Contributions to audio source separation and content description

Emmanuel Vincent

▶ **To cite this version:**

Emmanuel Vincent. Contributions to audio source separation and content description. Signal and Image processing. Université Rennes 1, 2012. tel-00758517v1

HAL Id: tel-00758517

https://theses.hal.science/tel-00758517v1

Submitted on 28 Nov 2012 (v1), last revised 18 Sep 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANNÉE 2012

HDR / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

Mention : Traitement du Signal
présentée par

Emmanuel Vincent

préparée et soutenue à l'IRISA - UMR 6074 le 23/11/2012

Contributions à la séparation de sources et à la description des contenus audio.

devant le jury composé de :

Christine GUILLEMOT
Directeur de Recherche, INRIA Rennes – Bretagne Atlantique/président

Yves GRENIER
Professeur, Télécom ParisTech/ rapporteur

Christian JUTTEN
Professeur, Université Joseph Fourier/ rapporteur

Maurizio OMOLOGO
Senior Researcher, FBK-irst/rapporteur

Anssi KLAPURI
Lecturer, Queen Mary University of London/examinateur

Contents

Remerciements

Je tiens avant tout à remercier les membres du jury: Yves Grenier, Christian Jutten et Maurizio Omologo pour avoir accepté la tâche de rapporteurs, ainsi que Christine Guillemot et Anssi Klapuri pour l'intérêt qu'ils ont bien voulu porter à mon travail.

Mes remerciements vont ensuite aux fondateurs et enseignants du DEA ATIAM qui m'ont ouvert la porte vers ce domaine passionnant, à la frontière des mathématiques appliquées, de l'informatique, de l'acoustique et de la musique. Merci aussi à Xavier Rodet et Mark Plumbley, avec qui j'ai eu la chance d'effectuer ma thèse et mon postdoc et qui m'ont toujours laissé libre de choisir mes propres directions de recherche.

Un très grand merci à Kamil Adiloğlu, Alexis Benichoux, Ngoc Duong, Valentin Emiya, Nobutaka Ito, Dimitris Moreau, Alexey Ozerov, Stanisław Racsyński, Gabriel Sargent, Laurent Simon et Joachim Thiemann, sans oublier Charles Blandin et les autres stagiaires, pour leur motivation et leurs efforts. Ils reconnaîtront des aspects de leurs travaux dans ce document. Plus généralement, merci à tous les membres de l'équipe METISS pour l'esprit d'équipe qu'ils ont entretenu pendant toutes ces années. Mention particulière à Frédéric Bimbot et Rémi Gribonval pour m'avoir montré l'exemple par leurs qualités humaines et scientifiques exceptionnelles, et à Stéphanie Lemaile pour sa disponibilité et son efficacité à résoudre les casse-tête administratifs les plus divers. Je leur en suis hautement reconnaissant.

J'exprime en fin toute ma gratitude à mes proches pour leur soutien indéfectible: ma femme Julie, mes enfants Thaïs et Gabriel, mes parents, frère et soeurs, et mes amis de tous horizons.

Foreword

Audio data occupy a central position in our life, whether it is for spoken communication, personal videos, radio and television, music, cinema, video games, or live entertainment. This raises a range of application needs from signal enhancement to information retrieval, including content repurposing and interactive manipulation.

Real-world audio data exhibit a complex structure due to the superposition of several sound sources and the coexistence of several layers of information. For instance, speech recordings often include concurrent speakers or background noise and they carry information about the speaker identity, the language and the topic of the discussion, the uttered text, the intonation and the acoustic environment. Music recordings also typically consist of several musical instruments or voices and they carry information about the composer, the temporal organization of music, the underlying score, the interpretation of the performer and the acoustic environment.

When I started my PhD in 2001, the separation of the source signals in a given recording was considered as one of the greatest challenges towards successful application to real-world data of audio processing techniques originally designed for single-source data. Fixed or adaptive beamforming techniques for target signal enhancement were already established, but they required a large number of microphones which is rarely available in practice [BW01]. Blind source separation techniques designed for a smaller number of microphones had just started to be applied to audio in, e.g., [MIZ01, BR01, BGB01]. Eleven years later, much progress has been made and source separation has become a mature topic. Thanks in particular to some of the contributions listed in this document, the METISS team has gained a leading reputation in the field, as exemplified by a growing number of technology transfer collaborations aiming to enhance and remix speech and music signals in various use cases.

The use of source separation as a pre-processing step for the description of individual speech or music sources within a mixture raises the additional challenge of efficiently dealing with non-linear distortions over the estimated source signals. Robust methods interfacing source separation, feature extraction and classification have emerged in the last ten years based on the idea of uncertainty propagation. This topic was part of my research program when I joined Inria in 2006 and it is currently undergoing major growth due to the ubiquity of speech applications for hand-held devices [Den11]. Current methods have not yet reached the robustness of the human auditory system, though, and speech or speaker recognition in real-world non-stationary noise environments remains a very challenging problem.

By comparison with the above two challenges, joint processing of the multiple layers of information underlying audio signals has attracted less interest to date. It remains however a

fundamental problem for music processing in particular [AP04, DBC09], where tasks such as polyphonic pitch transcription and chord identification are typically performed independently of each other without accounting for the strong links between pitch and chord information.

My work has been focusing on these three challenges and is based in particular on the theoretical foundations of Bayesian modeling and estimation on the one hand and sparse modeling and convex optimization on the other hand. This document provides an overview of my contributions since the end of my PhD along four axes: Chapter 1 is devoted to the formalization and diagnostic assessment of certain studied problems, Chapter 2 to linear modeling of audio signals and to some associated algorithms, Chapter 3 to variance modeling of audio signals and to some associated algorithms, and Chapter 4 to the description of multisource and multilayer contents. Chapter 5 summarizes the research perspectives arising from this work.

Throughout the document, I have chosen to adopt a rather unconventional writing style using either first-person singular pronouns for the studies conducted as the unique or main researcher or first-person plural pronouns for the studies conducted as supervisor or collaborator.

Chapter 1

Problems, evaluation, and diagnostic assessment

Supervision: Valentin Emiya (postdoc), Gabriel Sargent (PhD student)
Main collaborations: Carl von Ossietzky Universität Oldenburg (Germany), NTT Communication Science Labs (Japan), Queen Mary University of London (United Kingdom)

1.1 Source separation

Audio source separation is the problem of extracting the signals of one or more target sound sources from a given recording. Back in 2006, the assessment and the comparison of different separation algorithms remained challenging because the test signals and (sometimes unknowingly) the definition of the task and the evaluation metrics changed from one author to another. Independent component analysis (ICA)-based techniques either provided estimates of the source signals up to arbitrary filtering [MLS07] or projected them back into different input channels [MN01]. Time frequency masking-based techniques typically outputted a masked version of a single channel of the mixture instead [YR04]. We had already designed in [24] a set of evaluation metrics that were applicable to all algorithms and thereby overcame some of the limitations of previous algorithm-specific metrics [STS99, YR04]. However, these metrics were not attached to a specific task and left to the choice of the user instead.

It is when organizing one of the first evaluation campaigns in the field, namely the 2007 *Stereo Audio Source Separation Evaluation Campaign* (SASSEC) [35], that I proposed a reference methodology for the evaluation of source separation algorithms based on a set of standard tasks, evaluation criteria and performance bounds. This methodology has later been expanded in the series of *Signal Separation Evaluation Campaigns* (SiSEC) which I founded and then co-organized [10, 30, 34]. I also specified the information to be mentioned about the test data so that the results are reproducible [10]. This includes in particular the room *reverberation time*, that is the time taken by the echoes of a sound impulse to decay by 60 decibels (dB).

The following notations are used in the rest of the document. Scalars are denoted by plain letters, vectors by bold lowercase letters, and matrices by bold uppercase letters.

1.1.1 Tasks

Audio mixtures result from the process of simultaneously recording several sound sources and/or mixing multiple recordings using appropriate hardware or software. In any case, the mixing process can be approximated as a linear, time-varying process. Denoting by J and I the number of sources and mixture channels, the observed $I \times 1$ mixture signal $\mathbf{x}(t)$ at time t can be expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1.1}$$

where $\mathbf{c}_j(t)$ is the contribution of the j-th source to the mixture. We call $\mathbf{c}_j(t)$ the *spatial image* of the j-th source. This formulation, which was first proposed in the context of ICA [Car98], is in fact very general and does not rely on any assumption about the sources. In particular, it is not restricted to *point sources* emitting sound from a single point in space, but it is also applicable to spatially *diffuse* sources.

