Musical source separation using time-frequency source priors

Emmanuel Vincent

Musical Source Separation Using Time-Frequency Source Priors

Emmanuel Vincent*

***Abstract*— This article deals with the source separation problem for stereo musical mixtures using prior information about the sources (instrument names and localization). After a brief review of existing methods, we design a family of probabilistic mixture generative models combining modified positive Independent Subspace Analysis (ISA), localization models and Segmental Models (SM). We express source separation as a Bayesian estimation problem and we propose efficient resolution algorithms. The resulting separation methods rely on a variable number of cues including harmonicity, spectral envelope, azimuth, note duration and monophony. We compare these methods on two synthetic mixtures with long reverberation. We show that they outperform methods exploiting spatial diversity only and that they are robust against approximate localization of the sources.**

***Index Terms*— Music, source separation, source priors, independent subspace analysis, localization, segmental model.**

I. Introduction

USER needs regarding large music databases available today raise many issues, among which interactive modification of the data for applications such as karaoke, automatic soloist accompaniment, broadcast of CDs on multichannel devices, post-production of raw recordings, restoration of old recordings and music creation by instrument sampling. Most music signals are mixtures of several sources active simultaneously (musical instruments, voices, synthetic sounds), acquired by synthetic mixing of solo source signals or by recording of real audio scenes. Thus addressing these applications means separating the sources and remixing them accordingly.

A. Problem definition

The mixing operation can generally be expressed as a linear filtering followed by a summation. For simplicity, we consider here time-invariant filtering only. The mixture channels $(x_i)_{1\leq i\leq I}$ are defined by $x_i = \sum_{j=1}^{J} a_{ij} \star s_j + n_i$, where $\star$ denotes convolution, $(s_j)_{1\leq j\leq J}$ the source signals, $(a_{ij})_{1\leq i\leq I,1\leq j\leq J}$ the mixing filters and $(n_i)_{1\leq i\leq I}$ background noises. Source separation consists in estimating the images $(s_{\mathrm{img}\,ij})_{1\leq i\leq I,1\leq j\leq J}$ of the sources defined by $s_{\mathrm{img}\,ij} = a_{ij} \star s_j$ with the best possible audio quality. When prior information about the sources is available (instruments or localization), ordering of the source signals is also required.

Separation of music recordings is particularly difficult since sources generally overlap in the time-frequency plane due to long reverberation and note intervals favored by harmony rules. Moreover sources may have close spatial locations and very similar characteristics, *e.g.* in a violin duo.

Submitted for publication in IEEE *Transactions on Speech and Audio Processing* Special Issue on Statistical and Perceptual Audio Processing on January 31st 2005. Accepted on April 30th 2005.

The author is with the Centre for Digital Music, Queen Mary, University of London, Mile End Road, London E1 4NS (United Kingdom) (Phone: +44 20 7882 5528, Fax: +44 20 7882 7997, Email: emmanuel.vincent@elec.qmul.ac.uk).

B. Existing methods

Many algorithms aim at addressing this problem. They can be classified into three broad (sometimes overlapping) categories: Computational Auditory Scene Analysis (CASA), statistical spatial models and statistical spectral models.

CASA aims at identifying perceived auditory objects (*e.g.* notes in the case of music recordings) and grouping them into auditory streams using psycho-acoustical cues [1]. Basic methods focus on mono recordings, characterize note objects by harmonicity, common onset, correlated modulation and duration of sinusoidal partials, and build note streams based on pitch proximity. Hence they can hardly segregate instruments playing in the same pitch range into different streams. Proximity of spectral centroids [2], matching of timbre features learnt on solo excerpts [3], [4], [5] (spectral envelope, onset duration, *vibrato* amplitude), and more recently location similarity [6], [7] were proposed as supplementary cues to improve instrument segregation. Some of these cues were also shown to improve note transcription by avoiding mistaking chords formed by notes at harmonic intervals with single notes [3], [6]. CASA has a few drawbacks though: the precedence rules between grouping cues are sometimes hard to assess and the correlogram front-end used by most methods prevents identification of masked auditory objects and good separation of sinusoidal partials within the same critical band. Moreover existing methods for multichannel recordings are restricted to non-reverberant mixtures.

Statistical spatial models use simple probabilistic source models to exploit channel diversity on multichannel recordings. Independent Component Analysis (ICA) can be applied when there are less sources than channels and estimates time-invariant demixing filters based on independence of the source waveforms. Generally the mixture is split into frequency subbands and estimated source subbands are grouped based on location [8], correlation of amplitudes across subbands [9] or both [10]. Many ICA algorithms have been proposed, but their performance decreases fast when reverberation increases, since demixing filters become longer and harder to estimate. Detailed information about the sources, such as the text pronounced by the speakers in a speech mixture [11], can help separation but is not always available. Other methods cope with a larger number of sources by supposing they are nearly disjoint in the time-frequency plane and separate them using time-frequency binary masks. Interchannel

Intensity Difference (IID) and Interchannel Time Difference (ITD) [12] are summarized across time-frequency points to estimate the source azimuths and then used to derive optimal masks [13], [14], [12]. These methods fail in time-frequency zones where several sources overlap: time-frequency points containing energy from both "left" and "right" sources are erroneously associated with "center" sources [12], [14] and binary masks generate "burbling" noise [13]. Also these methods are not suited to reverberant mixtures since reverberation creates virtual sources with random locations that need to be associated with the original sources.

Finally, statistical spectral models have been proposed for the separation of mono recordings. Independent Subspace Analysis (ISA) decomposes the mixture power spectrogram as a sum of typical spectra with time-varying weights, builds the source power spectrograms by grouping these spectra into subspaces and computes the source waveforms by inverting their spectrograms [15] or by adaptive Wiener filtering [16]. Typical spectra are either learnt on solo excerpts [16] or estimated from the mixture using ICA [15], positive ICA [17] or Non-negative Matrix Factorization (NMF) [18]. Good results were reported for note transcription on solo recordings [19], [18] and separation of percussions from other instruments [16], [17]. However ICA and NMF badly separate low-intensity notes [20] and produce spurious notes with short duration, and their ability to segregate non-percussive instruments has not been studied. Hidden Markov Models (HMM) solve these issues by learning accurate priors for the log-power spectra of the sources on solo data and by setting a prior on event duration. Satisfying separation results were obtained on speech mixtures with factorial combination of source models [21]. But complex parameter sharing procedures are needed on musical mixtures to avoid overlearning [22], since the number of hidden states for each source (*i.e.* the number of chords it can play) may be very large.

