) or (17) and H through (19) or (20)

We note that recent work by Smaragdis et al. (2009) can be considered as another illustration of our framework. It is based on a spectrogram model which is very close to NMF. In the same way as we present here, this work relies on the EM algorithm to estimate the parameters on incomplete data, and it is thus very similar to our NMF2D example. As mentioned above for the NMF2D illustration, the optimization could have been solved with multiplicative updates as well, although the EM interpretation is more general and elegant. Several applications to the reconstruction of missing audio data are considered there,

investigating in particular the use of separate training data to improve the estimation of the bases, which we did not consider here. Irregular repartitions of the missing data are also considered, while we focus here on situations where some frames are entirely missing: such situations exploit the convolutive nature of the NMF2D model, relying on the time structure of the bases or on information from other notes appearing at different pitches, and are thus appropriate to illustrate our framework.

5. Examples of application of missing-data NMF2D

5.1. Toy example: reconstructing a 2D image

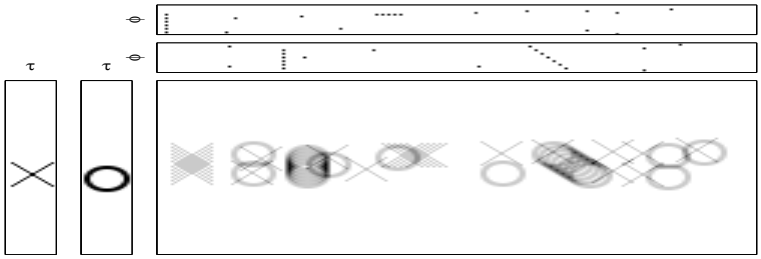
We first tested our algorithm on simulated data used in Mørup and Schmidt (2006). The data, shown in Fig. 2 (a), were created with W consisting of one cross in the first factor and one circle in the second, convolved with H given in the top of the figure to yield the full data matrix V. The NMF2D algorithm was used in the same conditions as in Mørup and Schmidt (2006), with $\tau = \{0, \ldots, 16\}$, $\phi = \{0, \ldots, 16\}$ and an L^1-norm sparseness penalty with coefficient 1 for the least-squares algorithm and 0.001 for the $\mathcal{I}$-divergence algorithm. The circle and cross templates span roughly 15 frames in both horizontal and vertical directions, while the whole data is 200 frames wide. To construct the incomplete data, we erased 3 frames horizontally and 2 frames every 10 frames vertically, as shown in Fig. 2 (b). Note that none of the occurrences of the structures (circle and cross) is fully available. However, in this ideal case where the original data is a strict convolution of the templates, the proposed algorithm is able to extract the original templates and their occurrences and to reconstruct the original data, as can be seen in Fig. 2 (c) (least-squares update equations) and Fig. 2 (d) ($\mathcal{I}$-divergence update equations). This shows that the reconstruction is based on global features of the data learnt by gathering information from the whole domain.

5.2. Audio example: reconstructing gaps in a sound

5.2.1. Experimental setting

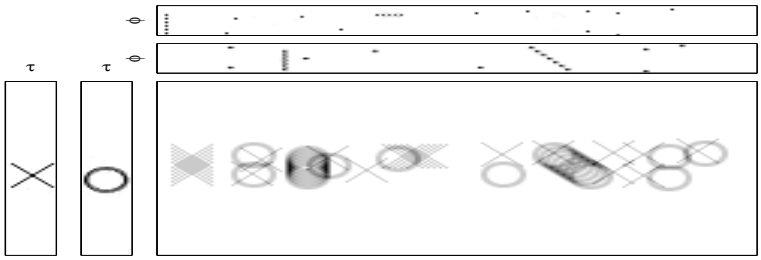
For auditory restoration experiments, contrary to what is done in Schmidt and Mørup (2006), we did not use the short-time Fourier transform afterwards converted into a log-frequency magnitude spectrogram, but a wavelet transform, which directly gives a log-frequency spectrogram. More precisely, the magnitude spectrogram was calculated from the input signals digitized at a 16 kHz sampling rate using a Gabor wavelet transform with a time resolution of 16 ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The frequency range extended from 50 Hz to 8 kHz and was covered by 200 channels, for a frequency resolution of 44 cent.

We used a 4.8 s piece of computer generated polyphonic music containing a trumpet and a piano, already used in Schmidt and Mørup (2006). Its spectrogram can be seen in Fig. 3 (a). The incomplete waveform was built by erasing

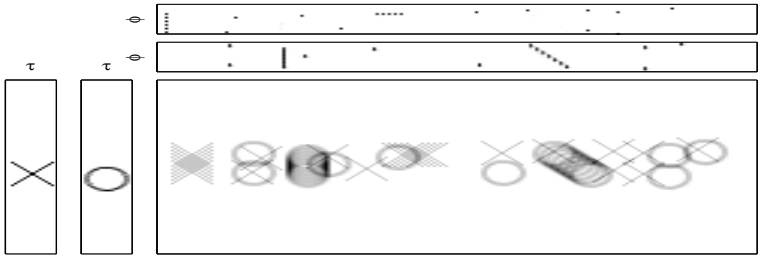


(a) Original simulated data.

(b) Truncated data with truncated regions in black.



(c) Reconstruction using least-squares.



(d) Reconstruction using the $\mathcal{I}$-divergence.

Figure 2: *NMF2D with missing data on a toy problem. (a) Original simulated data. W consists of one cross in the first factor and one circle in the second. They are convolved with H given in the top of the figure to yield the full data matrix V. (b) Truncated data. The truncated areas are indicated in black. (c) Estimated factors and reconstructed image using the least-squares algorithm. (d) Estimated factors and reconstructed image using the $\mathcal{I}$-divergence algorithm.*

80 ms of signal every 416 ms, leading to a signal with about 20 % of data missing. Its spectrogram is shown in Fig. 3 (b).

The mask indicating the region J to inpaint was built according to the erased portions of the waveform. With a Gabor wavelet transform, the influence of a local modification of the signal theoretically spans the whole interval. However, as the windows are Gaussian, one can consider that the influence becomes almost null further than about three times the standard deviation. This standard deviation is inversely proportional with the frequency, and the influence should thus be considered to span a longer interval for lower frequencies. Although it leaves some unreliable portions of the spectrogram out of the mask in the lower frequencies, for simplicity, we did not consider here this dependence on frequency, and simply considered unreliable, for each 80 ms portion of waveform erased, 6 whole spectrogram frames (corresponding to about 96 ms of signal in the highest frequencies). The incomplete spectrogram is shown in Fig. 3 (c), with areas to inpaint in black.

The NMF2D parameters were as follows. As in Schmidt and Mørup (2006), we used two factors, $d = 2$, since we are analyzing a scene with two instruments, and the number of convolutive components in pitch was set to $\phi = \{0, \ldots, 11\}$, as the pitch of the notes in the data spans three whole notes. For the convolutive components in time, we used empirically $\tau = \{0, \ldots, 31\}$, for a time range of about 500 ms, thus roughly spanning the length of the eighth notes in the music sample. The $\mathcal{I}$-divergence was used as the distortion measure, and the L^1 norm as the sparseness term with the coefficient λ set to 0.001. The algorithm was ran for 100 iterations.