Studies in the blind source separation community have mostly focused on point sources. In this case, the spatial image of each source is the result of the convolution process

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{a}_j(t-\tau, \tau) s_j(t-\tau) \tag{1.2}$$

where $s_j(t)$ is the single-channel source signal emitted by the source and the entries of $\mathbf{a}_j(t, \tau)$ are time-varying *mixing filters*. When the source is immobile, these filters become time-invariant and are denoted as $\mathbf{a}_j(\tau)$. Stacking all mixing filters into a matrix $\mathbf{A}(\tau)$ and all sources into a vector $\mathbf{s}(t)$, this yields the classical formulation of *convolutive* source separation $\mathbf{x}(t) = \mathbf{A} \star \mathbf{s}(t)$, where $\star$ stands for multichannel convolution. This boils down to $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ in the simple case when $\mathbf{A}(\tau)$ is an *instantaneous* mixing matrix $\mathbf{A}$. Using the terminology of linear algebra, the problem is then classically called over-determined, determined or *under-determined* depending whether $J < I$, $J = I$ or $J < I$, respectively [CJ10].

This brief analysis shows that the problem of source separation translates into two different tasks at least depending on the nature of the sources and on the intended application: the estimation of the single-channel source signals $s_j(t)$ or that of their multichannel spatial images $\mathbf{c}_j(t)$ [10, 35].

1.1.2 Evaluation criteria

While the evaluation criteria defined in [24] turned out to be appropriate for the assessment of source signal estimation, new criteria remained to be found for the assessment of source spatial image estimation. I defined such criteria in [10, 35] assuming reference spatial image signals $\mathbf{c}_{j'}(t)$ to be available for all sources, $1 \leq j' \leq J$, and released them as version 3.0 of the BSS Eval toolbox. Because source spatial image estimation does not suffer from any indeterminacy [Car98], it has a unique ideal solution that is $\mathbf{c}_j(t)$ itself. An estimated source spatial image $\widehat{\mathbf{c}}_j(t)$ can be decomposed as

$$\widehat{\mathbf{c}}_j(t) = \mathbf{c}_j(t) + \mathbf{e}_j^{\mathrm{spat}}(t) + \mathbf{e}_j^{\mathrm{interf}}(t) + \mathbf{e}_j^{\mathrm{artif}}(t) \tag{1.3}$$

where $\mathbf{e}_j^{\text{spat}}(t)$, $\mathbf{e}_j^{\text{interf}}(t)$ and $\mathbf{e}_j^{\text{artif}}(t)$ are distortion components representing spatial distortion, interference, and artifacts also known as "musical noise". Similarly to [24], $\mathbf{e}_j^{\text{spat}}(t)$ and $\mathbf{e}_j^{\text{interf}}(t)$ can be expressed as filtered versions of the reference source images and computed by time-invariant least-squares projection of the estimated source image onto the corresponding signal subspaces. The amount of spatial distortion, interference and artifacts is then measured by the *image-to-spatial distortion ratio* (ISR), the *signal-to-interference ratio* (SIR) and the *signal-to-artifacts ratio* (SAR) expressed in dB

$$\text{ISR}_j = 10\log_{10}\frac{\sum_t \|\mathbf{s}_j(t)\|^2}{\sum_t \|\mathbf{e}_j^{\text{spat}}(t)\|^2} \tag{1.4}$$

$$\text{SIR}_j = 10\log_{10}\frac{\sum_t \|\mathbf{s}_j^{\text{img}}(t) + \mathbf{e}_j^{\text{spat}}(t)\|^2}{\sum_t \|\mathbf{e}_j^{\text{interf}}(t)\|^2} \tag{1.5}$$

$$\text{SAR}_j = 10\log_{10}\frac{\sum_t \|\mathbf{s}_j^{\text{img}}(t) + \mathbf{e}_j^{\text{spat}}(t) + \mathbf{e}_j^{\text{interf}}(t)\|^2}{\sum_t \|\mathbf{e}_j^{\text{artif}}(t)\|^2}. \tag{1.6}$$

while the total distortion is measured by the *signal-to-distortion ratio* (SDR)

$$\text{SDR}_j = 10\log_{10}\frac{\sum_t \|\mathbf{s}_j^{\text{img}}(t)\|^2}{\sum_t \|\mathbf{e}_j^{\text{spat}}(t) + \mathbf{e}_j^{\text{interf}}(t) + \mathbf{e}_j^{\text{artif}}(t)\|^2}. \tag{1.7}$$

Although these metrics were shown to correlate with perception reasonably well [FSPZ07], the distortion components obtained by time-invariant least-squares projection are not always perceptually relevant and energy ratios do not account for auditory phenomena such as loudness perception and spectral masking. In order to complement these metrics, we proposed in [11] a specific listening test protocol for the subjective evaluation of audio source separation. The proposed protocol is inspired from the "multiple stimuli with hidden reference and anchor" (MUSHRA) protocol [ITU03] for the assessment of medium and large impairments, with the difference that the original mixture signal is also made available for listening. Listeners are asked to rate the global quality of the test sounds compared to the reference on a scale from 0 to 100, as well as their quality in terms of preservation of the target source, suppression of other sources, and absence of additional artificial noise. As an essential feature of MUSHRA, we developed a set of task-specific low-quality *anchor* sounds to be hidden among the test sounds in order to avoid fluctuation of the scoring scale depending on the test sounds. Using this protocol, we collected the subjective scores given by 20 listeners to 80 sounds, including the outputs of actual source separation algorithms of the SiSEC 2008 campaign.

Building upon these results, we defined four objective evaluation metrics aiming to predict the results of the listening test for any estimated source spatial image signal and released them as a software call PEASS [11]. These metrics exhibit two main differences compared to the SDR, ISR, SIR and SAR. Firstly, we replaced the time-invariant decomposition of the estimation error in [24] by a computationally efficient auditory-motivated time-varying decomposition. The estimated and the reference source spatial image signals are passed through a gammatone filterbank mimicking the frequency-dependent bandwidth of the ear and subsequently windowed into

overlapping time frames whose duration is inversely proportional to the bandwidth. The distortion components in (1.3) are then separately estimated in each subband and each time frame by least-squares projection as above and the fullband distortion components are reconstructed by overlap and add and filterbank inversion. Secondly, we proposed to assess the perceptual salience of each distortion component via the PEMO-Q auditory model [HK06] and to combine the resulting saliences via a neural network trained on the collected subjective scores.

We assessed the performance of the resulting *overall perceptual score* (OPS), *target-related perceptual score* (TPS), *interference-related perceptual score* (IPS) and *artifacts-related perceptual score* (APS) in terms of linear correlation and rank correlation with the subjective scores in a cross-validation experiment. The results displayed in Table 1.1 show that, except for the IPS with respect to the SIR, the subjective relevance of the OPS, TPS and APS is greatly improved with respect to the SDR, ISR and SAR. I further refined these criteria in [43] by optimizing the parameters of PEMO-Q and proposing a different training procedure.

	BSS Eval				PEASS			
	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS
Linear correlation	0.37	-0.14	**0.72**	0.31	**0.61**	**0.46**	0.60	**0.43**
Rank correlation	0.36	-0.07	**0.67**	0.31	**0.55**	**0.44**	0.59	**0.43**

Table 1.1: Linear and rank correlation between BSS Eval or PEASS metrics and the subjective scores given to actual separation algorithms [11].

1.1.3 Performance bounds

The separation performance of a given algorithm depends on several factors: the intrinsic difficulty of separation of the considered mixture, the choice of an underlying model, the constraints on this model (window size, number of parameters, etc) and the choice of an algorithm to estimate the parameters of the model. In order to assess the impact of the former three factors with respect to the latter, I introduced in [22] the concept of *oracle* estimator, that is an estimator of the best SDR possibly achievable by a class of algorithms on a given mixture signal. I designed explicit algorithms to compute oracle estimators for three particular classes of algorithms, namely multichannel time-invariant filtering and single-channel or multichannel *time-frequency masking*. I then implemented them into a toolbox called BSS Oracle and evaluated their performance on various data.

The results for multichannel time-frequency masking are particularly interesting to analyze. This class of algorithms operates as follows [Gri03, YR04, AAG07]. The signals are transformed into the time-frequency domain via the short-time Fourier transform (STFT). In each time-frequency bin (n, f), the convolutive mixing process can be approximated under a narrowband assumption (see Chapter 2) as $\widetilde{\mathbf{x}}(n, f) = \widetilde{\mathbf{A}}(f)\widetilde{\mathbf{s}}(n, f)$ where $\widetilde{\mathbf{x}}(n, f)$ are $\widetilde{\mathbf{s}}(n, f)$ are the STFT coefficients of the mixture and the sources and $\widetilde{\mathbf{A}}(f)$ is the Fourier transform of the mixing filters. Due to the sparsity of audio sources in this domain, it can be assumed that only $J' < J$ sources are actually active in this bin. The SFTF coefficients of these sources are then recovered by pseudo-inversion of the $I \times J'$ matrix consisting of the corresponding columns of

$\widetilde{\mathbf{A}}(f)$, while the STFT coefficients of the other sources are set to zero. Time-domain signals are finally obtained by the inverse STFT. Figure 1.1 shows the resulting oracle SDR for convolutive two-channel three-source mixtures as a function of the STFT window length and the assumed number of active sources J' per time-frequency bin. The results indicate that the performance of popular *binary masking* algorithms [YR04], which assume a single active source per time-frequency bin, is limited to about 12 dB SDR and that it is well worth developing algorithms allowing a greater number of active sources, as we shall present in the following.