C. Overview of the proposed method

In this article, we propose a new musical source separation method that integrates a variable number of CASA-like cues into a sound statistical framework. More precisely we modify existing statistical spectral and spatial source models and combine them into a single multilayer Bayesian network. Then we express source separation as a Bayesian estimation problem, using instrument-specific parameters learnt on solo recordings. This method does not aim at modeling auditory sound processing, but rather at improving the results of existing statistical models by combining spectral, spatial and temporal cues. In particular it is to our knowledge the first method able to separate mixtures with long reverberation.

In the following we focus on AB narrow instrumental recordings, *i.e.* stereo mixtures of instrumental sources (non vocal) recorded with two far-field omnidirectional microphones spaced about 40 cm. However the proposed method can cope with other stereo recording setups after minor modifications. We suppose that the number of sources, their approximate azimuths and the names of the corresponding instruments are known. Our preliminary work on panoramic mixtures [23] (*i.e.* mixtures where the mixing filters are simply gains) is extended in several ways: we propose a new spatial model that is applicable to AB narrow mixtures and more robust against erroneous source localization, and we derive a new temporal model that better models note durations and takes into account monophony. Also the performance comparison between our method and existing ones gives very different results on reverberant mixtures.

D. Structure of the article

The rest of the article has the following structure. In Section II we present a Bayesian network model of music recordings and propose several joint parametric distributions for observed and hidden variables. We design corresponding source separation algorithms in Section III and test them on synthetic mixtures in Section IV. We conclude in Section V by pointing out other applications of the proposed model.

II. Mixture modeling with time-frequency source priors

A. Front-end

Similarly to most of the methods mentioned above, we do not model mixture waveforms directly. Instead we transform them in the time-frequency domain and compute three relevant observed quantities related to the amount of power, the spatial location and the number of active sources in each time-frequency point. Let $(H_f)_{0\leq f\leq F-1}$ be a set of complex bandpass filters and w a rectangular window of length L. We split each mixture channel x_i into subband signals $(x_i^f)_{0\leq f\leq F-1}$ defined by $x_i^f = H_f \star x_i$ and then into finite support signals $(x_i^{tf})_{0\leq t\leq T-1,\, 0\leq f\leq F-1}$ defined by $x_i^{tf}(u) = w(u - tL)\, x_i^f(u)$. Let us denote $\langle .,. \rangle$ the inner product between two complex signals, $\|.\|^2$ the energy of a signal and $\angle .$ the angle in $]-\pi, \pi]$ of a complex number. We define the log-power spectrum $(o_{tf}^{\mathrm{pow}})_{0\leq t\leq T-1,\, 0\leq f\leq F-1}$ by

$$o_{tf}^{\mathrm{pow}} = \log\left(\frac{\|x_1^{tf}\|^2 + \|x_2^{tf}\|^2}{g_f} + 1\right), \tag{1}$$

where $(g_f)_{0\leq f\leq F-1}$ is a power threshold that prevents o_{tf}^{pow} from dropping to $-\infty$ by forcing $o_{tf}^{\mathrm{pow}} \approx 0$ whenever $\|x_1^{tf}\|^2$ and $\|x_2^{tf}\|^2$ are small compared to g_f. In the following this threshold is set to the absolute auditory masking threshold after appropriate normalization of the mixture signal. We also define the interchannel phase difference $(o_{tf}^{\mathrm{pha}})_{0\leq t\leq T-1,\, 0\leq f\leq F-1}$ by

$$o_{tf}^{\mathrm{pha}} = \angle \langle x_1^{tf}, x_2^{tf} \rangle, \tag{2}$$

and the interchannel coherence $(o_{tf}^{\mathrm{coh}})_{0\leq t\leq T-1,\, 0\leq f\leq F-1}$ by

$$o_{tf}^{\mathrm{coh}} = \frac{|\langle x_1^{tf}, x_2^{tf} \rangle|}{\|x_1^{tf}\| \|x_2^{tf}\|}. \tag{3}$$

Interchannel phase difference is related to ITD [14] and IID is irrelevant in AB narrow mixtures since both channels have nearly equal power spectra. When a single source j is active in (t, f), o_{tf}^{coh} is close to 1 [13] and o_{tf}^{pha} equals the relative phase response a_{jf}^{pha} of the mixing filters of this source at frequency

f. On the contrary, when several sources or reverberation are active in (t, f), o_{tf}^{coh} is notably lower than 1 [13].

Due to the separation method chosen in the following, the filterbank $(H_f)_{0\leq f\leq F-1}$ is subject to two constraints: narrowband filters are needed in the lower frequency range to identify notes with close pitch and separate them by Wiener filtering, but a limited number of subbands is also necessary to avoid overlearning of instrument-specific note spectra. We use a set of two hundred modulated Hanning windows with center frequencies linearly spaced between 30 Hz and 11 KHz on the ERB scale [12] defined by $f_{ERB} = 9.26\log(0.00437 f_{Hz}+1)$, we set the width of the filter main lobes to four times the spacing between central frequencies of adjacent filters, and we partition subbands into $L = 11$ ms time frames.

B. Three-layer generative model

We propose to model observed quantities by a three-layer Bayesian network. The network structure is the following: on each time frame each instrument plays a finite number of notes among possible notes on a discrete semitone pitch scale, then these notes are described more precisely with instantaneous parameters that characterize the corresponding source power spectrum, and in the end the power spectra of all sources are linked to the observed quantities. This generative model generalizes many generative models underlying existing statistical source separation methods. Indeed both the mixture power spectrum (as with ISA or HMM) and the interchannel phase difference (as with binary masking) are modelled, and the source power spectra are related both to continuous descriptors (as with ISA) and to discrete states (as with HMM).