5.2.2. Results and discussion

To evaluate the reconstruction accuracy of the spectrogram, we use two measures: Signal to Noise Ratio (SNR), defined as $10 \log_{10}(||\hat{S} - S||^2/||S||^2)$ where S denotes the reference magnitude spectrogram and $\hat{S}$ the reconstructed magnitude spectrogram, and Segmental SNR (SSNR), computed as the median of the individual SNRs of all the frames. We note that computing the SNR directly on the magnitude spectrogram amounts to assuming that the phase is perfectly reconstructed. The results are summarized in Table 1, where "in" refers to the measure computed inside the gaps (the inpainted part), "out" to the measure computed outside the gaps (the part more classically reconstructed based on observed data), "M" refers to the proposed Missing-data NMF2D, "O" to the original NMF2D on the whole data with missing data (if any) assumed to be zero, "C" to the magnitude spectrogram of the complete waveform, and "I" to the one of the incomplete waveform. Finally, "WX" refers to the spectrogram reconstructed by applying algorithm W on spectrogram X, and "Y/Z" to the comparison of spectrogram Y with spectrogram Z as a reference. For example, the SNR of "MI/C" is the SNR of the spectrogram reconstructed using our missing-data approach on the spectrogram of the incomplete data w.r.t. the spectrogram of the full waveform.

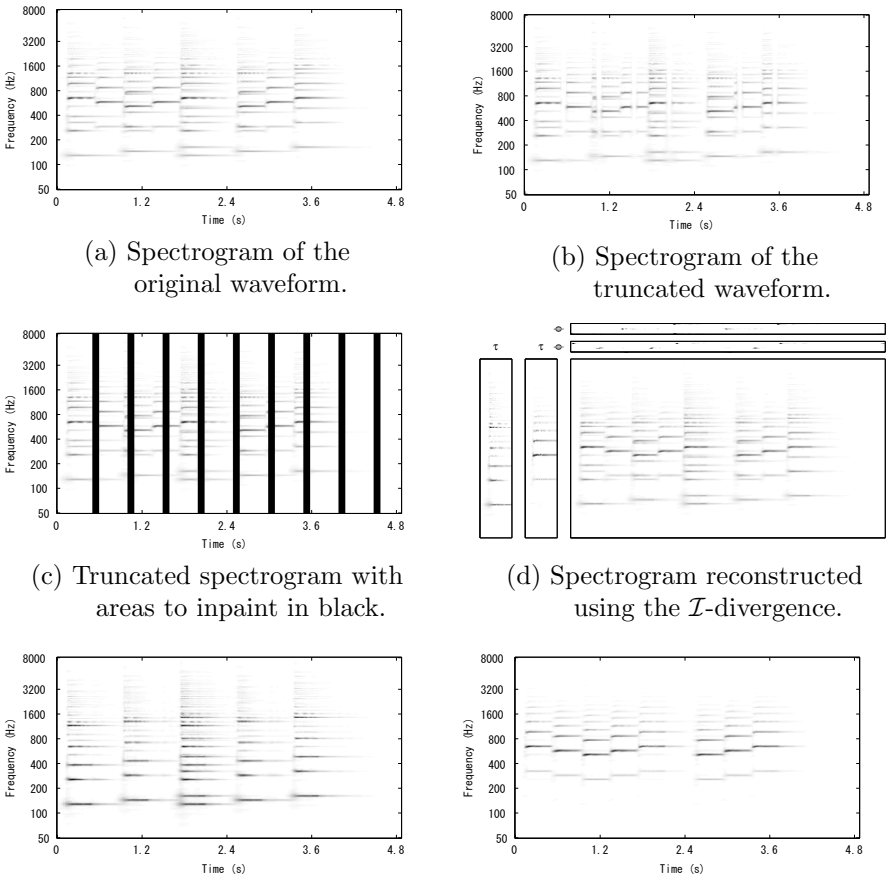
The OC spectrogram is the spectrogram reconstructed by the original method applied to the complete spectrogram C. Comparing it to the complete spectro-

Table 1: *Results of the reconstruction experiment. "in" and "out" refer to the measures computed inside and outside the gaps, respectively; "M" refers to the proposed Missing-data NMF2D, "O" to the original NMF2D on the whole data with missing data (if any) assumed to be zero, "C" to the magnitude spectrogram of the complete waveform, and "I" to that of the incomplete waveform. Finally, "WX" refers to the spectrogram reconstructed by applying algorithm W on spectrogram X, and "Y/Z" to the comparison of spectrogram Y with spectrogram Z as a reference.*

	SNR		SSNR	
	in	out	in	out
OC / C	13.2	13.1	12.6	12.4
I / C	2.5	21.3	2.7	28.2
OI / C	4.3	10.6	4.1	10.1
OI / I	5.8	10.8	7.3	10.5
MI / C	10.6	13.1	10.5	12.1
MI / I	-3.8	13.1	3.3	12.1

gram (OC/C results) thus gives us the modeling accuracy of NMF2D in ideal conditions, and a reference on the performance that we should aim for when trying to analyze the incomplete scene and to reconstruct its missing parts. On the other hand, comparing the spectrogram I of the incomplete waveform to that of the complete waveform inside the gaps (I/C "in" results) indicates the starting point before any reconstruction is done. Let us now look at the performance of the original NMF2D applied to the spectrogram of the incomplete waveform, under the crude assumption that data in the gaps are equal to zero. Comparing the reconstructed spectrogram OI to either that of the complete waveform (OI/C) or the incomplete waveform (OI/I) shows first that a bias is introduced even in the reconstruction of the observed data ("out"), and second that, as expected, the missing data are not reconstructed (OI/C "in"). We finally look at the performance of the proposed missing-data NMF2D also applied to the spectrogram of the incomplete waveform. Comparing the reconstructed spectrogram MI to that of the incomplete waveform (MI/I), we see that the proposed algorithm correctly performs its task of reconstructing the observed data ("out"). This result is important in itself as it shows that the proposed framework enables NMF2D, designed for complete data, to be used on incomplete data without decrease of the performance measured on the observed part of the data, in particular without letting the missing regions introduce a bias in the analysis of the observed regions. One could actually think of applications for which reconstruction of the missing parts may be unnecessary, for example if only the spectro-temporal templates or their activations themselves are desired. Comparing now MI to the spectrogram of the complete waveform (MI/C), we see that the formerly erased regions ("in") are correctly inpainted, with a great improvement over the incomplete spectrogram, as seen earlier with the I/C results, and that our method performs closely to NMF2D applied on the complete spectrogram, as we saw above with the OC/C results.

Graphical results are shown in Fig. 3 (d), (e), (f), where one can see in particular that the acoustical scene analysis (i.e., the learning of a spectro-



(a) Spectrogram of the original waveform.

(b) Spectrogram of the truncated waveform.

(c) Truncated spectrogram with areas to inpaint in black.

(d) Spectrogram reconstructed using the $\mathcal{I}$-divergence.

(e) Reconstructed and separated spectrogram of the piano part.

(f) Reconstructed and separated spectrogram of the trumpet part.

Figure 3: *NMF2D with missing data on the spectrogram of a truncated waveform. (a) Spectrogram of the original waveform (a mixture of piano and trumpet sounds). (b) Spectrogram of the truncated waveform. (c) Truncated spectrogram, with truncated regions indicated in black. (d) Estimated factors and reconstructed spectrogram using the $\mathcal{I}$-divergence algorithm. (e) Reconstructed and separated spectrogram of the piano part. (f) Reconstructed and separated spectrogram of the trumpet part.*

temporal template for each instrument and the estimation of the pitch and onset time of each note) is performed correctly, and that blind source separation is also performed in spite of the presence of gaps.