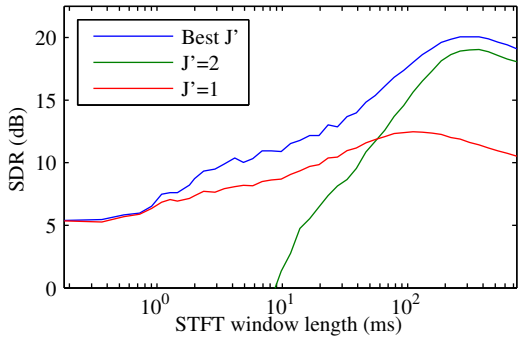


Figure 1.1: Performance of the multichannel time-frequency masking oracle on two-channel convolutive mixtures of three audio sources as a function of the STFT window length and the number of active sources J', assuming that $\widetilde{\mathbf{A}}(f)$ is known [22].

In subsequent studies, we developed oracle estimators for other classes of algorithms, namely time-frequency masking using adaptive time-frequency transforms [77, 79] and single-channel source separation based on Gaussian mixture models (GMMs) of the source spectra [98].

1.1.4 Evaluation campaigns

The proposed tasks, evaluation metrics and oracle estimators have been exploited in the series of SASSEC and SiSEC evaluation campaigns [10, 30], which I founded and then co-organized. Each edition of these campaigns featured 1 to 6 different datasets and attracted on the order of 15 to 30 submissions. These campaigns have had a great impact on the audio source separation community, as they have made it easier to adopt common datasets, measure progress over time, identify established solutions and focus on the remaining challenges. I provided a detailed analysis of these trends in [10].

As an example, let us analyze the evolution of performance over the "Under-determined speech and music mixtures" dataset. Due to the growing of the dataset over the years, we consider the fixed subset of two-channel mixtures of four speech sources and two-channel mixtures of three music sources of SiSEC 2008 and similar mixtures for SASSEC. These mixtures were either mixed instantaneously or recorded in a room with 250 ms reverberation time via a pair of microphones spaced by 5 cm. For each campaign and each mixing condition, the performance of the system leading to the best average SDR is reported in Table 1.2. It can be seen that the separation of instantaneous mixtures is close to be solved, with an average SDR of 14 dB, while

Performance	SASSEC 2007	SiSEC 2008	SiSEC 2010	SiSEC 2011	Binary masking oracle
Instantaneous mixtures					
Method	[80]	[OF10]	[8]	[8]	[22]
SDR (dB)	10.3	14.0	13.4	13.8	10.4
ISR (dB)	19.2	23.3	23.4	23.6	19.4
SIR (dB)	16.0	20.4	20.0	20.7	21.1
SAR (dB)	12.2	15.4	14.9	15.1	11.4
Microphone recordings					
Method	[SAM11]	[EPSG08]	[SAM11]	[NO12]+[80]	[22]
SDR (dB)	1.8	2.6	3.5	3.8	9.2
ISR (dB)	7.0	5.7	8.4	8.2	16.9
SIR (dB)	4.2	2.4	7.0	7.2	18.5
SAR (dB)	6.8	7.3	6.3	7.2	9.9

Table 1.2: Evolution of the average performance of the best source spatial image estimation method on the "Under-determined speech and music mixtures" dataset of SiSEC compared to oracle binary masking.

that of reverberant recordings remains much more difficult, with an average SDR of 4 dB. For instantaneous mixtures, a performance gain of 3 to 4 dB was achieved in 2008 when replacing the sparse decomposition algorithm in [80] (see Section 2.2.1) by the algorithms in [OF10, 8] which exploit the spectral structure of the sources (see Section 3.4). All these algorithms assume multiple active sources per time-frequency bin, which enables them to outperform the binary masking oracle. For convolutive mixtures, steady progress has been made but the best algorithm so far [NO12] still relies on the sparse decomposition algorithm in [80] and room is left for further progress as shown by the oracle.

Table 1.3 presents the best results achieved over the other datasets of SiSEC 2011, which reflect the current state of the art. For each dataset, the algorithm yielding the best average SDR was selected among the algorithms that were able to separate all sources of all mixtures. With an average SDR of 8 dB, the separation of a speech source mixed with real-world background noise appears easier than that of the convolutive mixtures in Table 1.2. Indeed, although the number of background noise sources may be large, the *spectral diversity* between the target and the noise sources is greater so that time-frequency masking is easier. The separation of mixtures of one immobile source and one moving source yields an SDR of 4 dB, which is significantly smaller than that achieved for mixtures of two immobile sources [10]. Finally, the separation of real-world music recordings appears most difficult, with an SDR of 3 dB only. This can be attributed in particular to the fact that most musical instruments are mixed to the center of the stereo space. The sources then lack *spatial diversity* and only their spectral structure can be exploited to discriminate them, as achieved by the algorithm in [8] (see Section 3.4.2).

From these results, it can be concluded that lack of spatial diversity, reverberation, source movements and background noise are the main remaining challenges to be addressed.

Dataset	Number of channels and sources	Method	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)
Mixed speech and real-world domestic noise	$I = 2$ J unknown	[NM11]	8.0	11.0	14.7	12.0
Mixtures of an immobile and a moving source	$I = 2$ $J = 2$	[NO12]	4.3	5.5	12.8	7.0
Professionally produced music recordings	$I = 2$ $J = 2$ to 10	[8]	2.8	6.9	5.7	4.0

Table 1.3: Average performance of the best source spatial image separation method on the other datasets of SiSEC 2011.

1.2 Music structure estimation

On an entirely different topic, the processing of music signals also features a number of ill-defined problems. The notions of music genre, rhythm or music structure are difficult to define in a univocal manner, for instance. We have focused our methodological efforts on the notion of music structure, which refers to the global temporal organization of a piece of music [PMK10]. Popular thinking associates music structure with the notions of chorus and verse. However, these notions have a limited scope and they are more difficult to formalize than it would appear. In the field of music theory, the structure of a piece derives instead from a number of composition rules which have varied over the centuries and still vary today from one music style to another. Finally, in the field of *music information retrieval*, it is typically considered as a subjective concept depending on the listener [BMK06]. This variety of viewpoints prevents the rigorous comparison of different algorithms for automatic music structure estimation, as the existing ground truth annotations differ from one annotator to another.

In [40, 47, 57], we proposed an operational definition of music structure called *semiotic structure* based on axioms from the theory of linguistics known as structuralism [dS16]. This definition is applicable to a variety of music genres and does not require any experience in music theory from the annotator. We posit that conventional music pieces are formed by concatenation of a number of *structural blocks* at a time scale on the order of 15 s, which exhibit three categories of properties: *morphological* properties relating to the internal temporal organization of each block, *paradigmatic* properties relating to the repetition of certain blocks elsewhere in the piece, and *syntagmatic* properties relating to the relationships between neighboring blocks in the piece.

We proposed a *system and contrast* model [40] to characterize the morphology of the blocks, by which a block is assumed to consist of a sequence of (typically) three elements forming a logical *carrier system* completed by a fourth element contrasting with the first three. Paradigmatic analysis is conducted by looking for blocks corresponding to the same carrier system and grouping them into *equivalence classes*, while syntagmatic properties translate into *structural patterns*, i.e., sequences of block classes being more likely than others. We proposed a practical annotation procedure based on joint morphological, syntagmatic and paradigmatic analysis of the piece and a set of annotation conventions. This approach was validated on a database of 20

music pieces and resulted in a concordance of 91% among annotators about the block boundaries [57].

Chapter 2

Linear modeling and associated algorithms

Supervision: Alexis Benichoux (PhD student), Pascal Bado (MSc intern)
Main collaborations: Université Paris 6 - LAM, Supélec - L2S (France), Queen Mary University of London (United Kingdom)

Building upon the definition of the source separation problem in Chapter 1, we now present our contributions towards addressing this and other audio signal processing problems. Following our categorization of source separation algorithms in [28], we divide these contributions into two categories depending on the underlying signal modeling paradigm. The current chapter focuses on *linear modeling*, while Chapter 3 focuses on variance modeling. We especially concentrate on under-determined mixtures, whose number of channels I is strictly smaller than the number of sources J.

The organization of both chapters is similar. After summarizing the general principle of the considered modeling paradigm, we present a baseline model accounting for the *local* time-frequency characteristics of the signals. We then introduce more advanced models accounting respectively for their *spatial* and their *spectral* characteristics. For each model, we briefly describe the associated estimation algorithms.

2.1 General principle

Linear modeling is a general modeling paradigm which consists of representing the signals in a (possibly undercomplete or overcomplete) basis of signals $\mathbf{\Phi}$ and of assuming a certain prior distribution or cost function over their coefficients in this basis [Mal98]. The basis $\mathbf{\Phi}$ may be fixed, selected from a finite library of bases, learned from the signal itself, or defined as the output of a nonlinear parametric function. In the latter case, the resulting model is often called generalized linear model [DGI06]. The challenges associated with linear modeling include designing the prior or the cost over the representation coefficients, computing the representation coefficients for a given signal and selecting or learning an appropriate basis for a family of signals.