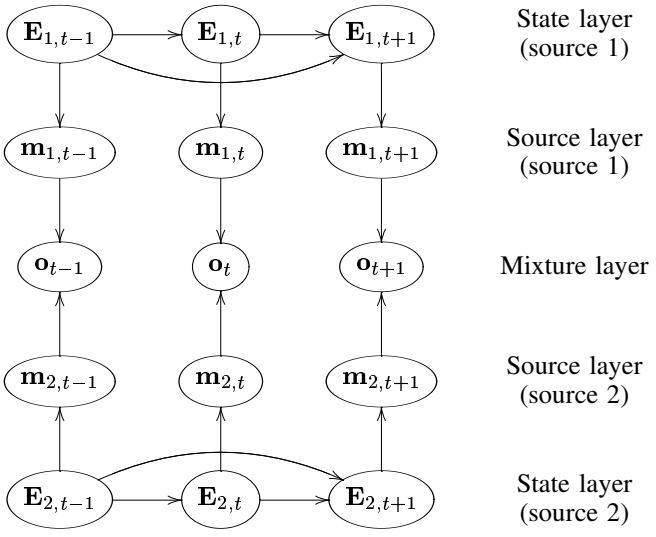


Fig. 1. Simplified graphical representation of a two-source mixture model.

The model is represented in figure 1 in the two-source case. We index by h the MIDI pitch of the possible notes that instrument j can play ($H_j \leq h \leq H'_j$) and we denote by E_{jht} the binary activity state of note h at time t (*i.e.* whether it is actually played or not). We further denote $\mathbf{E}_{jt} = (E_{jht})_{H_j\leq h\leq H'_j}$ the state of source j at time t, $\mathbf{m}_{jt} = (m_{jtf})_{0\leq f\leq F-1}$ its power spectrum and $\mathbf{o}_{jt} = (o_{tf}^{\mathrm{pow}}, o_{tf}^{\mathrm{pha}}, o_{tf}^{\mathrm{coh}})_{0\leq f\leq F-1}$ the observed quantities. The three layers are termed respectively state layer, source layer and mixture layer, and we design parametric priors for each of them in the rest of this section. Note that these priors have not been chosen arbitrarily, but that their shapes have been validated on isolated notes and on synthetic mixtures of solo excerpts (see [24] for details).

C. Source layer

The source layer is the core of the model. Our hypothesis is that the waveforms of different notes are uncorrelated in each time-frequency point, so that the source power spectrum is equal to the sum of note power spectra. We also suppose that the spectrum of each note is stationary and that only its intensity varies. This results in the non-negative additive model

$$m_{jtf} = \sum_{h=H_j}^{H'_j} e_{jht}\,\Phi_{jhf}, \tag{4}$$

where $\mathbf{\Phi}_{jh} = (\Phi_{jhf})_{0\leq f\leq F-1}$ is the normalized power spectrum of note h and e_{jht} its power at time t. The note spectra $(\mathbf{\Phi}_{jh})_{H_j\leq h\leq H'_j}$ account for harmonicity and instrument-dependent spectral envelope cues. The decomposition of the source power spectrum as a weighted sum allows to represent a large set of source spectra with a limited set of note spectra, avoiding complex parameter sharing procedures as with HMMs. Also the weights carry more meaningful information than in ISA, since they are explicitly defined as note powers.

By definition the note power e_{jht} is constrained to 0 when note h is inactive. We suppose that the powers of active notes follow independent log-Gaussian laws. The conditional distribution of the source layer at time t $P_t^{\mathrm{src}} = P((\mathbf{m}_{jt})|(\mathbf{E}_{jt}), (\mathbf{\Phi}_{jh}), (\mu_{jh}^e), (\sigma_{jh}^e))$ equals

$$P_t^{\mathrm{src}} = \prod_{j=1}^{J} \prod_{h\in\mathcal{A}_{jt}} \mathcal{N}(\log e_{jht}; \mu_{jh}^e, \sigma_{jh}^e), \tag{5}$$

where $\mathcal{A}_{jt}$ is the set of active notes for source j at time t and $\mathcal{N}(a; \mu, \sigma)$ is the Gaussian density of mean μ and standard deviation σ evaluated in a.

D. Mixture layer

The mixture layer defines the observed quantities as functions of the source power spectra and the mixing parameters. We parameterize the mixing filters by their common power response $\mathbf{a}^{\mathrm{pow}} = (a_f^{\mathrm{pow}})_{0\leq f\leq F-1}$ and by their relative phase responses $(\mathbf{a}_j^{\mathrm{pha}})_{1\leq j\leq J} = (a_{jf}^{\mathrm{pha}})_{1\leq j\leq J, 0\leq f\leq F-1}$ computed from the source azimuths by the beamforming equation [14]

$$a_{jf}^{\mathrm{pha}} = 2\pi\frac{fd}{c}\sin\theta_j \mod 2\pi, \tag{6}$$

where f is the subband central frequency, d the distance between sensors, c the speed of sound and θ_j the azimuth of source j. We also suppose that the background noise is due to stationary noise sources and we describe it by its power spectrum $\mathbf{n} = (n_f)_{0\leq f\leq F-1}$ and by the corresponding relative phase response $\mathbf{b}^{\mathrm{pha}} = (b_f^{\mathrm{pha}})_{0\leq f\leq F-1}$ (possibly corresponding to different azimuths in different subbands). We propose two models of the layer distribution.

1) Mono model: The mono model takes into account only the fit between source and mixture power spectra. We suppose that the waveforms of different sources are uncorrelated in each time-frequency-point so that sources add in the power spectral domain. The observed log-power spectrum writes

$$o_{tf}^{\mathrm{pow}} = \log\Big(\sum_{j=1}^{J} a_f^{\mathrm{pow}} m_{jtf} + n_f\Big) + \epsilon_{tf}^{\mathrm{pow}}, \tag{7}$$

where $\epsilon_{tf}^{\mathrm{pow}}$ is the residual error modelled as a Gaussian noise with fixed standard deviation $\sigma^{\epsilon\,\mathrm{pow}}$. The conditional distribution of the mixture layer at time t $P_t^{\mathrm{mix}} = P(\mathbf{o}_t|(\mathbf{m}_{jt}), \mathbf{n}, \mathbf{a}^{\mathrm{pow}})$ equals

$$P_t^{\mathrm{mix}} = \prod_{f=0}^{F-1} \mathcal{N}(\epsilon_{tf}^{\mathrm{pow}}; 0, \sigma^{\epsilon\,\mathrm{pow}}). \tag{8}$$

Usual ISA and NMF generative models represent residual error by Gaussian noise or Poisson noise in the power domain [19], [18], which allots more importance to residual error in high-intensity time-frequency zones. This results in low-intensity notes being considered as absent [20] and notes of the same pitch from different instruments being badly discriminated. Indeed note power spectral envelopes are very similar for different instruments with only the first few partials having significantly nonzero power. Representing the residual error by Gaussian noise in the log-power domain addresses these issues. Note that this is close to the modelling of log-power spectra by mixtures of Gaussians in HMM [21].