Here again, it is interesting to note that the missing information is not reconstructed from the neighboring parts as in classical interpolation techniques, but indirectly from similar patterns in other regions of the spectrograms, using the local information mainly to determine what similar patterns to use in the reconstruction. Although one may think that in this particular simple exam-

ple, where pitches are flat and spectral envelopes rather steady, interpolation techniques could be used as well, it is important to note that such techniques are helpless when higher-order structure is necessary to reconstruct the missing regions: this is for example the case when these regions include the beginning or the end of a note and typical note length is thus a crucial key, when a complete harmonic is missing and needs to be reconstructed by inferring it from the typical spectro-temporal envelope and the strength of the current note, or in general when the missing information is too complex to be inferred from the neighboring parts only. One can relate this behavior of our model to the phonemic illusion phenomenon mentioned earlier, where humans arguably also use higher-order models (e.g., language models, production models) to infer the missing phonemes.

6. Missing-data Harmonic-Temporal Clustering

The Harmonic-Temporal Clustering framework attempts to perform the analysis of an acoustical scene by modeling its power spectrogram as a constrained Gaussian mixture model. It has been introduced by Kameoka et al. (2007) for music signals and later extended to signals with continuously varying pitch such as speech by Le Roux et al. (2007a). We explain here how it can be further extended to deal with acoustical scenes with incomplete data, and how the inherent continuity constraint on the fundamental frequency contour imposed by the use of cubic spline functions and an extra Markovian prior enables us to perform robust F_0 estimation on incomplete speech data.

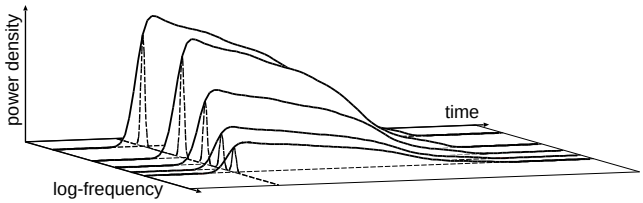
6.1. Overview of the HTC model

Consider the wavelet power spectrum $W(x,t)$ of a signal recorded from an acoustical scene, defined on a domain of definition $D = \{x,t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$. The problem considered is to approximate the power spectrum as well as possible as the sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$ modeling the power spectrum of K "objects" each with its own F_0 contour $\mu_k(t)$ and its own harmonic-temporal structure. As described in Kameoka et al. (2007) and Le Roux et al. (2007a), the source models $q_k(x,t;\mathbf{\Theta})$ are expressed as a Gaussian Mixture Model (GMM) with constraints on the characteristics of the kernel distributions: supposing that there is harmonicity with N partials modeled in the frequency direction, and that the power envelope is described using Y kernel functions in the time direction, we can rewrite each source model in the form

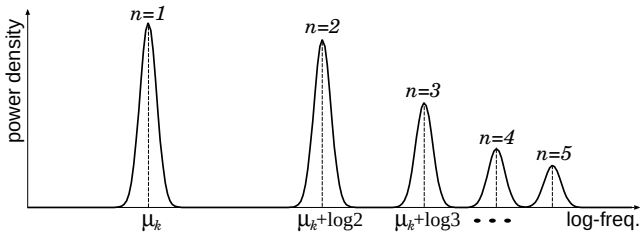
$$q_k(x,t;\mathbf{\Theta}) = \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\mathbf{\Theta}), \tag{23}$$

where $\mathbf{\Theta}$ is the set of all parameters and with kernel densities $S_{kny}(x,t;\mathbf{\Theta})$ which are assumed to have the following shape:

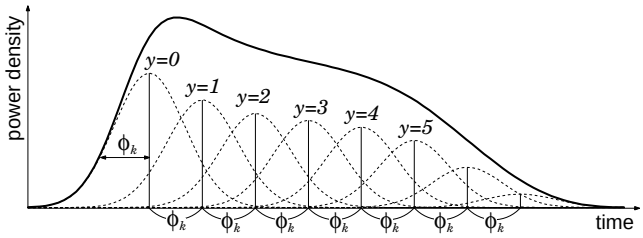
$$S_{kny}(x,t;\mathbf{\Theta}) \triangleq \frac{w_k v_{kn} u_{kny}}{2\pi\sigma_k\phi_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}}, \tag{24}$$



(a) Profile of an HTC source model $q_k(x, t; \mathbf{\Theta})$



(b) Cross-section of $q_k(x, t; \mathbf{\Theta})$ at constant time



(c) Power envelope function

Figure 4: *Graphical representation of an HTC source model. Fig. (a) shows the time-frequency profile of the model, while Fig. (b) shows a cross-section of the model at constant time and Fig. (c) the evolution in time of the power envelope function. The harmonic structure of the model can be seen in Fig. (b), and the approximation of the power envelope in the time direction as a sum of Gaussian kernels can be seen in Fig. (c).*

where the weight parameters w_k, v_{kn} and u_{kny} are normalized such that $\sum_k w_k = 1$, $\sum_n v_{kn} = 1, \forall k$ and $\sum_y u_{kny} = 1, \forall k, n$. The parameter τ_k gives the onset time of the source model q_k, w_k its energy, v_{kn} the ratio of energy inside its n-th partial, and the parameters u_{kny} together with the duration paramater ϕ_k determine the shape of the temporal envelope of this n-th partial. A graphical representation of an HTC source model $q_k(x, t; \mathbf{\Theta})$ can be seen in Fig. 4. The F_0 contours $\mu_k(t)$ can be expressed using piece-wise flat functions or cubic spline functions according to the signal to be modeled.

The goal is to minimize the difference between $W(x, t)$ and $Q(x, t; \mathbf{\Theta}) = \sum_{k=1}^{K} q_k(x, t; \mathbf{\Theta})$ according to a certain criterion. We use the $\mathcal{I}$-divergence (Csiszár, 1975) as a classical way to measure the distance between two non-

negative distributions:

$$\mathcal{I}(W|Q(\mathbf{\Theta})) \triangleq \iint_D \left(W(x,t) \log \frac{W(x,t)}{Q(x,t;\mathbf{\Theta})} - \Big(W(x,t) - Q(x,t;\mathbf{\Theta}) \Big) \right) dx\, dt, \tag{25}$$

and we are thus looking for $\mathbf{\Theta}_{\text{opt}} = \text{argmin}_{\mathbf{\Theta}}\, \mathcal{I}(W|Q(\mathbf{\Theta}))$.

6.2. Formulation of the model on incomplete data

The optimization process in HTC (Kameoka et al., 2007; Le Roux et al., 2007a) is nothing else than the fitting of a model (in particular a constrained Gaussian mixture model) to an observed distribution (the wavelet power spectrum of an acoustical scene), using the $\mathcal{I}$-divergence as a measure of the goodness of fit.