2.2 Local sparse modeling

In the context of audio source separation, the chosen basis $\mathbf{\Phi}$ is typically an overcomplete STFT basis with a fixed analysis window. In each time-frequency bin (n, f), the STFT coefficients $\widetilde{\mathbf{x}}(n, f)$ of the mixture signal in time frame n and frequency bin f satisfy

$$\widetilde{\mathbf{x}}(n, f) = \sum_{j=1}^{J} \widetilde{\mathbf{c}}_j(n, f) \tag{2.1}$$

where $\widetilde{\mathbf{c}}_j(n, f)$ are the STFT coefficients of spatial source images. Let us assume that all sources are point sources whose spatial images can be expressed as in (1.2). Under a *narrowband assumption* which is valid with low reverberation, $\widetilde{\mathbf{c}}_j(n, f)$ can be approximated as [MLS07]

$$\widetilde{\mathbf{c}}_j(n, f) = \widetilde{s}_j(n, f)\, \widetilde{\mathbf{a}}_j(f) \tag{2.2}$$

where $\widetilde{\mathbf{a}}_j(f)$ are *steering vectors* representing the frequency response of the mixing filters and $\widetilde{s}_j(n, f)$ are the STFT coefficients of the source signals. Stacking these quantities into a matrix $\widetilde{\mathbf{A}}(f)$ and a vector $\widetilde{\mathbf{s}}(n, f)$, the mixture coefficients then satisfy the linear model $\widetilde{\mathbf{x}}(n, f) = \widetilde{\mathbf{A}}(f)\widetilde{\mathbf{s}}(n, f)$. In the case of an instantaneous mixture, a similar model may also alternatively be obtained using a complete modified discrete cosine transform (MDCT) basis.

In the STFT or the MDCT domain, speech and music signals are *sparse* [YR04]: few coefficients are large and most are close to zero, as can be seen in Figure 2.1. As a consequence, in each time-frequency bin, the observed mixture $\widetilde{\mathbf{x}}(n, f)$ is close to collinear to the steering vector $\widetilde{\mathbf{a}}_j(f)$ of the dominant source j in that bin. Source separation algorithms for under-determined mixtures generally involve two steps [MLS07, CJ10]: in the first step, the steering vectors $\widetilde{\mathbf{a}}_j(f)$ are estimated by, e.g., clustering of $\widetilde{\mathbf{x}}(n, f)$ and, in the second step, the source STFT coefficients $\widetilde{s}_j(n, f)$ are derived and transformed back to the time domain by inverse STFT or MDCT.

We have mostly focused on the second step, assuming a *semi-blind* scenario where the mixing filters or the steering vectors are known. Because of under-determinacy, $\widetilde{\mathbf{A}}(f)$ is non-invertible. The classical time-frequency masking approach assumes that only $J' \leq I$ sources are active in each time-frequency bin and performs pseudo-inversion of the matrix made of the corresponding columns of $\widetilde{\mathbf{A}}(f)$ [Gri03, YR04, AAG07]. An alternative approach is to minimize

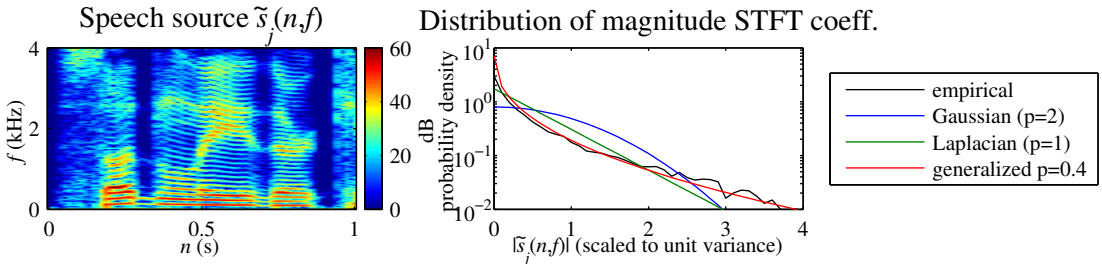


Figure 2.1: Distribution of the magnitude STFT coefficients of a speech source.

the ℓ_1 norm of the source STFT coefficients, which is typically justified as maximum likelihood (ML) estimation of the source STFT coefficients under the assumption that they have a sparse Laplacian distribution [ZPBK01, WKSM07].

2.2.1 Complex-valued ℓ_p norm minimization

In [80], I studied the distribution of magnitude STFT coefficients of audio sources and showed that they are sparser than Laplacian. This distribution can be well approximated by the generalized Gaussian distribution [CCA98]

$$p(|\widetilde{s}_j(n,f)|) = p\frac{\beta^{1/p}}{\Gamma(1/p)}e^{-\beta\,|\widetilde{s}_j(n,f)|^p} \tag{2.3}$$

where the ML value of p is typically on the order of $p \approx 0.4$ as opposed to $p = 1$ for Laplacian or $p = 2$ for Gaussian data, as illustrated in Figure 2.1. ML estimation of the source STFT coefficients then translates into ℓ_p norm minimization of $\widetilde{\mathbf{s}}(n,f)$ under the constraint that $\widetilde{\mathbf{x}}(n,f) = \widetilde{\mathbf{A}}(f)\widetilde{\mathbf{s}}(n,f)$. This problem is more difficult than ℓ_1 norm minimization since it is nonconvex when $p < 1$. Also, the characterization of the solutions of ℓ_p norm minimization for real-valued data does not apply to complex-valued data [WKSM07].

I provided a mathematical characterization of the local minima of complex-valued ℓ_p norm minimization when $J = I + 1$ [80]. These minima are either singular, in which case at least one of the entries of $\widetilde{\mathbf{s}}(n,f)$ is zero, or nonsingular, in which case $\widetilde{\mathbf{s}}(n,f)$ belongs to a bounded set. I then built upon this characterization to derive an algorithm for the estimation of the global minimum and showed experimentally that, for small enough p, this minimum is almost surely singular, i.e., it involves at most I nonzero coefficients. The solution can then be very quickly estimated by selecting the singular local minimum with minimum ℓ_p norm out of J such minima.

Figure 2.3 shows the resulting source separation performance on two-channel mixtures of three speech sources. It turns out that for convolutive mixtures the best separation is achieved for the sparsest setting of p, that is $p \to 0$. For instantaneous mixtures, this setting also provides a good tradeoff between separation performance and computational cost. The discrepancy

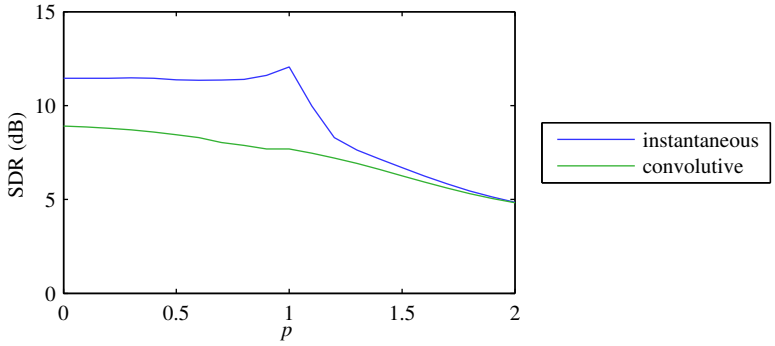


Figure 2.2: Semi-blind separation performance of ℓ_p norm minimization on two-channel instantaneous or convolutive mixtures of three speech sources [80].

between the value of p obtained by ML fitting of the source signals and that yielding the best performance can be explained by the recent analysis in [GCD12] together with the fact that the equality $\widetilde{\mathbf{x}}(n,f) = \widetilde{\mathbf{A}}(f)\widetilde{\mathbf{s}}(n,f)$ does not hold exactly. The algorithm with $p \to 0$ ranked first for the separation of instantaneous mixtures in SiSEC 2007 and was later reused as part of the best approach for under-determined convolutive mixtures in SiSEC 2011 [NO12] (see Table 1.2).

2.2.2 Time-frequency basis selection

Since both time-frequency masking and ℓ_p norm minimization with small p can separate at most I sources per time-frequency bin, the separation performance of these algorithms can be improved by representing the signals in a different basis $\mathbf{\Phi}$ in which they are more disjoint. Automatic selection of the best basis has been widely studied, e.g., in the field of audio coding [ISO05]. However, the best basis for audio coding is not necessarily the one for source separation. In [79], I proposed an algorithm for the selection of the best basis within the dyadic cosine packet library using a disjointness criterion tailored to source separation. Each basis in this library is similar to the MDCT, except that the window length varies over time following a dyadic partition of the time axis. We later proposed another algorithm in [26] by relaxing the dyadic constraint and considering window lengths similar to those used for audio coding.