2) Stereo model: The stereo model extends the mono model by taking into account the interchannel phase difference, which accounts for the azimuth cue. Again we suppose that the waveforms of different sources are uncorrelated in each time-frequency-point. After adding a residual error term to the expression in [12], this gives

$$o_{tf}^{\mathrm{pha}} = \angle\Big(\sum_{j=1}^{J} a_f^{\mathrm{pow}} m_{jtf} \exp(i a_{jf}^{\mathrm{pha}}) + n_f \exp(i b_f^{\mathrm{pha}})\Big) + \epsilon_{tf}^{\mathrm{pha}} \mod 2\pi. \tag{9}$$

We suppose that the residual $\epsilon_{tf}^{\mathrm{pha}}$ belongs to $]-\pi, \pi]$ and that its distribution is proportional to a Gaussian noise whose standard deviation $\sigma_{tf}^{\epsilon\,\mathrm{pha}}$ is defined from the interchannel coherence o_{tf}^{coh} by

$$\sigma_{tf}^{\epsilon\,\mathrm{pha}} = \sigma^{\epsilon\,\mathrm{pha}}\,(1 - o_{tf}^{\mathrm{coh}})^{\lambda^{\mathrm{pha}}}, \tag{10}$$

where $\sigma^{\epsilon\,\mathrm{pha}}$ is the maximal standard deviation and λ^{pha} is a positive exponentiation factor. The conditional distribution of the mixture layer then becomes

$$P_t^{\mathrm{mix}} = \prod_{f=1}^{F-1} \mathcal{N}(\epsilon_{tf}^{\mathrm{pow}}; 0, \sigma^{\epsilon\,\mathrm{pow}}) \frac{\mathcal{N}(\epsilon_{tf}^{\mathrm{pha}}; 0, \sigma_{tf}^{\epsilon\,\mathrm{pha}})}{\int_{-\pi}^{\pi} \mathcal{N}(\epsilon; 0, \sigma_{tf}^{\epsilon\,\mathrm{pha}})\, d\epsilon}. \tag{11}$$

Our previous work on non-reverberant panoramic mixtures modelled the residual on IID as a Gaussian noise with fixed standard deviation [23]. We found that this model was too simplistic because IID is less predictable in zones where several sources overlap. Indeed, when several sources have similar power in a given time-frequency point a small variation of their powers may result in a large deviation of IID, whereas when a single source is prominent IID does not depend on the power of this source. This remark also applies to interchannel phase difference. Moreover, interchannel phase difference information should be given less importance in reverberant zones since reverberation generates virtual sources with random azimuths which have to be grouped with the original sources relying on other cues than location. Equation 10 models these beliefs by allotting the residual a larger variance in zones with low interchannel coherence, which are precisely the zones containing several sources or reverberation.

E. State layer

The state layer describes the possible values taken by the state of each source on each frame and its temporal evolution. Again we propose two different models.

1) Factorial model: The simplest model, called factorial Bernoulli model or more simply factorial model, supposes that the states $(E_{jht})_{1\le j\le J,\, H_j\le h\le H'_j,\, 0\le t\le T-1}$ follow independent Bernoulli priors with same parameter Z. The distribution of the state layer $P^{\mathrm{sta}} = P((\mathbf{E}_{jt}))$ is expressed as

$$P^{\mathrm{sta}} = \prod_{j=1}^{J} \prod_{t=0}^{T-1} (1-Z)^{\#\mathcal{A}_{jt}} Z^{\#\mathcal{I}_{jt}}, \tag{12}$$

where $\#\mathcal{A}_{jt}$ and $\#\mathcal{I}_{jt}$ are respectively the number of active and inactive notes for source j at time t. Note that independence of the states implies independence of the note powers $(e_{jht})_{1\le j\le J,\, H_j\le h\le H'_j,\, 0\le t\le T-1}$ (instead of conditional independence only), which is the usual assumption in ISA. The parameter Z is then an explicit sparsity factor: the higher it is the more probable are state variables containing a low number of active notes on each time frame.

2) Segmental model: The factorial model can be improved by adding to the sparsity prior a temporal persistence prior. In our previous work [23], we supposed that the states of different notes or instruments were independent and that the state series $(E_{jht})_{0\le t\le T-1}$ of a each note h was a first order Markov chain with two states activity/inactivity. This model favors long activity/inactivity patterns for each note, but it has two drawbacks. Firstly, note duration is modelled by an exponential prior, which gives a large probability to very short and very long durations. The performance of the model for source separation is then only slightly better than the performance of the factorial model, since it still finds spurious notes with short duration or short silences within actual notes [23]. It is possible to concentrate note duration around a peak value and to impose a minimal duration by representing the state series of each note by a Markov chain with more than two states [22], however this increases even more the probability of very long durations. Secondly, the hypothesis that the states of different notes are independent is unrealistic. In practice, notes from the same instrument either begin at the same time

within chords or follow each other with a minimal delay within phrases. We address these two issues by building a new state layer model for monophonic instruments (*i.e.* instruments that cannot play chords). This so-called segmental model accounts for note duration and monophony cues using explicit duration distributions. It could also be applied to polyphonic instruments after minor modifications.