If some parts of the power spectrum are missing or corrupted, or if some parts of the HTC model are partially or entirely lying outside the boundaries of the spectrogram (for example if some harmonics of the model are above the maximum frequency and a prior is used to link the powers of the harmonics, fitting the upper harmonics to zero will bias the optimization), the estimation of the HTC model must be performed under an incomplete-data framework, as in Section 2.3. In the same way as we showed there, optimization can be performed in an iterative way by using the values of the model at the previous step as an estimation of the unobserved data. In the case of HTC, this results in a hierarchical algorithm with two levels. At the upper level is the iterative algorithm described above. At the lower level, inside the step 2 of the upper level, the EM algorithm is used as in the classical formulation of the HTC optimization. Let W be the observed part of the spectrogram and $I \subset D$ its domain of definition. The objective function to minimize here is the same as (25) but restricted to the domain where the spectrogram is observed:

$$\mathcal{I}(W, Q(\mathbf{\Theta})) \triangleq \iint_I \left(W(x,t) \log \frac{W(x,t)}{Q(x,t;\mathbf{\Theta})} - \Big(W(x,t) - Q(x,t;\mathbf{\Theta}) \Big) \right) dx\, dt. \tag{26}$$

We define the auxiliary function as

$$\begin{aligned} &\mathcal{I}^{+}(W, V, Q(\mathbf{\Theta})) \triangleq \\ &\mathcal{I}(W, Q(\mathbf{\Theta})) + \iint_{D \setminus I} \left(V(x,t) \log \frac{V(x,t)}{Q(x,t;\mathbf{\Theta})} - \Big(V(x,t) - Q(x,t;\mathbf{\Theta}) \Big) \right) dx\, dt. \end{aligned} \tag{27}$$

Then membership functions m can be further introduced as in the classical formulation of HTC to build the final auxiliary function $\mathcal{I}^{++}(W, V, Q(\mathbf{\Theta}), m)$. These membership functions are non-negative and sum up to 1 for each (x,t): $\sum_{k,n,y} m_{kny}(x,t) = 1$. If we note

$$Z(x,t) = \begin{cases} W(x,t) & \text{if } (x,t) \in I \\ V(x,t) & \text{if } (x,t) \in D \setminus I \end{cases}$$

we define

$$\mathcal{I}^{++}(W, V, Q(\boldsymbol{\Theta}), m) \triangleq \iint_D \left(\sum_{k,n,y} m_{kny}(x,t) Z(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t) Z(x,t)} - \Big(Z(x,t) - Q(x,t;\boldsymbol{\Theta}) \Big) \right) dx\, dt. \quad (28)$$

Using the concavity of the logarithm, one can see that

$$\mathcal{I}^{+}(W, V, Q(\boldsymbol{\Theta})) \leq \mathcal{I}^{++}(W, V, Q(\boldsymbol{\Theta}), m) \quad (29)$$

with equality for

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}. \quad (30)$$

Altogether, the optimization process can be formulated as follows.

Step 1 Estimate V such that $\mathcal{I}(W, Q(\boldsymbol{\Theta})) = \mathcal{I}^{+}(W, V, \boldsymbol{\Theta})$:

$$\hat{V}(x,t) = Q(x,t;\boldsymbol{\Theta}), \forall (x,t) \in D \setminus I. \quad (31)$$

Step 2 Update $\boldsymbol{\Theta}$ with $\hat{V}$ fixed:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, \mathcal{I}^{+}(W, \hat{V}, \boldsymbol{\Theta}). \quad (32)$$

To do so, perform one iteration of the classical formulation of HTC:

E-Step

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}, \quad (33)$$

M-Step

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \Big(\mathcal{I}^{++}(W, \hat{V}, \boldsymbol{\Theta}, \hat{m}) - \log P(\boldsymbol{\Theta}) \Big) \quad (34)$$

where $P(\boldsymbol{\Theta})$ is a prior distribution on the parameters.

6.3. Optimization of the model

Analytical update equations for the M-step are derived in Kameoka et al. (2007) and Le Roux et al. (2007a). However, when the F_0 contour is modeled using cubic spline functions, which is relevant for speech or musical instruments whose pitch can vary continuously, the spline parameters were updated in Le Roux et al. (2007a) not globally but one after the other. The corresponding optimization procedure, called the Expectation-Constrained Maximization algorithm (ECM) (Meng and Rubin, 1993), does not ensure the minimization in the M-step but nonetheless guarantees the decrease of the objective function. This spline parameter update was thus not optimal but yet led to very good results in F_0 estimation accuracy. However, it suffered from some instability problems in long regions with low harmonic energy (silence or unvoiced parts of speech for example). When dealing with missing-data problems, such issues become critical, and we thus need to use better update equations for the spline parameters, briefly introduced in Le Roux et al. (2007b) and which we present in detail in Appendix Appendix B. Contrary to the update equations previously described in Le Roux et al. (2007a), the ones presented here are global analytical update equations which lead to the minimum of the auxiliary function in the M-step. They ensure a greater stability of the spline model and a better F_0 estimation accuracy, as shown in the next section.

7. F_0 estimation on incomplete data with HTC

7.1. Importance of F_0 estimation accuracy

Ensuring a very good accuracy for the F_0 estimation is not only important as a necessary step for computational auditory induction by HTC, but it is also in itself a primary issue. Indeed, being able to estimate the F_0 accurately is important as well for some previous audio interpolation methods such as Maher's sinusoidal model based method (Maher, 1994), in which the harmonics before and after the gap need to be linked, or Vaseghi and Rayner's extended AR model (Vaseghi and Rayner, 1990), which takes advantage of the long-term correlation structure of the signals by introducing extra predictor parameters around the pitch period.

7.2. Relevance of HTC's F_0 contour model

When applied to speech, HTC is based on a spline F_0 contour. A Markovian prior, presented in Appendix Appendix B, is used on the parameters of the contour to ensure that it will not move too abruptly. This Markovian prior penalizes the deviation of a spline parameter from the linear interpolation of its neighbors. Altogether, HTC's F_0 contour model is somewhere between a spline interpolation and a linear interpolation, depending on the strength of the matching between the HTC source model and the observed data.

Attempting to use this model for reconstruction of incomplete data implies that the F_0 contour inside the gap is close to an interpolation based on the values of the contour outside the gap. To confirm the relevance of this interpolation,

we thus need to ensure that, assuming that the F_0 estimation on complete parts of the data is accurately performed, the F_0 of the missing parts of the data is accurately performed as well.

We thus evaluated as a preliminary experiment the accuracy of the F_0 contour obtained by interpolating the reference F_0 values outside the gaps on the whole interval using both natural splines and linear interpolation. This can be considered as an evaluation of what can be expected by HTC.

We then conducted experiments to confirm the accuracy of the proposed method for F_0 estimation. The goal here was first to confirm that the new spline update equations indeed outperform the former update equations on single-speaker F_0 estimation in clean environment for complete data, then to evaluate the F_0 accuracy with parts of the data missing.

7.3. Experimental setting

The general conditions of the experiments were exactly the same as in Le Roux et al. (2007a), and we shall briefly review them here, all the details being given there.

We used a database of speech recorded together with a laryngograph signal (Bagshaw et al., 1993), consisting of one male and one female speaker who each spoke 50 English sentences for a total of 5 min and 37 s of speech, for the purpose of evaluation of F_0-estimation algorithms. The power spectrum $W(x, t)$ was calculated from an input signal digitized at a 16 kHz sampling rate (the original data of the database was converted from 20 kHz to 16 kHz) using a Gabor wavelet transform with a time resolution of 16 ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The lower bound of the frequency range and the frequency resolution were respectively 50 Hz and 14 cent. The spline contour was initially flat and set to 132 Hz for the male speaker and 296 Hz for the female speaker. The length of the interpolation intervals was fixed to 4 frames. For HTC, we used $K = 10$ source models, each of them with $N = 10$ harmonics. We used as ground truth the F_0 estimates and the reliability mask derived by de Cheveigné and Kawahara (2002). As the spline function gives an analytical expression for the F_0 contour, we compare our result with the reference values at a sampling rate of 20 kHz although all the analysis was performed with a time resolution of 16 ms. Deviations over 20 % from the reference were deemed to be gross errors.