These algorithms offer an increase of separation performance on the order of 1 dB for instantaneous mixtures outweighed by a considerable increase of computation time. Limited further improvements are to be expected from this direction for the fundamental reason illustrated in Figure 2.3. Although ℓ_p norm minimization can improve upon binary masking by estimating the I predominant sources in each time-frequency bin, these sources are most often incorrectly estimated. In order to address this issue, it appears necessary to move from a local time-frequency model to a more global model accounting for spatial or spectral dependencies between different time-frequency bins.

2.3 Wideband modeling of the mixing filters

It is in that spirit that we pioneered in [16] a new approach to under-determined source separation by replacing the narrowband approximation (2.2) with the exact *wideband* mixing process (1.2) for point sources. Indeed, time-domain filtering induces dependencies between the STFT coefficients of the source spatial images in neighboring time frames (and to a lesser extent in neighboring frequency bins), especially when the mixing filters are longer than the STFT window length. Wideband processing is common for determined mixtures [KB03] but it had not yet been considered for under-determined mixtures due to the difficulty of simultaneously handling time-domain modeling of the mixing filters and STFT-domain modeling of the source signals.

2.3.1 Estimation of the source signals

When the model for the source STFT coefficients corresponds to a convex regularization term such as the ℓ_p norm with $p \geq 1$, this difficulty may be addressed by replacing the equality (1.1)

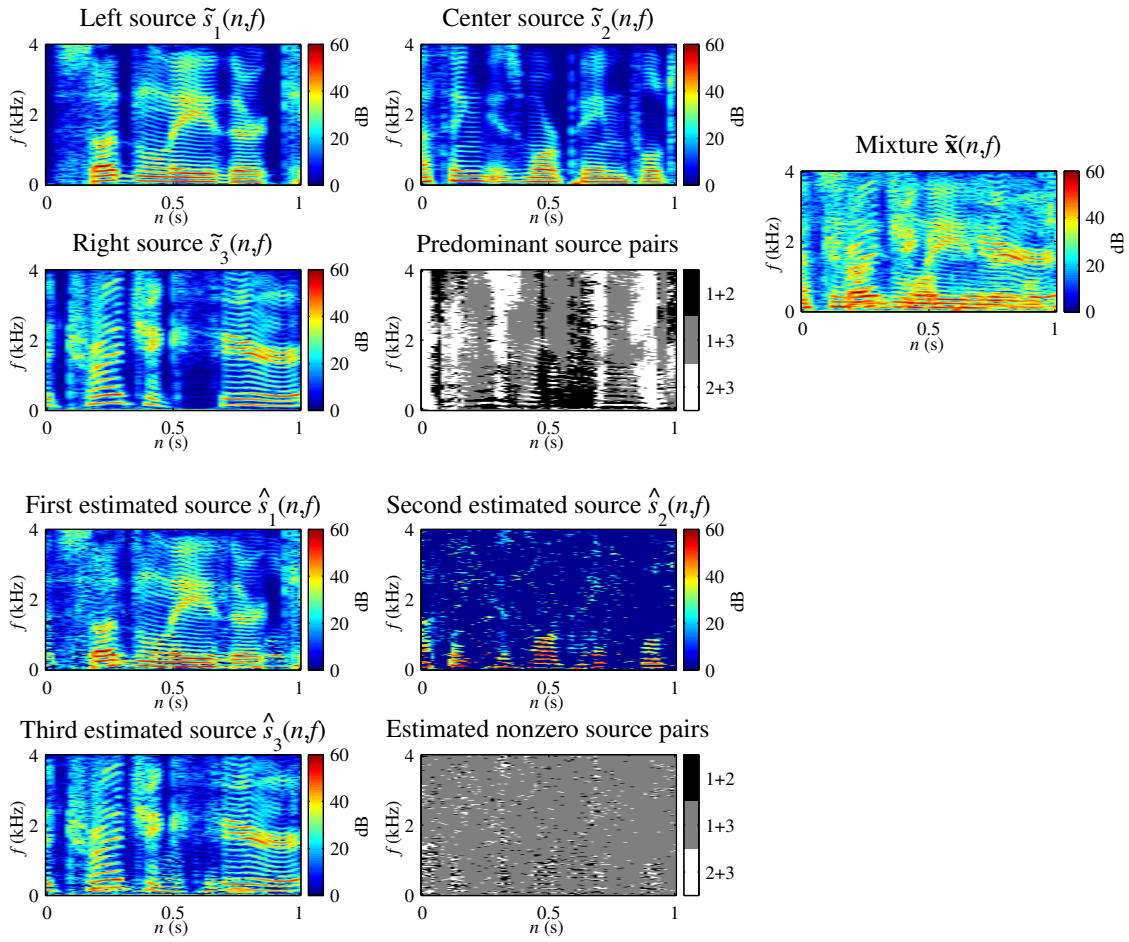


Figure 2.3: Semi-blind separation of a two-channel instantaneous mixture of three speech sources by ℓ_p norm minimization with $p \to 0$ [80].

by a ℓ_2 loss term and using convex optimization techniques such as the fast iterative thresholding algorithm (FISTA) proposed in [BT09]. In the case of ℓ_1 norm regularization, this results in the following minimization problem [16]

$$\min_{\widetilde{\mathbf{s}}} \frac{1}{2} \sum_t \|\mathbf{x}(t) - \mathbf{A} \star \mathrm{iSTFT}(\widetilde{\mathbf{s}})(t)\|_2^2 + \lambda \sum_{n,f} \|\widetilde{\mathbf{s}}(n,f)\|_1. \tag{2.4}$$

The solution of ℓ_1 norm minimization under the constraint (1.1) is attained in the limit when $\lambda \to 0$. We also proposed a variant where the ℓ_1 norm is replaced by the squared $\ell_{1,2}$ mixed norm $\sum_{n,f} \|\widetilde{\mathbf{s}}(n,f)\|_1^2$. While the ℓ_1 norm induces sparsity over the time-frequency plane, $\ell_{1,2}$ norm regularization induces sparsity over the sources within each time-frequency bin only, in a similar way to time-frequency masking [KT08].

The application of FISTA to these two minimization problems is detailed in Algorithm 1.

Algorithm 1: Wideband semi-blind source separation via FISTA. Parenthesized superscripts indicate the iteration number.

Initialization: $\widetilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N \times B}$, $\mathbf{z}^{(0)} = \widetilde{\mathbf{s}}^{(0)}$, $\tau^{(0)} = 1$, $k = 1$.
repeat
  $\mathbf{b}^{(k)} = \mathbf{z}^{(k-1)} - \frac{1}{L} \operatorname{STFT}[\mathbf{A}^* \star (\mathbf{x} - \mathbf{A} \star \operatorname{iSTFT}(\widetilde{\mathbf{s}})^{(k-1)})]$
  switch *regularization term* **do**
    case ℓ_1
      $\widetilde{\mathbf{s}}^{(k)} = \operatorname{prox}_{\frac{\lambda}{L}\|.\|_1}(\mathbf{b}^{(k)})$
    case $\ell_{1,2}$
      $\widetilde{\mathbf{s}}^{(k)} = \operatorname{prox}_{\frac{\lambda}{2L}\|.\|_{1,2}^2}(\mathbf{b}^{(k)})$
  $\tau^{(k)} = \frac{1+\sqrt{1+4\tau^{(k-1)^2}}}{2}$
  $\mathbf{z}^{(k)} = \widetilde{\mathbf{s}}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\widetilde{\mathbf{s}}^{(k)} - \widetilde{\mathbf{s}}^{(k-1)})$
  $k = k + 1$
until *convergence* ;

In each iteration, the current estimate is first linearly updated using a combination of the STFT, the iSTFT and filtering by $\mathbf{A}(\tau)$ and its adjoint $\mathbf{A}^*(\tau) = \mathbf{A}(-\tau)^T$. The result is then subject to nonlinear shrinkage using the proximal operator $\operatorname{prox}_{\mathcal{P}}$ for the consider regularization $\mathcal{P}$. This operator defined as

$$\operatorname{prox}_{\mathcal{P}}(\mathbf{z}) = \frac{1}{2} \arg\min_{\mathbf{u}} \|\mathbf{z} - \mathbf{u}\|_2^2 + \mathcal{P}(\mathbf{u}) \tag{2.5}$$

amounts to entrywise soft thresholding with a fixed threshold for the ℓ_1 norm and with a data-dependent threshold for the squared $\ell_{1,2}$ norm. The tradeoff parameter λ was initially set to 10^{-1} and decreased over the iterations down to 10^{-8}.

Separation results are reported in Table 2.1 for two-channel convolutive mixtures of four speech sources. Both wideband algorithms improve the SDR by as much as 4 dB compared to narrowband binary masking and ℓ_1 norm minimization. The SIR and SAR are also improved and wideband ℓ_1 norm minimization performs slightly better than wideband $\ell_{1,2}$ norm minimization. We also studied the robustness of these algorithms to inaccurate estimation of the mixing filters $\mathbf{A}(\tau)$ by truncating them or perturbating them with exponentially decaying Gaussian noise [16].