For each source j, let us partition the time line into successive segments indexed by r ($0 \leq r \leq R_j - 1$), where the start time t_r of each segment corresponds to the attack of a new note (whose pitch is different from the pitches of active notes in the previous time frame $t_r - 1$). The duration d_r of this note may be greater than the duration $t_{r+1} - t_r$ of the segment due to reverberation, or it may be less in case the end of the segment contains silence. We model the durations of all notes and segments by independent log-Gaussian distributions with lower thresholds. The note duration prior $\mathcal{D}^{\mathrm{not}}$ is defined by

$$\mathcal{D}^{\mathrm{not}}(d) = \begin{cases} \dfrac{\mathcal{N}(\log d; \mu^n, \sigma^n)}{\sum_{d' \geq d^n} \mathcal{N}(\log d'; \mu^n, \sigma^n)} & \text{if } d \geq d^n, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

and the segment duration prior $\mathcal{D}^{\mathrm{seg}}$ is defined by a similar equation with parameters μ^s, σ^s and d^s. We suppose that the source state on the first time frame $t = 0$ follows a factorial Bernoulli prior and that the pitch of each new note is uniformly distributed among the pitches of inactive notes in the previous time frame. The distribution of the state layer reads

$$P^{\mathrm{sta}} = \prod_{j=1}^{J} \prod_{r=0}^{R_j - 1} \mathcal{D}^{\mathrm{not}}(d_r) \prod_{r=0}^{R_j - 1} \mathcal{D}^{\mathrm{seg}}(t_{r+1} - t_r) \\ (1 - Z)^{\#\mathcal{A}_{j,0}} Z^{\#\mathcal{I}_{j,0}} \prod_{r=1}^{R_j - 1} (\#\mathcal{I}_{j,t_r - 1})^{-1}. \tag{14}$$

This model generalizes segmental models of speech, where each segment contains a single phoneme [25]. Here each segment contains not only a new note, but also reverberation of the previous notes. This modeling of reverberation by temporal continuity is the main cue that allows to group it with the corresponding sources. In practice, separation errors may arise near observations boundaries $t = 0$ and $t = T - 1$, due to the fact that notes that are active on these boundaries may still be active beyond them and thus have a shorter measured duration than expected. We circumvent this issue by modelling the duration of these notes with another distribution $\mathcal{Q}^{\mathrm{not}}$ which is proportional to the cumulative distribution of $\mathcal{D}^{\mathrm{not}}$. We also model the durations of the first and last segments by another distribution $\mathcal{Q}^{\mathrm{seg}}$ defined similarly.

F. Weighted Bayes law

We gather the distributions of all layers into a single joint distribution using Bayes law. In the following we suppose that the parameters $(\theta_j)_{1 \leq j \leq J}$, $\sigma^{\epsilon\,\mathrm{pow}}$, $\sigma^{\epsilon\,\mathrm{pha}}$, λ^{pha}, Z, μ^n, σ^n, d^n, μ^s, σ^s and d^s are fixed. We denote $\mathcal{M} = ((\mathbf{\Phi}_{jh}), (\mu^e_{jh}), (\sigma^e_{jh}))$ instrument-specific parameters and $\mathbf{\Theta} = (\mathbf{a}^{\mathrm{pow}}, \mathbf{n}, \mathbf{b}^{\mathrm{pha}})$ unknown mixing parameters and we set uniform priors on them. Experimentally the parametric distributions defining each layer do not model observed data perfectly. For example, the residual errors on the observed quantities are generally loosely correlated between adjacent time-frequency points. Representing them as independent variables increases the importance given to the probability of the mixture layer compared to the probability of other layers. Similarly, the note powers on successive time frames are correlated and representing them as independent variables exaggerates the importance of the source layer probability. We take this into account by replacing the "naive" Bayes law by a weighted Bayes law [26]

$$P^{\mathrm{tot}} = P((\mathbf{o}_t), (\mathbf{m}_{jt}), (\mathbf{E}_{jt}), \mathbf{\Theta}, \mathcal{M}) \\ \propto \Big(\prod_{t=0}^{T-1} P_t^{\mathrm{mix}}\Big)^{w_{\mathrm{mix}}} \Big(\prod_{t=0}^{T-1} P_t^{\mathrm{src}}\Big)^{w_{\mathrm{src}}} P^{\mathrm{sta}}, \tag{15}$$

where w_{mix} and w_{src} are weights comprised between 0 and 1. Some ISA algorithms use similar weights without providing this probabilistic interpretation [17], [16]. In the following, we use $w_{\mathrm{src}} = 0.5$, $w_{\mathrm{mix}} = 0.5$ for the mono model and $w_{\mathrm{mix}} = 0.25$ for the stereo model since we believe that this better models the true distribution of the data. However the performance of the results is exactly the same on average than the performance obtained using the "naive" Bayes law.

III. Bayesian source separation

A. Approximate Bayesian source estimation

Now that we have built probabilistic models of music mixtures, we apply these models to source separation in the following way. The Maximum *A Posteriori* (MAP) estimator of the image of source j on channel i writes $\widehat{s_{\mathrm{img}\,ij}} = \arg\max P(s_{\mathrm{img}\,ij} | (x_i), \mathcal{M})$. Model variables may be introduced by developing $P(s_{\mathrm{img}\,ij} | (x_i), \mathcal{M})$ as the integral of $P(s_{\mathrm{img}\,ij}, (\mathbf{m}_{jt}), \mathbf{\Theta} | (x_i), \mathcal{M})$ over $(\mathbf{m}_{jt})_{1 \leq j \leq J,\, 0 \leq t \leq T-1}$ and $\mathbf{\Theta}$. Unfortunately this integral is far from tractable. Approaching it by the largest value of the integrand, we obtain

$$\widehat{s_{\mathrm{img}\,ij}} \approx \arg\max P(s_{\mathrm{img}\,ij} | (x_i), \widehat{\mathbf{\Theta}}, \widehat{(\mathbf{m}_{jt})}), \tag{16}$$

where

$$\widehat{\mathbf{\Theta}}, \widehat{(\mathbf{m}_{jt})} = \arg\max P(\mathbf{\Theta}, (\mathbf{m}_{jt}) | (\mathbf{o}_t), \mathcal{M}). \tag{17}$$

From these equations, we see that source separation involves three successive steps: learning model parameters $\mathcal{M}$ on learning data, inferring jointly source power spectra $(\mathbf{m}_{jt})_{1 \leq j \leq J,\, 0 \leq t \leq T-1}$ and unknown mixing parameters $\mathbf{\Theta}$ on a mixture, and filtering it to extract the source waveforms $(s_{\mathrm{img}\,ij})_{1 \leq i \leq I,\, 1 \leq j \leq J}$. These steps are described separately in the rest of this section, ending by the first one for convenience.