For the incomplete data, we prepared four sets of data by replacing segments of the utterances of different lengths by silence. The sets are prepared such that approximately 20 % of the data is lost, erasing segments of length L ms every $5L$ ms of data (the obtained utterances would then be successions of L ms of silence, $4L$ ms of speech, L ms of silence, etc.). The four lengths we selected are 25 ms, 50 ms, 75 ms and 100 ms, and the corresponding data sets are denoted by Erase-25ms, Erase-50ms, Erase-75ms and Erase-100ms respectively. For example, the utterances in Erase-50ms were produced by erasing 50 ms every 250 ms of data, leading to utterances which are successions of 50 ms of silence, 200 ms of speech, 50 ms of silence, etc. We shall note that gaps from

Table 2: *Gross error rates (%) for F_0 interpolation inside the gaps based on reference F_0*

Data set		Cubic Splines	Linear interpolation
Erase-25ms	Male	0.9	0.0
	Female	0.2	0.0
	Total	**0.5**	**0.0**
Erase-50ms	Male	1.5	0.5
	Female	1.0	0.5
	Total	**1.2**	**0.5**
Erase-75ms	Male	4.0	1.3
	Female	3.0	0.5
	Total	**3.5**	**0.9**
Erase-100ms	Male	7.6	4.2
	Female	6.2	1.6
	Total	**6.9**	**2.9**

30 ms to 50 ms are already considered very long gaps in the audio restoration literature (Esquef and Biscainho, 2006; Godsill and Rayner, 1998; Maher, 1994).

7.4. Preliminary experiment on F_0 interpolation

We first performed a preliminary experiment based on the reference F_0 and the reliability mask derived in de Cheveigné and Kawahara (2002). The reliability mask was used to determine the voiced regions of the speech utterance, and a global contour over the whole utterance was derived by interpolating the values of the F_0 reference which were both inside the reliability mask and outside the erased segments of the data. We used both linear interpolation and cubic spline interpolation. We then computed the gross error rates of the interpolated F_0 values inside the gaps (by construction the values outside the gaps are equal to the reference and thus no error can occur there). Results for the incomplete data sets can be seen in Table 2. Spline interpolation does not perform as well as linear interpolation due to the large variations that can occur depending on the slope of the contour at the beginning or end of a voiced region. This is precisely what the Markovian prior in HTC's F_0 contour model aims to avoid.

7.5. Accuracy on complete data

We first used the classical HTC formulation on complete data, using the new spline update equations. Here, only step 2 of the algorithm devised in 6.2 is thus used (iteration of equations (33) and (34)).

The results can be seen in Table 3, with for comparison the results obtained with HTC using the former spline update equations as well as the ones obtained with the state-of-the-art algorithm YIN (de Cheveigné and Kawahara, 2002). We note that we obtained 2.1 % gross error rate for YIN using the code made available by its authors, as opposed to 1.3 % reported in the original paper. We can see that HTC with the newly proposed spline update equations now performs comparably to YIN.

Table 3: *Gross error rates for F_0 estimation on complete data (clean single-speaker speech)*

Method		**Gross error (%)**
YIN	Male	3.2
	Female	1.0
	Total	**2.1**
HTC (former spline update)	Male	3.2
	Female	3.7
	Total	**3.5**
HTC (proposed spline update)	Male	1.1
	Female	1.3
	Total	**1.2**

7.6. Accuracy on incomplete data

The wavelet transforms were performed on the truncated waveforms of the data sets introduced above. The regions $D \setminus I$ which are to be considered missing in the spectrogram were defined as the frames corresponding to the erased parts of the waveform. The influence of the erased portion is larger for low frequencies, but we neglect this and consider missing a whole frame regardless of the frequency bin.

In such situations where part of the data is irrelevant, one might think that algorithms which perform F_0 estimation more locally should be used, using interpolation between the preceding and following voiced portions to obtain F_0 values inside the gap. If the estimation can be accurately performed outside the gaps, such a method should lead to very good results, as we saw in 7.4. However, one needs to note that if such algorithms are used, a robust Voice Activity Detection (VAD) must be performed as well to determine which points should be used in the interpolation. A poor VAD accuracy could lead to very bad results in the interpolation process, as unreliable values for the F_0 could be used as base points for the interpolation, leading to wrong results on the whole interpolation region. To illustrate this and as a comparison with HTC, we used the algorithm YIN to perform F_0 estimation outside the gaps, and used a linear interpolation to obtain values inside the gaps, using the closest voiced regions outside the gaps as boundaries. The positions of the gaps were given, and the voiced regions were determined using the aperiodicity measure given by YIN, with a threshold of 0.2. The results given here were obtained using linear interpolation, but cubic spline interpolation gave similar results.

Results for HTC and YIN are given in Table 4, with gross error rates for the whole file as well as for the erased segments only. We can see that the performance of HTC degrades as the gaps become longer. HTC performs better than the algorithm based on YIN for the total accuracy as well as for the accuracy inside the gaps with 25 ms and 50 ms erased segments, while the algorithm based on YIN performs better inside the gaps with 75 ms and 100 ms erased segments but is still outperformed on the total error.

Table 4: *Gross error rates for F_0 estimation on incomplete data with HTC and YIN (clean single-speaker speech). The results for YIN are indicated in parentheses.*

Data set		**Error in the gaps (%)**	**Total error (%)**
Erase-25ms	Male	6.0 (12.0)	3.9 (10.0)
	Female	4.6 (4.7)	3.1 (3.3)
	Total	**5.3 (8.3)**	**3.5 (6.5)**
Erase-50ms	Male	8.8 (14.2)	4.2 (9.5)
	Female	6.9 (5.7)	2.7 (3.2)
	Total	**7.9 (9.9)**	**3.4 (6.3)**
Erase-75ms	Male	14.1 (15.5)	4.5 (9.5)
	Female	13.4 (6.4)	4.1 (3.3)
	Total	**13.7 (10.9)**	**4.3 (6.3)**
Erase-100ms	Male	22.1 (19.2)	7.0 (10.3)
	Female	19.5 (7.4)	5.6 (3.6)
	Total	**20.9 (13.5)**	**6.3 (6.9)**

These results raise several remarks. We note first that the accuracy of YIN on the whole waveform is stable as gaps become longer while decreasing inside the gaps, meaning that it tends to increase outside the gaps. Although this may first sound surprising, it is related to the fact that interpolation is performed by using as anchors the closest regions outside the gaps with a sufficiently low aperiodicity measure; the presence of gaps will thus influence the interpolation process and the F_0 estimates around them. But as gaps become longer, the results inside and outside become less dependent: each gap inducing a loss of information which influences a small neighborhood through interpolation, many small gaps are likely to harm the estimates outside the gaps more than a few long ones. On the other hand, HTC is influenced in a different way by the length of the gaps. Results outside the gaps are stable, while accuracy inside the gaps decreases faster than for YIN when gaps become too long. One reaches here the limits of the above-mentioned "prediction power" of the model inside the gaps, and its degree of freedom becomes too high to systematically converge to a relevant solution. This problem could be coped with by investigating the introduction of more complex pitch contour models which try to encompass the long-term dynamics of the F_0, such as the Fujisaki model (Fujisaki and Nagashima, 1969) for example, or more complex priors on the spline contours, although in both cases one faces the risk of making the optimization intractable.