Algorithm	narrowband		wideband	
	Binary	ℓ_1 min.	$\ell_{1,2}$ min.	ℓ_1 min.
SDR (dB)	3.4	2.8	7.2	**7.6**
SIR (dB)	10.0	6.4	13.9	**14.0**
SAR (dB)	5.1	6.5	8.5	**9.1**

Table 2.1: Semi-blind separation performance of wideband vs. narrowband algorithms on two-channel convolutive mixtures of four speech sources [16].

2.3.2 Estimation of the mixing filters

In order to move towards a *blind* scenario where the mixing filters are unknown, we recently extended the wideband approach to the reverse problem of estimating the mixing filters when the source signals are known [46]. While sparse regularization of time-domain filters had already been used for deconvolution [LCKL07, NBJ10], we extended this approach to multiple sources and introduced new regularization terms accounting both for the sparsity and for the decreasing temporal envelope of acoustic room impulse responses. This algorithm allowed us to propose a competitive way of recording large sets of room impulse responses compared to traditional recording techniques based on sine sweeps.

2.4 Harmonic sinusoidal modeling of the source signals

Independently of our work on wideband modeling, we also sought ways of better modeling the source signals. A natural approach within the linear modeling paradigm is to learn a signal-dependent basis from the time frames of the observed signal, which we applied to source separation in [18]. This approach is however not fully satisfactory, since it does not account for *translation invariance*, i.e., the fact that all time-shifted versions of every atom of the basis should also be part of the basis, and it is prone to overfitting for short signals.

In the context of my postdoctoral research assignment which was to estimate, compress and manipulate parametric sound objects, I focused on harmonic sinusoidal modeling instead. In a given time frame n, the single-source signal $s(t)$ windowed by $w_n(t)$ can be expressed as a sum of periodic sound objects indexed by p and an aperiodic residual $e_n(t)$

$$w_n(t)s(t) = w_n(t)\sum_p \sum_{m=1}^{M_p} a_{npm}\cos(2\pi m f_{np} t + \phi_{npm}) + e_n(t) \tag{2.6}$$

where f_{np} is the *fundamental frequency* of the p-th periodic object and (a_{npm}, ϕ_{npm}) are the amplitude and the phase of its m-th harmonic partial [Ros03, DGI06]. The parameters of this model are typically estimated either by extracting sinusoidal tracks [MQ86] and subsequently grouping them or by estimating the fundamental frequencies and deriving the amplitudes and phases of their harmonics. A number of multiple pitch estimation techniques have been proposed for the latter goal, ranging from spectral peak clustering [RG04, PI08] and harmonic sum [Kla06] to neural networks [Mar04] and support vector machines [PE07].

2.4.1 Bayesian estimation

Motivated by the assigned parametric coding application, I pursued the Bayesian approach studied in [Ros03, DGI06] instead, which can be seen as a form of analysis by synthesis. This approach consists of setting probabilistic priors over the variables and estimating them in the maximum a posteriori (MAP) sense. One difficulty is that of *model selection*: due to their higher number of parameters, the joint MAP criterion peaks at submultiples of the fundamental frequencies instead of the actual fundamental frequencies, which results in so-called octave

errors. This problem is aggravated by the choice of priors which do not penalize missing harmonic partials. In [20], I suggested to estimate the MAP fundamental frequencies in a first step by *marginalizing* over, i.e., integrating out, the amplitudes and phases of the partials and to derive the MAP amplitudes and phases conditionally to the estimated fundamental frequencies in a second step. I defined an auditory-weighted Gaussian prior for the aperiodic residual and log-Gaussian priors for the fundamental frequencies and the harmonic amplitudes, so as to penalize missing partials. I then proposed an efficient algorithm for the computation of the high-dimensional marginalization integral based on automatic identification of those variables which are independent a posteriori. Compared to conventional marginalization techniques, this algorithm was shown to be more accurate than Laplace approximation [CH96] and faster than Markov chain Monte Carlo (MCMC) [CR05]. Also, it is not specific to harmonic models and can be seen as a fundamental tool for marginalization in other contexts than audio.

Using this approach for harmonic sinusoidal modeling, I developed a parametric *sound object coder* for music in [21]. I added a Markov prior over the fundamental frequencies so as to extract continuous sequences of frequencies. I then developed a specific vector quantization technique based on adaptive interpolation of the amplitudes of the harmonic partials over a set of signal-dependent temporal and frequency breakpoints. Figure 2.4 depicts the resulting compression performance. A huge bitrate reduction is achieved compared to conventional transform or sinusoidal coders, with good quality down to 8 kbit/s and fair quality at 2 kbit/s. Application to source separation [36] and manipulations of individual sound objects, such as pitch shifting, time stretching or spectral envelope modification, were also investigated.

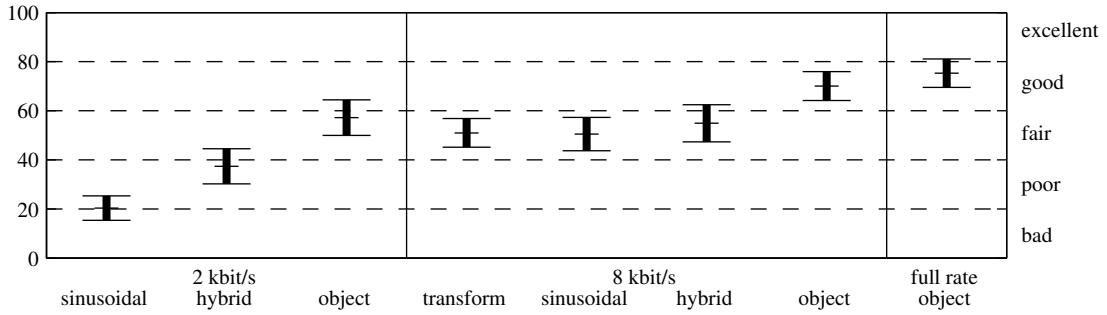


Figure 2.4: Subjective comparison of the proposed sound object coder with baseline coders via MUSHRA [ITU03]. Lame[2]was used for transform coding. The sinusoidal and hybrid coders were implemented in a similar way as the MPEG-4 sinusoidal coding (SSC) and harmonic and individual lines plus noise (HILN) standards [ISO05]. Bars indicate 95% confidence intervals over 10 test items and 7 subjects [21].

2.4.2 Greedy estimation

In a different application context for which high resynthesis quality was not required, we explored the use of a much faster but less accurate greedy algorithm for harmonic sinusoidal mod-

[2]http://lame.sourceforge.net/

eling [19]. Inspired from the harmonic matching pursuit (MP) algorithm introduced in [GB03], our algorithm extracts harmonic sound atoms one by one. However, contrary to [GB03], the amplitudes of the harmonic partials of each atom are not determined from the observed signal but trained on a separate set of isolated note signals of several musical instruments, which reduces octave errors. In addition, constraints over the temporal structure are employed in order to extract sequences of atoms with similar fundamental frequency corresponding to musical notes. We applied this algorithm to multiple pitch estimation and polyphonic musical instrument identification [19].

Chapter 3

Variance modeling and associated algorithms

Supervision: Alexey Ozerov, Laurent Simon, Joachim Thiemann (postdocs), Ngoc Duong, Nobutaka Ito (PhD students), Charles Blandin (MSc intern)
Main collaborations: Télécom ParisTech - LTCI (France), University of Tokyo, NTT Communication Science Labs (Japan)

While linear modeling of the mixture signal or the source signals appears suitable for point sources and periodic sounds, respectively, it is not directly applicable to diffuse sources or random sounds. Also, wideband modeling of the mixing filters or harmonic sinusoidal modeling of the sources incur a large computational cost. The alternative paradigm of *variance modeling* was designed to address these limitations.

This chapter follows a similar structure to the previous one. After summarizing the general principle, we introduce a local variance model and proceed to more advanced spatial and spectral models. Whenever possible, we compare the performance of the proposed models to their linear counterparts.

3.1 General principle

Many speech or music processing techniques do not operate on time-domain signals but on their short-term power spectra, discarding their phase [RJ93, KD06]. Such *phase invariant* processing relies on the observation that the phase spectrum bears less information than the power spectrum for certain applications and it is sometimes justified by the fact that the human auditory system is phase invariant to a certain extent. The fit between the observed spectrum and the model spectrum may be assessed in terms of Itakura-Saito (IS) divergence, Kullback-Leibler (KL) divergence or other divergences. In statistical terms, this is equivalent to assuming that the STFT coefficients of the observed signal have been drawn from a zero-mean phase-invariant distribution whose variance is equal to the model spectrum. Similarly, in the field of source separation, it has been shown that the samples of each source signal may be modeled

as drawn either from identical sparse distributions or from a Gaussian distribution with sample-dependent variance. Both models result in sparsely distributed samples and provide alternative routes to source separation [Car01]. The latter model has been exploited for under-determined audio source separation in [FC05, OF10] in combination with narrowband approximation of the mixing process.