B. Inference step

Estimating the source power spectra is equivalent to estimating the note states and powers. Let us suppose for a moment that the mixture state at time t $\mathbf{E}_t = (\mathbf{E}_{jt})_{1 \leq j \leq J}$ is fixed. If note h from instrument j is active at this time, then the derivative of $\log P^{\mathrm{tot}}$ versus its log-power $\log e_{jht}$ equals

$$\frac{\partial \log P^{\mathrm{tot}}}{\partial \log e_{jht}} = \frac{w_{\mathrm{mix}}}{\sigma^{\epsilon\,\mathrm{pow}\,2}} \sum_{f=0}^{F-1} \epsilon^{\mathrm{pow}}_{tf} \pi_{jhtf} - \frac{w_{\mathrm{src}}}{\sigma^{e\,2}_{jh}} (\log e_{jht} - \mu^e_{jh}) \tag{18}$$

when the mono model of the mixture layer is used, where

$$\pi_{jhtf} = \frac{a_f^{\mathrm{pow}} e_{jht}\, \Phi_{jhf}}{\sum_{j=1}^{J} a_f^{\mathrm{pow}} m_{jtf} + n_f} \tag{19}$$

is the proportion of the total power due to this note in time-frequency point (t, f). The optimal powers $(e_{jht})_{1 \leq j \leq J,\, h \in \mathcal{A}_{jt}}$ cannot be found analytically. We select their initial values from a few random trials and we reestimate them iteratively using a joint approximate second-order Newton method (or fixed-point algorithm). After supposing that the cross-derivatives $(\partial^2 \log P^{\mathrm{tot}} / \partial e_{jht} \partial e_{j'h't})_{1 \leq j,j' \leq J,\, h.h' \in \mathcal{A}_{jt},\, j' \neq j \text{ or } h' \neq h}$ are nearly equal to zero and modifying slightly the double-derivatives $(\partial^2 \log P^{\mathrm{tot}} / \partial e_{jht}^2)_{1 \leq j \leq J,\, h \in \mathcal{A}_{jt}}$, we obtain a stable update rule $\log e_{jht} \leftarrow \log e_{jht} + \Delta \log e_{jht}$ for each note h with

$$\Delta \log e_{jht} = \frac{\dfrac{w_{\mathrm{mix}}}{\sigma^{\epsilon\,\mathrm{pow}\,2}} \displaystyle\sum_{f=0}^{F-1} \epsilon_{tf}^{\mathrm{pow}} \pi_{jhtf} - \dfrac{w_{\mathrm{src}}}{\sigma_{jh}^{e\,2}} (\log e_{jht} - \mu_{jh}^{e})}{\dfrac{w_{\mathrm{mix}}}{\sigma^{\epsilon\,\mathrm{pow}\,2}} \displaystyle\sum_{f=0}^{F-1} \pi_{jhtf} + \dfrac{w_{\mathrm{src}}}{\sigma_{jh}^{e\,2}}}. \tag{20}$$

This update rule performs a compromise between the power predicted from the data and the one predicted from the prior model, with the relevance of the observed power spectrum o_{tf}^{pow} in each subband f being determined by the weight π_{jhtf}. When the estimated values of the note powers imply that note h from instrument j is masked by other notes or by background noise in time-frequency point (t, f), then $\pi_{jhtf} \approx 0$ and the value of o_{tf}^{pow} is considered irrelevant for the purpose of estimating e_{jht} more precisely. This means that the model naturally deals with missing data without needing explicit time-frequency masks as in CASA methods [4], [5]. A similar update rule may be derived when the stereo model is chosen. In this case, the relevance of the observed interchannel phase difference o_{tf}^{pha} is inversely proportional to $\sigma_{tf}^{\epsilon\,\mathrm{pha}\,2}$.

Generally the mixture states are unknown and the state space is too large to test all possible combinations. For instance, the factorial state model contains approximately 3.10^8 mixture states on each time frame for a mixture of cello and violin whose playing ranges span forty-six semitones each when the number of simultaneous notes is limited to three per instrument (this limit being often reached due to reverberation). Thus we use heuristic criteria to reduce the size of the search space. When the factorial state model is used, mixture states are estimated using an iterative jump procedure on each time frame inspired from [19]. At first, the estimated mixture state is supposed to contain only inactive notes. Then, at each iteration, the unknown mixing parameters $\boldsymbol{\Theta}$ are updated based on the current state estimation (using Newton update rules), the value of P^{tot} is computed for all states containing one active note more or less than the current estimated state (using the Newton algorithm described above) and the best state is selected as the current state for the following iteration. This procedure converges when further activating or disactivating notes does not increase the value of P^{tot}. The optimal number of active notes is not fixed *a priori*. When the segmental state model is chosen, a preliminary estimation is performed using the factorial state model to estimate unknown mixing parameters and rule out very improbable states. Then mixture states are estimated using standard beam search. The algorithm hypothesizes partial state paths by scanning time frames in ascending order and ruling out unprobable paths using "acoustic pruning" and "histogram pruning" heuristics [27]. The optimal path is selected once all time frames have been observed. This technique is similar to the blackboard architecture often used in CASA [1].

C. Extraction step

Once the note states and powers and the mixing parameters have been estimated, the source power spectra are reconstructed using equation 4 and the source images are extracted by adaptive Wiener filtering of each mixture channel. The image $s_{\mathrm{img}\,ij}^{tf}$ of source j on channel i in time-frequency point (t, f) is estimated by

$$s_{\mathrm{img}\,ij}^{tf} = \frac{a_f^{\mathrm{pow}}\, m_{jtf}}{\sum_{j=1}^{J} a_f^{\mathrm{pow}}\, m_{jtf} + n_f}\, x_i^{tf}, \tag{21}$$

the first term of this equation being the proportion of the total power in this point due to source j. In each subband f, the finite support signals $(s_{\mathrm{img}\,ij}^{tf})_{0 \leq t \leq T-1}$ are added together to form the subband signal $s_{\mathrm{img}\,ij}^{f}$. Finally the source image $s_{\mathrm{img}\,ij}$ itself is reconstructed from the subband signals $(s_{\mathrm{img}\,ij}^{f})_{0 \leq f \leq F-1}$ using standard filterbank inversion [28]. Source images extracted by Wiener filtering sound more natural than those obtained by inverting the source power spectrograms, because their characteristics are picked with full details from the mixture instead of being only grossly predicted by the model.