Altogether, the results show that HTC's F_0 estimation accuracy, while degrading in extreme cases, is very good even in the presence of long gaps, and that, although other F_0 estimation algorithms could be used as well, it is not obvious, regardless of their performance on complete data, whether they can be turned into effective algorithms on incomplete data, due in particular to the importance of a robust VAD for the interpolation to be effective.

8. Conclusion

We presented a computational framework to model auditory induction, i.e., the human auditory system's ability to estimate the missing parts of a continuous auditory stream briefly covered by noise, by extending acoustical scene analysis methods based on global statistical models such as HTC and NMF2D to handle unobserved data. We related the method to the EM algorithm, enabling the use of priors on the parameters. We illustrated on a simple example how the proposed framework was able to simultaneously perform acoustical scene analysis and gap interpolation in a musical piece with NMF2D, and how a robust F_0 estimation could be performed on incomplete data with HTC. While we assumed here that the gap locations were known, future work will investigate their joint estimation together with the model parameters and the missing data, in a similar way to Barker et al. (2005) for missing-feature speech recognition.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments which greatly helped improve the clarity and the general quality of this article.

Appendix A. Derivation of the Q-function of Section 3.2

We compute here the explicit form of the Q-function $Q(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})$ involved in the application of the EM algorithm to the ML problem introduced in Section 3.2 in which the data are supposed to have been generated independently at each x from a probability distribution of an exponential family $\mathcal{F}_\phi$ with expectation parameter $g(x, \boldsymbol{\Theta})$. In the following, we will denote by $\nu_{x,\phi,\boldsymbol{\Theta}}(z)$ the density of this probability distribution, which can be written as explained in 3.1 directly using the corresponding Bregman divergence:

$$\nu_{x,\phi,\boldsymbol{\Theta}}(z) = e^{-d_\phi(z, g(x;\boldsymbol{\Theta}))} b_\phi(z). \tag{A.1}$$

As the data are supposed to have been generated independently at each x from the probability distribution with density $\nu_{x,\phi,\boldsymbol{\Theta}}(z)$, observed and unobserved data are in particular independent conditionally to $\boldsymbol{\Theta}$, and the Q-function can be written as follows:

$$\begin{aligned}
Q(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) &= \mathbb{E}(\log P(h|\boldsymbol{\Theta}))_{P(h|f,\bar{\boldsymbol{\Theta}})} + \mathbb{E}(\log P(f|\boldsymbol{\Theta}))_{P(h|f,\bar{\boldsymbol{\Theta}})} \\
&= \int_{\mathbb{R}^n \setminus I} \mathbb{E}(\log P(h(x)|\boldsymbol{\Theta}))_{P(h(x)|\bar{\boldsymbol{\Theta}})}\, dx \\
&\qquad + \left(\int P(h|f, \bar{\boldsymbol{\Theta}}) \right) \int_I \log P(f(x)|\boldsymbol{\Theta})\, dx \\
&= \int_{\mathbb{R}^n \setminus I} \int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z) \log \nu_{x,\phi,\boldsymbol{\Theta}}(z)\, dz\, dx + \int_I \log P(f(x)|\boldsymbol{\Theta})\, dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^n \setminus I} \int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z) \Big(\log b_\phi(z) - d_\phi(z, g(x;\boldsymbol{\Theta})) \Big)\, dz\, dx \\
&\qquad + \int_I \Big(\log b_\phi(f(x)) - d_\phi(f(x), g(x;\boldsymbol{\Theta})) \Big)\, dx \\
&= -\int_{\mathbb{R}^n \setminus I} \int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z) d_\phi(z, g(x;\boldsymbol{\Theta}))\, dz\, dx \\
&\qquad - \int_I d_\phi(f(x), g(x;\boldsymbol{\Theta}))\, dx + C_1(f, \bar{\boldsymbol{\Theta}}), \qquad \text{(A.2)}
\end{aligned}
$$

where $C_1(f, \bar{\boldsymbol{\Theta}})$ does not depend on $\boldsymbol{\Theta}$. If we now rewrite $d_\phi(z, g(x;\boldsymbol{\Theta}))$ as

$$
\begin{aligned}
d_\phi(z, g(x;\boldsymbol{\Theta})) &= d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) \\
&\quad + \phi(z) - \phi(g(x;\bar{\boldsymbol{\Theta}})) - \langle z - g(x;\bar{\boldsymbol{\Theta}}); \nabla\phi(g(x;\boldsymbol{\Theta})) \rangle, \qquad \text{(A.3)}
\end{aligned}
$$

we can simplify the first term in Eq. (A.2):

$$
\begin{aligned}
&\int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z) d_\phi(z, g(x;\boldsymbol{\Theta}))\, dz \\
&\qquad = \Big(\int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz \Big) d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) \\
&\qquad\quad - \Big\langle \int (z - g(x;\bar{\boldsymbol{\Theta}})) \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz; \nabla\phi(g(x;\boldsymbol{\Theta})) \Big\rangle + C_2(f, \bar{\boldsymbol{\Theta}}) \\
&\qquad = d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) + C_2(f, \bar{\boldsymbol{\Theta}}),
\end{aligned}
$$

where $C_2(f, \bar{\boldsymbol{\Theta}})$ does not depend on $\boldsymbol{\Theta}$. To lead the calculation above, we used the fact that the mass of a probability distribution of an exponential family with expectation parameter $g(x;\bar{\boldsymbol{\Theta}})$ is 1 and its mean is $g(x;\bar{\boldsymbol{\Theta}})$:

$$
\int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz = 1, \qquad \text{(A.4)}
$$

$$
\int z \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz = g(x;\bar{\boldsymbol{\Theta}}). \qquad \text{(A.5)}
$$

We then obtain for the Q-function

$$
\begin{aligned}
Q(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) &= -\int_{\mathbb{R}^n \setminus I} d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta}))\, dx \\
&\qquad - \int_I d_\phi(f(x), g(x;\boldsymbol{\Theta}))\, dx + C(f, \bar{\boldsymbol{\Theta}}) \\
&= -\mathcal{L}^+(\boldsymbol{\Theta}, g(x;\bar{\boldsymbol{\Theta}})) + C(f, \bar{\boldsymbol{\Theta}}), \qquad \text{(A.6)}
\end{aligned}
$$

where $C(f, \bar{\boldsymbol{\Theta}})$ again does not depend on $\boldsymbol{\Theta}$.