I generalized this idea to non-point sources by proposing to model the multichannel source spatial images $\widetilde{\mathbf{c}}_j(n,f)$ instead of the single-channel source signals by a zero-mean distribution invariant to global phase rotation in each time-frequency bin [28]. The complex-valued Gaussian distribution is particularly suitable since it results in a closed-form expression for the likelihood of the mixture. The distribution of $\widetilde{\mathbf{c}}_j(n,f)$ is given by

$$p(\widetilde{\mathbf{c}}_j(n,f)) = \frac{1}{\det(\pi \mathbf{R}_{\mathbf{c}_j}(n,f))} e^{-\widetilde{\mathbf{c}}_j(n,f)^H \mathbf{R}_{\mathbf{c}_j}^{-1}(n,f) \widetilde{\mathbf{c}}_j(n,f)} \tag{3.1}$$

where the covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n,f)$ can be factored as the product of a scalar *spectral variance* $v_j(n,f)$ encoding its power spectrum and a *spatial covariance matrix* $\mathbf{\Sigma}_j(f)$:

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\mathbf{\Sigma}_j(f). \tag{3.2}$$

Assuming that the sources are independent conditionally to their covariance matrices, the STFT coefficients of the mixture $\widetilde{\mathbf{x}}(n,f)$ also follow a zero-mean Gaussian distribution with covariance matrix

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} \mathbf{R}_{\mathbf{c}_j}(n,f). \tag{3.3}$$

In addition, we proposed to exploit the local zero-mean *empirical covariance matrix* $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ of the mixture in the context of under-determined source separation. This matrix, which can be computed by local averaging of $\widetilde{\mathbf{x}}(n,f)\widetilde{\mathbf{x}}(n,f)^H$ over the time-frequency plane, can be seen as a form of quadratic time-frequency representation. The parameters $\boldsymbol{\theta} = \{v_j(n,f), \mathbf{\Sigma}_j(f)\}_{jnf}$ may be fit to $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ instead of $\widetilde{\mathbf{x}}(n,f)$ via the log-likelihood

$$\log p(\widehat{\mathbf{R}}_{\mathbf{x}}|\boldsymbol{\theta}) = \sum_{n,f} -\log\det(\pi \mathbf{R}_{\mathbf{x}}(n,f)) - \mathrm{tr}(\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)) \tag{3.4}$$

which accounts for the local correlation between the channels of the mixture [59, 75]. Figure 3.1 illustrates the benefit of this extra information for the discrimination of two possible solutions to a source separation problem. Whereas the observation of the mixture STFT coefficients does not suffice to discriminate these solutions without additional sparsity assumptions, the observation of the mixture covariance matrix suffices. If the data have been generated by the "green" solution in which one source predominates, then the two channels of the mixture are strongly correlated. Conversely, if the data have been generated by the "red" solution in which two sources contribute to a similar extent, then the two channels of the mixture are weakly correlated.

Once the model parameters have been estimated, the STFT coefficients of the source spatial images are derived via *multichannel Wiener filtering* $\widehat{\mathbf{c}}_j(n,f) = \mathbf{W}_j(n,f)\mathbf{x}(n,f)$ with

$$\mathbf{W}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{3.5}$$

and transformed back to the time domain via the inverse STFT.

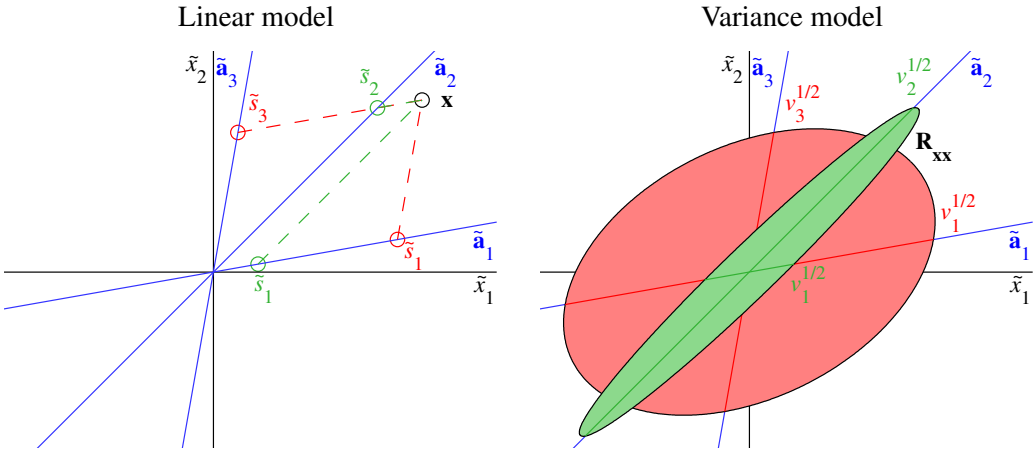


Figure 3.1: Theoretical comparison of linear modeling and variance modeling for the separation of a two-channel mixture of three sources in a given time-frequency bin under the narrowband approximation. For the ease of representation, time-frequency indices are omitted and all quantities are assumed to be real-valued.

3.2 Local Gaussian modeling

The simplest approach within the above Gaussian modeling paradigm is to assume that the spectral variances $v_j(n,f)$ are independent between sources and between time-frequency bins. We termed this approach local Gaussian modeling. Figure 3.2 shows the resulting separation performance for the two-channel instantaneous mixture of three speech sources previously considered in Figure 2.3, under the constraint that at most two sources are nonzero in each time-frequency

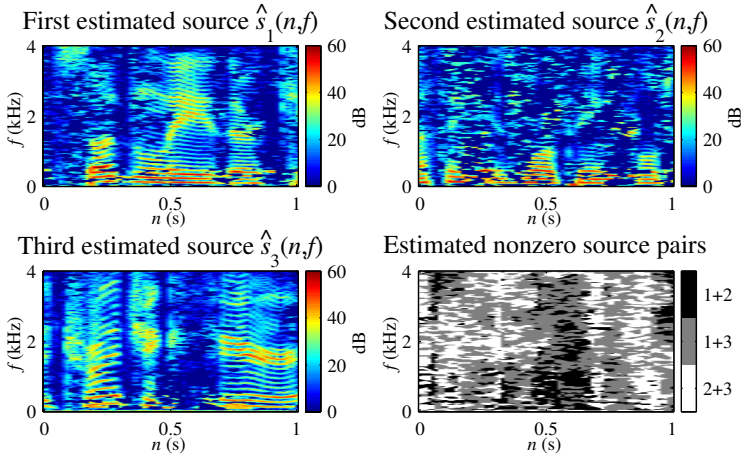


Figure 3.2: Separation of the two-channel instantaneous mixture of three speech sources in Figure 2.3 by local Gaussian modeling [75].

bin [75]. Although not perfect, local Gaussian modeling does not suffer from the limitation of ℓ_p norm minimization discussed in Section 2.2.2 and it is able to correctly estimate the predominant pair of sources in many time-frequency bins from the observation of the mixture covariance matrix.

3.2.1 ML spatial covariance estimation

In [15], we introduced four possible parameterizations of $\mathbf{\Sigma}_j(f)$. The narrowband approximation (2.2) is here equivalent to assuming that the channels of the source spatial images are perfectly correlated with each other and that $\mathbf{\Sigma}_j(f)$ is a *rank-1* matrix equal to $\widetilde{\mathbf{a}}_j(f)\widetilde{\mathbf{a}}_j(f)^H$. For diffuse or reverberated sources, the channels of the source spatial images are more weakly correlated and we proposed to model $\mathbf{\Sigma}_j(f)$ as an unconstrained *full-rank* covariance matrix instead. The principal vector of $\mathbf{\Sigma}_j(f)$ then points to the apparent direction of sound, while the ratio between its largest and smallest eigenvalues is related to the ratio of direct to reverberant sound power. The parameters $v_j(n,f)$ and $\mathbf{\Sigma}_j(f)$ may be estimated in the ML sense by the following iterative algorithm [15] based on Expectation-Maximization (EM) [MK97]: in the E-step, the zero-mean covariance matrices of the estimated source images are computed as

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \tag{3.6}$$

and in the M-step, the parameters are updated as

$$v_j(n,f) = \frac{1}{I}\,\mathrm{tr}(\mathbf{\Sigma}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{3.7}$$

$$\mathbf{\Sigma}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)} \tag{3.8}$$

where $\mathbf{R}_{\mathbf{c}_j}(n,f)$, $\mathbf{R}_{\mathbf{x}}(n,f)$ and $\mathbf{W}_j(n,f)$ are defined in (3.2), (3.3) and (3.5) respectively, $\mathbf{I}$ is the $I \times I$ identity matrix and N the number of time frames.

In semi-blind conditions when $\mathbf{\Sigma}_j(f)$ is known, this algorithm increased the SDR on the order of 2 dB compared to binary masking on the two-channel convolutive mixtures of three to six sources in [16], that is 30% to 60% of the increase observed via wideband linear modeling. This improvement compared to binary masking is significant, considering the fact that the model involves only one additional parameter per frequency bin and that it operates locally in the time-frequency plane, avoiding the computationally expensive convolution operations of wideband modeling. It is in fact comparable to the improvement observed by wideband modeling when truncating the length of the mixing filters to that of the STFT window [16], a comparison that may be more relevant given the difficulty of blindly estimating long mixing filters.