D. Learning step

Prior to separation of mixtures, model parameters have to be learnt on separate mono data labelled with note states and mixing parameters. The Maximum Likelihood (ML) estimator of instrument-specific parameters is $\widehat{\mathcal{M}} = \arg\max P(\mathcal{M} | (\mathbf{o}_t), \boldsymbol{\Theta}, (\mathbf{E}_{jt}))$. This estimator involves again computing an integral over $(\mathbf{m}_{jt})_{1 \leq j \leq J,\, 0 \leq t \leq T-1}$, which is intractable. The same approximation technique as above yields

$$\begin{cases} \widehat{\mathcal{M}} \approx \arg\max P(\mathcal{M} | (\mathbf{o}_t), \boldsymbol{\Theta}, \widehat{(\mathbf{m}_{jt})}), \\ \widehat{(\mathbf{m}_{jt})} = \arg\max P((\mathbf{m}_{jt}) | (\mathbf{o}_t), \boldsymbol{\Theta}, (\mathbf{E}_{jt}), \mathcal{M}). \end{cases} \tag{22}$$

We solve these coupled equations using an approximate Expectation/Maximization (EM) algorithm that reestimates $(\mathbf{m}_{jt})_{1 \leq j \leq J,\, 0 \leq t \leq T-1}$ and $\mathcal{M}$ alternatively till convergence. The E step consists in estimating the note powers as described above. The M step involves updating the note spectra $(\boldsymbol{\Phi}_{jh})_{1 \leq j \leq J,\, H_j \leq h \leq H'_j}$ by an approximate second-order Newton algorithm and replacing the parameters of the note power priors $(\mu_{jh}^{e}, \sigma_{jh}^{e})_{1 \leq j \leq J,\, H_j \leq h \leq H'_j}$ by their empirical values.

Model parameters may be learnt on all kinds of labelled learning data, ranging from isolated notes to mixtures of several instruments. However, learning on isolated notes is safer because the whole pitch range and playing styles are available for each instrument, and because note spectra are

fully observed instead of being possibly masked by other notes. Databases of isolated notes were used by some CASA algorithms [3], [4] and gave better results than databases of solos in our experiments.

IV. Evaluation examples

We evaluate our proposals on synthetic two-source mixtures with long reverberation. A mixture of clarinet and violin and a mixture of cello and violin were created using ten-second solo excerpts from music CDs and AB narrow impulse responses recorded at IRCAM ($T_{60} = 800$ ms reverberation time). Instrument-specific parameters were learnt on isolated notes from the RWC database [29]. After observing a few solo excerpts and synthetic mixtures, parameters of the state layer were fixed to $Z = 0.96$, $\mu^n = \log(50)$, $\mu^s = \log(30)$, $\sigma^n = \sigma^s = 0.2$ and $d^n = d^s_j = 20$ and parameters of the mixture layer to $\sigma^{\epsilon\,\mathrm{pow}} = 1.4$, $\sigma^{\epsilon\,\mathrm{pha}} = 2.4$ and $\lambda^{\mathrm{pha}} = 0.2$. Depending on the models chosen for the mixture and state layers, four source separation methods result from our proposals. We compare their performance with two simpler methods: ICA (using the algorithm defined in [9] on 2048 subbands) and spatial masking (*i.e.* estimation of the unconstrained source power spectra minimizing $|\epsilon^{\mathrm{pha}}_{tf}|$ for each (t, f) followed by Wiener filtering). Other ICA algorithms resulted in a similar performance. Note that spatial masking exploits prior information about the source azimuths, whereas ICA does not. Note also that the proposed method could cope with more than two sources, contrary to ICA.

In a first experiment, the source interchannel delays $(\frac{d}{c}\sin\theta_j)_{1\leq j\leq 2}$ were fixed to 4.92 and -1.15 samples, corresponding to approximate azimuths of -20° and +5°. Figure 2 shows that the corresponding relative phase responses $(\mathbf{a}^{\mathrm{pha}}_j)_{1\leq j\leq 2}$ are close to the real unknown responses. However large differences appear around 350 Hz, which corresponds to the fundamental frequency of the cello notes. Thus the azimuth cue is unreliable in this frequency range. Source separation results are available for listening on `http://www.elec.qmul.ac.uk/people/emmanuelv/IEEE05/` and evaluated in table I using the median values of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) [30] computed on 200ms frames. Separation performance varies a lot depending on the method. ICA fails completely to separate both mixtures, and the estimated sources sound pretty much as the original mixtures. Spatial masking better reduces crosstalk in both mixtures, however the separated sources contain continuous "burbling" noise at a very annoying level. The mono factorial model provides good results on the first mixture, which proves its ability to separate instruments with distinct spectral envelope characteristics. In this case, the stereo factorial model improves the separation performance only slightly. On the contrary, the mono factorial model fails to separate the second mixture and associates it nearly wholly to the violin source because cello and violin notes have similar spectral envelopes. The stereo factorial model then demonstrates its ability to exploit spatial information efficiently by providing better results on this mixture. Finally segmental models remove most of the time-localized artifacts generated by factorial models and result in a better performance on both mixtures. On average, separation based on stereo segmental models performs about 19 dB better than ICA and 11 dB better than spatial masking according to the SDR measure.

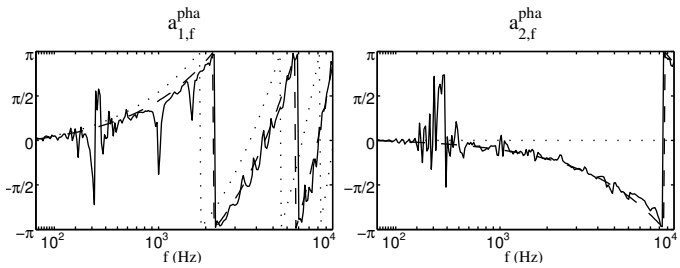


Fig. 2. Relative phase responses of the mixing filters (plain) compared to the responses predicted from the real source azimuths (dashed) and from the erroneous source azimuths (dotted).

TABLE I
Source separation performance on two-source mixtures.