Appendix B. Derivation of the spline parameter update equations in HTC

Appendix B.1. Spline contour

The analysis interval is divided into subintervals $[t_i, t_{i+1})$ of equal length ϵ. The parameters of the spline contour model are then the values z_i of the F_0 at

each bounding point t_i. Assuming that the second derivative vanishes at the bounds of the analysis interval leads to the so-called natural splines. Under this assumption, one can explicitly compute offline a matrix M linking the values z_i'' of the second derivative of the contour at t_i with the values z_i, such that $\boldsymbol{z}'' = \boldsymbol{M}\boldsymbol{z}$. An analytical expression for the contour $\mu(t;\boldsymbol{z})$ as a concatenation of third order polynomials can then be classically obtained. For $t \in [t_i, t_{i+1})$:

$$\mu(t;\boldsymbol{z}) \triangleq \frac{1}{t_{i+1}-t_i}\Big(z_i(t_{i+1}-t) + z_{i+1}(t-t_i) \\ - \frac{1}{6}(t-t_i)(t_{i+1}-t)\big[(t_{i+2}-t)z_i'' + (t-t_{i-1})z_{i+1}''\big]\Big). \quad \text{(B.1)}$$

One can notice that the expression of $\mu(t;\boldsymbol{z})$ is actually linear in $\boldsymbol{z}$:

$$\mu(t;\boldsymbol{z}) = \boldsymbol{A}(t)^\top \boldsymbol{z} \quad \text{(B.2)}$$

where $\boldsymbol{A}(t)$ is a column vector such that, for $t \in [t_i, t_{i+1})$,

$$\boldsymbol{A}(t) = \frac{1}{t_{i+1}-t_i}\Big((t_{i+1}-t)\boldsymbol{e}_i + (t-t_i)\boldsymbol{e}_{i+1} \\ - \frac{(t-t_i)(t_{i+1}-t)}{6}\big[(t_{i+2}-t)\boldsymbol{M}_i^\top + (t-t_{i-1})\boldsymbol{M}_{i+1}^\top\big]\Big) \quad \text{(B.3)}$$

where $\boldsymbol{M}_j$ denotes the j-th row of the matrix $\boldsymbol{M}$ and $\boldsymbol{e}_j$ denotes the j-th vector of the canonical basis. We note furthermore that $\boldsymbol{A}(t) = \nabla_{\boldsymbol{z}}\,\mu(t;\boldsymbol{z})$.

Appendix B.2. Optimization of the objective function

During the M-step of the EM algorithm, one wants to minimize $\mathcal{J}(\boldsymbol{\Theta}) = \mathcal{I}^{++}(W, \hat{V}, \boldsymbol{\Theta}, \hat{m}) - \log P(\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$. We can compute the gradient with respect to $\boldsymbol{z}$:

$$\begin{aligned}\nabla_{\boldsymbol{z}}\mathcal{J} &= -\iint_D \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \mu(t,\boldsymbol{z}) - \log n)\boldsymbol{A}(t)\,dx\,dt - \nabla_{\boldsymbol{z}}\log P(\boldsymbol{\Theta}) \\ &= -\iint_D \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \boldsymbol{A}(t)^\top\boldsymbol{z} - \log n)\boldsymbol{A}(t)\,dx\,dt - \nabla_{\boldsymbol{z}}\log P(\boldsymbol{\Theta})\end{aligned}$$

where $\ell_{kny}(x,t) = m_{kny}(x,t)Z(x,t)$. Note that the term $\iint_D Q(x,t;\boldsymbol{\Theta})\,dx\,dt$ in (28) does not contribute to the gradient w.r.t. $\boldsymbol{z}$ as the spline parameters do not influence the normalization of the model.
Let

$$\begin{aligned}\phi(t) &= \int \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \log n)\,dx, \\ \gamma(t) &= \int \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}\,dx.\end{aligned}$$

Then

$$\nabla_{\boldsymbol{z}}\mathcal{J} \;=\; -\int \phi(t)\boldsymbol{A}(t)\,dt + \left(\int \gamma(t)\boldsymbol{A}(t)\boldsymbol{A}(t)^\top\,dt\right)\boldsymbol{z} - \nabla_{\boldsymbol{z}}\log P(\boldsymbol{\Theta}).$$

One can then obtain the Hessian matrix:

$$H_{\boldsymbol{z}}\mathcal{J} = \int \gamma(t)\boldsymbol{A}(t)\boldsymbol{A}(t)^\top\,dt - H_{\boldsymbol{z}}\log P(\boldsymbol{\Theta}). \tag{B.4}$$

If one uses a Markov assumption on the spline parameters with Gaussian distributions for the state transitions, the prior distribution becomes

$$P(\boldsymbol{z}) = P(z_0)\prod_{j=1}^{|\boldsymbol{z}|} P(z_j|z_{j-1}),$$

with z_0 following a uniform distribution and

$$P(z_j|z_{j-1}) = \frac{1}{\sqrt{2\pi}\sigma_s}e^{-\frac{(z_j-z_{j-1})^2}{2\sigma_s^2}}.$$

Then

$$\begin{aligned}\nabla_{\boldsymbol{z}}\log P(\boldsymbol{\Theta}) &= -\frac{1}{\sigma_s^2}\begin{pmatrix} 1 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 1 \end{pmatrix}\boldsymbol{z} \\ &= H_{\boldsymbol{z}}\log P(\boldsymbol{\Theta})\boldsymbol{z}. \end{aligned}\tag{B.5}$$

Putting to 0 the gradient w.r.t. $\boldsymbol{z}$, one can find the update equation for $\boldsymbol{z}$:

$$\boldsymbol{z} = (H_{\boldsymbol{z}}\mathcal{J})^{-1}\int \phi(t)\boldsymbol{A}(t)\,dt. \tag{B.6}$$

The convexity can be studied by looking at $H_{\boldsymbol{z}}\mathcal{J}$in Eq. (B.4). The first term is indeed non-negative, as $\gamma(t) \geq 0, \forall t$. For the second term, coming from the prior distribution, we recognize a tridiagonal matrix, for which the principal minors can be easily calculated. If $T = (t_{ij})$ is a tridiagonal matrix and α_n its n-th principal minor, then

$$\alpha_n = t_{n,n}\alpha_{n-1} + t_{n,n-1}t_{n-1,n}\alpha_{n-2}. \tag{B.7}$$

In our case, we see that the principal minors of $H_{\boldsymbol{z}}\log P(\boldsymbol{\Theta})$ are all non-positive. The matrix $-H_{\boldsymbol{z}}\log P(\boldsymbol{\Theta})$ is thus positive semi-definite. Altogether, $H_{\boldsymbol{z}}\mathcal{J}$ is at least positive semi-definite, and the update (B.6) thus corresponds to a minimum.

References

Achan, K., Roweis, S., Hertzmann, A., Frey, B., Mar. 2005. A segment-based probabilistic generative model of speech. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 5. pp. 221–224.

Bagshaw, P. C., Hiller, S. M., Jack, M. A., Sep. 1993. Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech). pp. 1003–1006.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., Oct. 2005. Clustering with Bregman divergences. Journal of Machine Learning Research 6, 1705–1749.

Barker, J., Cooke, M. P., Ellis, D. P. W., 2005. Decoding speech in the presence of other sources. Speech Communication 45 (1), 5–25.

Barker, J. P., 2006. Robust automatic speech recognition. In: Wang, D.-L., Brown, G. J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley, pp. 297–350.

Barlow, H. B., Aug. 2001. The exploitation of regularities in the environment by the brain. Behavioral and Brain Sciences 24 (4), 602–607.

Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C., Jul. 2000. Image inpainting. In: SIGGRAPH 2000, Computer Graphics Proc. pp. 417–424.

Bregman, A. S., 1990. Auditory Scene Analysis. The MIT Press.

Bregman, L. M., 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. Computational Mathematics and Mathematical Physics 7 (3), 620–631.

Cemgil, A. T., Jul. 2008. Bayesian inference in non-negative matrix factorization models. Tech. Rep. CUED/F-INFENG/TR.609, University of Cambridge.

Cemgil, A. T., Godsill, S. J., Sep. 2005. Probabilistic phase vocoder and its application to interpolation of missing values in audio signals. In: Proceedings of the European Signal Processing Conference (EUSIPCO).

Cherry, E. C., 1953. Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustical Society of America 25 (5), 975–979.

de Cheveigné, A., Kawahara, H., Apr. 2002. YIN, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111 (4), 1917–1930.

Clark, P., Atlas, L., Apr. 2008. Modulation decompositions for the interpolation of long gaps in acoustic signals. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 3741–3744.

Cooke, M. P., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.

Criminisi, A., Pérez, P., Toyama, K., Sep. 2004. Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing 13 (9), 1200–1212.

Csiszár, I., Jan. 1975. I-divergence geometry of probability distributions and minimization problems. The Annals of Probability 3 (1), 146–158.

Eggert, J., Körner, E., Jul. 2004. Sparse coding and NMF. In: Proceedings of the IEEE International Joint Conference on Neural Networks. Vol. 4. pp. 2529–2533.

Ellis, D. P. W., Oct. 1993. Hierarchic models of sound for separation and restoration. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

Ellis, D. P. W., Jun. 1996. Prediction-driven computational auditory scene analysis. Ph.D. thesis, Massachusetts Institute of Technology.

Esquef, P. A. A., Biscainho, L. W. P., Jul. 2006. An efficient model-based multirate method for reconstruction of audio signals across long gaps. IEEE Transactions on Audio, Speech, and Language Processing 14 (4), 1391–1400.

Févotte, C., Bertin, N., Durrieu, J.-L., Mar. 2009. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation 21 (3), 793–830.

Fujisaki, H., Nagashima, S., 1969. A model for synthesis of pitch contours of connected speech. Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo 28, 53–60.

Godsill, S. J., Rayner, P. J. W., 1998. Digital Audio Restoration: A Statistical Model Based Approach. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Griffin, D. W., Lim, J. S., Apr. 1984. Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (2), 236–243.

Grünwald, P. D., 2007. The Minimum Description Length Principle. The MIT Press.

von Helmholtz, H., 1954. On the Sensations of Tone as a Physiological Basis for the Theory of Music. (Second English edition, translated by A. J. Ellis, 1885. First German edition: 1863). Reprinted by Dover Publications.

Janssen, A. J. E. M., Veldhuis, R., Vries, L. B., Apr. 1986. Adaptive interpolation of discrete-time signals that can be modeled as AR processes. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (2), 317–330.

Kameoka, H., Mar. 2007. Statistical approach to multipitch analysis. Ph.D. thesis, The University of Tokyo.

Kameoka, H., Nishimoto, T., Sagayama, S., Mar. 2007. A multipitch analyzer based on harmonic temporal structured clustering. IEEE Transactions on Audio, Speech, and Language Processing 15 (3), 982–994.

Kameoka, H., Ono, N., Kashino, K., Sagayama, S., Apr. 2009. Complex NMF: A new sparse representation for acoustic signals. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 3437–3440.

Kashino, M., Jun. 2006. Phonemic restoration: The brain creates missing speech sounds. Acoustical Science and Technology 27 (6), 318–321.

Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., Sagayama, S., May 2007a. Single and multiple F_0 contour estimation through parametric spectrogram modeling of speech in noisy environments. IEEE Transactions on Audio, Speech, and Language Processing 15 (4), 1135–1145.

Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., Sagayama, S., Sep. 2007b. Monaural speech separation through harmonic-temporal clustering of the power spectrum. In: Proceedings of the Acoustical Society of Japan Autumn Meeting. No. 3-4-3. pp. 351–352.

Le Roux, J., Kameoka, H., Ono, N., Sagayama, S., Mar. 2008a. On the interpretation of I-divergence-based distribution-fitting as a maximum-likelihood estimation problem. Tech. Rep. METR 2008-11, The University of Tokyo.

Le Roux, J., Ono, N., Sagayama, S., Sep. 2008b. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In: Proceedings of the ISCA Workshop on Statistical and Perceptual Audition (SAPA). pp. 23–28.

Lee, D. D., Seung, H. S., Oct. 1999. Learning of the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Lee, D. D., Seung, H. S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems 13 (Proc. NIPS*2000). The MIT Press, Cambridge, MA, pp. 556–562.

Lu, L., Mao, Y., Liu, W., Zhang, H.-J., Apr. 2003. Audio restoration by constrained audio texture synthesis. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 5. pp. 636–639.

Maher, R. C., Oct. 1993. A method for extrapolation of missing digital audio data. In: 95th AES Convention. New York, pp. 1–19.

Maher, R. C., May 1994. A method for extrapolation of missing digital audio data. Journal of the Audio Engineering Society 42 (5), 350–357.

McAulay, R. J., Quatieri, T. F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (4), 744–754.

Meng, X.-L., Rubin, D. B., Jun. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.

Mørup, M., Schmidt, M. N., 2006. Sparse non-negative matrix factor 2-D deconvolution. Tech. rep., Technical University of Denmark.

Raj, B., Seltzer, M. L., Stern, R. M., 2004. Reconstruction of missing features for robust speech recognition. Speech Communication 43 (4), 275–296.

Raj, B., Stern, R. M., Sep. 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine 22 (5), 101–116.

Rajan, J. J., Rayner, P. J. W., Godsill, S. J., Aug. 1997. A Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler. IEE Proceedings on Vision, Image, and Signal Processing 144 (4), 249–256.

Rayner, P. J. W., Godsill, S. J., Oct. 1991. The detection and correction of artefacts in archived grammophone recordings. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

Reyes-Gomez, M., Jojic, N., Ellis, D. P. W., Oct. 2004. Towards single-channel unsupervised source separation of speech mixtures: the layered harmonics/formants separation-tracking model. In: Proceedings of the ISCA Workshop on Statistical and Perceptual Audition (SAPA). pp. 25–30.

Sajda, P., Du, S., Parra, L. C., Aug. 2003. Recovery of constituent spectra using non-negative matrix factorization. In: Proceedings of SPIE Wavelets X. pp. 321–331.

Schmidt, M. N., Mørup, M., Apr. 2006. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In: Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006). pp. 700–707.

Smaragdis, P., Sep. 2004. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004). pp. 494–499.

Smaragdis, P., Raj, B., Shashanka, M., Sep. 2009. Missing data imputation for spectral audio signals. In: Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP).

Vaseghi, S. V., Rayner, P. J. W., Feb. 1990. Detection and suppression of impulsive noise in speech communication systems. Communications, Speech and Vision, IEE Proceedings I 137 (1), 38–46.

Veldhuis, R., 1990. Restoration of Lost Samples in Digital Audio Signals. Prentice-Hall, Englewood Cliffs, NJ.

Virtanen, T., Mesaros, A., Ryynänen, M., Sep. 2008. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In: Proceedings of the ISCA Workshop on Statistical and Perceptual Audition (SAPA). pp. 17–22.

Wang, D.-L., Brown, G. J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley.

Warren, R. M., Jan. 1970. Perceptual restoration of missing speech sounds. Science 167 (3917), 392–393.

Warren, R. M., 1982. Auditory Perception: A New Synthesis. Pergamon, New York, NY.

Wolfe, P. J., Godsill, S. J., Mar. 2005. Interpolation of missing data values for audio signal restoration using a Gabor regression model. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. V. pp. 517–520.