Separation method	SDR	SIR	SAR	SDR	SIR	SAR
First mixture	Clarinet (dB)			Violin (dB)		
ICA	-1	4	4	-9	-3	3
Spatial masking	4	15	4	1	6	8
Mono factorial model	14	25	16	9	21	12
Mono segmental model	18	37	19	14	28	15
Stereo factorial model	15	28	16	11	25	12
Stereo segmental model	16	37	16	13	28	14
Second mixture	Cello (dB)			Violin (dB)		
ICA	-3	7	0	-3	2	4
Spatial masking	4	14	5	5	12	7
Mono factorial model	-17	12	-15	1	1	26
Mono segmental model	N/S	N/S	N/S	2	3	23
Stereo factorial model	7	26	7	7	11	13
Stereo segmental model	13	39	14	16	34	17

In a second experiment, the source interchannel delays were fixed to 6 and 0 samples, corresponding to 5° azimuth error on both sources. Figure 2 shows that the corresponding relative phase responses $(\mathbf{a}^{\mathrm{pha}}_j)_{1\leq j\leq 2}$ are now very different from the real unknown responses above 3 KHz, and thus the observed azimuth becomes a very unreliable cue in the upper frequency range. Despite this, the performance of the proposed algorithms varied about 1 dB only. Observation of the data shows that this is because the interchannel coherence (as defined in equation 3) naturally takes small values in the upper frequency range of reverberant mixtures, even in time-frequency zones containing energy from a single source, so that the stereo models give a smaller weight to spatial information in this range. This robustness towards erroneous source azimuths suggests that the proposed model would also perform well on real mixtures involving small source movements.

V. Conclusion and perspectives

In this article we proposed a family of source separation methods for stereo mixtures of instrumental sources based on multilayer Bayesian network models of short term power

spectrum and interchannel phase difference. We defined new models for each layer and designed corresponding algorithms that exploit instrument-specific parameters learnt on isolated notes and approximate source azimuths given *a priori*. These algorithms may be seen as generalizing some existing statistical source separation algorithms in order to take into account a larger number of CASA-like cues. To our knowledge, they are the first algorithms able to provide a satisfying separation performance on mixtures with long reverberation.

The main advantage of our approach lies in the generality of the Bayesian network formalism. The proposed models may be improved by modifying only some parts of the layer models and the estimation algorithms depending on the kind of mixture and on the wanted tradeoff between performance and computational tractability. For example the mixture layer could be modified to cope with other kinds of stereo recording setups, and the other layers could be complexified to represent tempo and rhythm, chord probabilities, note attack and release, and temporal continuity of the note intensities.

The most straightforward application of this work concerns the extraction of particular sounds from music databases where the number, names and approximate azimuths of the instruments in each mixture are available as metadata. However the algorithms proposed for source separation may also be applied to polyphonic transcription, since they estimate note states and intensities in their first stage. Moreover the models could be used to estimate automatically the needed metadata. We already obtained satisfying preliminary results for Bayesian instrument identification in solo and duo recordings using the mono factorial model [31]. But we suppose that the stereo model could perform even better. Also the histogram estimation method for source azimuths proposed in [12] failed on the test mixtures of this article (even after removing non-relevant interchannel phase difference information in time-frequency points with low interchannel coherence). But the time-frequency models of the sources could possibly provide better estimation of source azimuths since harmonicity and spectral envelope cues help summarizing the observed interchannel phase difference across time-frequency points.

ACKNOWLEDGEMENTS

Most of this work was performed while the author was with the Analysis-Synthesis Group of IRCAM, Paris, France. The author also wants to thank N. Mitianoudis and M. Davies for providing their code for convolutive ICA.

REFERENCES

[1] D. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, MIT, 1996.

[2] D. Godsmark and G. Brown, "A blackboard architecture for computational auditory scene analysis," *Speech Communication*, vol. 27, 1999.

[3] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of bayesian probability network to music scene analysis," in *Working notes of IJCAI Workshop on CASA*, 1995.

[4] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proc. IJCAI Worshop on CASA*, 1999.

[5] J. Eggink and G. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in *Proc. ISMIR*, 2003.

[6] T. Nakatani, "Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition," Ph.D. dissertation, Kyoto University, 2002.

[7] Y. Sakuraba and H. Okuno, "Note recognition of polyphonic music by using timbre similarity and direction proximity," in *Proc. ICMC*, 2003.

[8] L. Parra and C. Alvino, "Geometric source separation : merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, 2002.

[9] N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Speech and Audio processing*, vol. 11, no. 5, 2003.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," in *Proc. ICA*, 2003.

[11] M. Reyes-Gomez, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. ICASSP*, 2003.

[12] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the ASA*, vol. 114, no. 4, 2003.

[13] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Proc. AES 22nd Conference on Vitual, Synthetic and Entertainment Audio*, 2002.

[14] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. DAFx*, 2003.

[15] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. ICMC*, 2000.

[16] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non-negative sparse representation for Wiener based source separation with a single sensor," in *Proc. ICASSP*, 2003.

[17] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. ICMC*, 2003.

[18] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA*, 2003.

[19] S. Abdallah and M. Plumbley, "An ICA approach to automatic music transcription," in *Proc. 114th AES Convention*, 2003.

[20] D. Fitzgerald, E. Coyle, and B. Lawlor, "Independent subspace analysis using locally linear embedding," in *Proc. DAFx*, 2003.

[21] S. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000.

[22] C. Raphael, "Automatic transcription of piano music," in *Proc. ISMIR*, 2002.

[23] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. ICA*, 2004.

[24] E. Vincent, "Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux," Ph.D. dissertation, IRCAM, 2004.

[25] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, 1996.

[26] D. Hand and K. Yu, "Idiot's bayes - not so stupid after all ?" *International Statistical Review*, vol. 69, no. 3, 2001.

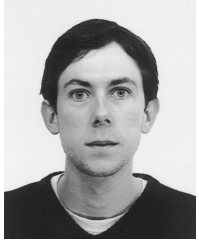
[27] S. Ortmanns, H. Ney, and A. Eiden, "Language-model look-ahead for large vocabulary speech recognition," in *Proc. ICSLP*, 1996.

[28] M. Slaney, D. Naar, and R. Lyon, "Auditory model inversion for sound separation," in *Proc. ICASSP*, 1994.

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. ISMIR*, 2003.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Speech and Audio Processing*, 2005, to appear.

[31] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. ISMIR*, 2004.



Emmanuel Vincent graduated from École Normale Supérieure, Paris, France in 2001. He obtained the Ph.D. degree in acoustics, signal processing and computer science applied to music from the University of Paris-VI Pierre et Marie Curie, Paris, France, in 2004. He is currently a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, London, United Kingdom. His research focuses on structured probabilistic modeling of audio signals applied to blind source separation, indexing and object coding of musical audio.