Figure 3.3 displays the separation results in blind conditions. We used a hierarchical clustering technique inspired from [WKSM07] to initialize the spatial covariance matrices and the technique in [SAMM07] to permute the estimated source images to the same order across all frequency bins. An SDR improvement on the order of 1 to 2 dB is still observed with respect to binary masking and ℓ_1 norm minimization for realistic reverberation times on the order of 130 ms or more. Compared to ℓ_1 norm minimization, the SIR and the SAR are both improved

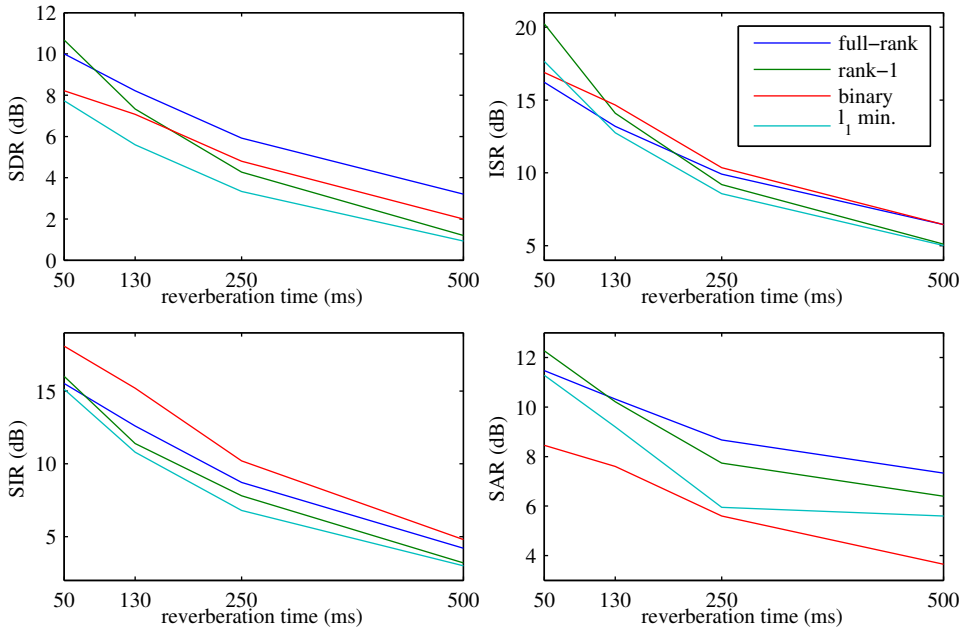


Figure 3.3: Experimental comparison of full-rank local Gaussian modeling, rank-1 local Gaussian modeling, binary masking and ℓ_1-norm minimization for the separation of two-channel convolutive mixtures of three speech sources as a function of the reverberation time [15].

while, for binary masking, this translates into a large increase of the SAR and a smaller decrease of the SIR.

The proposed local Gaussian model and EM algorithm have recently been adopted by other authors for source separation [AN11] and for other applications such as acoustic echo reduction [TH11] and dereverberation [TKT+12]. In [8], we also proposed a variant of this EM algorithm that is able to handle spatial covariance matrices of any rank R by expressing them as $\mathbf{\Sigma}_j(f) = \widetilde{\mathbf{A}}_j(f)\widetilde{\mathbf{A}}_j(f)^H$ and by updating $\widetilde{\mathbf{A}}_j(f)$ instead of $\mathbf{\Sigma}_j(f)$, where $\widetilde{\mathbf{A}}_j(f)$ is an $I \times R$ matrix.

3.2.2 Alternative time-frequency representations

Similarly to local sparse modeling, the separation performance of local Gaussian modeling may be improved by choosing the time-frequency representation such that the sources are maximally disjoint. In [59], we achieved an SDR improvement of 0.5 dB while halving the computational cost by using a fixed quadratic time-frequency representation on a nonlinear auditory-motivated frequency scale, which would not have been possible in the case of linear modeling.

3.3 Modeling of the spatial covariance matrices

Although we have validated its benefit in blind conditions, ML estimation of the source spatial covariance matrices is not fully satisfactory. Firstly, it is prone to overfitting in the frequency

bins where the corresponding sources are inactive. Secondly, as other convolutive blind source separation techniques based on a local model, it requires an additional permutation step in order to align the order of the estimated source images across all frequency bins[MLS07].

3.3.1 MAP spatial covariance estimation

In [2, 50], we studied these issues by investigating various possible probabilistic priors over the source spatial covariance matrices and by estimating them in the MAP sense, assuming that the sources directions of arrival (DOAs) are known. An SDR improvement on the order of 1 dB was obtained compared to ML estimation using an inverse-Wishart prior [MK00]. This prior has the effect of modifying the M-step of the EM algorithm as

$$\mathbf{\Sigma}_j(f) = \frac{1}{\gamma(m+I)+N}\left(\gamma\mathbf{\Psi}_j(f) + \sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)}\right) \tag{3.9}$$

where

$$\frac{\mathbf{\Psi}_j(f)}{m-I} = \mathbf{d}_j(f)\mathbf{d}_j(f)^H + \sigma^2_{\mathrm{rev}}\mathbf{\Omega}(f) \tag{3.10}$$

is the mean of the prior, m its number of degrees of freedom, and γ a tradeoff parameter. The expression (3.10) corresponds to a direct-plus-diffuse model, where $\mathbf{d}_j(f)$ is the anechoic steering vector corresponding to the source DOA, σ^2_{rev} is the intensity of echoes and reverberation and $\mathbf{\Omega}(f)$ is the spatial covariance matrix of a spherically diffuse sound field given by the theory of room acoustics [GRT03].

In order to move towards a blind setting where the source DOAs are unknown, we also conducted in [7] a large-scale evaluation of existing algorithms for multiple source localization that complements the one in [BOS08]. In this context, we proposed a frequency-weighted variant of conventional beamforming-based localization algorithms [BW01] that increases the localization accuracy in the case of closely spaced microphones.

3.3.2 Subspace constraints

In parallel to this study on probabilistic priors, we designed a family of models specifically for diffuse sources based on deterministic subspace constraints [92, 95]. The matrix $\mathbf{\Sigma}_j(f)$ is represented as the sparse combination of a number of basis matrices. In the case of a mixture of one point source and one diffuse source, the combination coefficients may be estimated by convex optimization algorithms for matrix completion [SJ03] or for trace norm minimization [TY10], which is a matrix generalization of ℓ_1 norm minimization. We modified these algorithms in order to deal with complex-valued Hermitian positive-semidefinite data. We successfully applied this approach to the denoising of speech signals and to the localization of multiple sources in noisy environments.

3.4 Factorization-based modeling of the short-term power spectra

Moving on to the modeling of the source spectral characteristics, the phase invariance property of variance modeling opens up many possibilities such as the handling of random noise sounds.

It also enables the parameterization of the source spectra with a fixed, smaller number of parameters, which reduces the risk of overfitting and the need for model selection without completely eliminating them though.

A popular model for single-channel data which I extended to multichannel data in my PhD [23] relies on nonnegative matrix factorization (NMF) [LS01, Vir06, OF10]. The spectral variance $v_j(n,f)$ of the sources is represented as the sum of K *basis spectra* $w_{jk}(f)$ multiplied by time-varying *activation coefficients* $h_{jk}(n)$

$$v_j(n,f) = \sum_{k=1}^{K} w_{jk}(f) h_{jk}(n). \tag{3.11}$$

Different basis spectra may represent different phones in the case of speech or different notes in the case of music. Codebook models, whereby the source spectrum is selected within a finite set of spectra in each time frame, or scaled variants thereof have also been employed [BBG06].

3.4.1 Harmonicity constraints

Although it improves upon the local Gaussian model, NMF lacks flexibility. The basis spectra may be either fixed using training data, in which case they may badly fit the test data, or adapted to the test data in the ML sense [OF10, 56], in which case a significant risk of overfitting remains. Using an idea initially proposed outside of the context of NMF in [VK02], I showed in [17, 78] how to deterministically enforce harmonicity of the basis spectra by representing them as the sum of a few *narrowband harmonic spectra* $n_{jkm}(f)$ associated with the same fundamental frequency multiplied by *spectral envelope* coefficients e_{jkm}

$$w_{jk}(f) = \sum_{m=1}^{M_k} n_{jkm}(f) e_{jkm}. \tag{3.12}$$

The spectra $n_{jkm}(f)$ are fixed and distributed over a fixed fundamental frequency scale (several possible definitions were investigated in [17]), while only the coefficients e_{jkm} are adapted to the test data, which greatly reduces the dimension of the model. This model is illustrated in Figure 3.4. The parameters $h_{jk}(n)$ and e_{jkm} can be estimated in the ML sense via a multiplicative update algorithm similar to those in [LS01], whose convergence we analyzed in [13].

I initially exploited this harmonic NMF algorithm for the task of multiple pitch estimation in polyphonic music signals, using simple thresholding of the activation coefficients $h_{jk}(n)$ to detect the active notes in each time frame [17]. The performance of this and other algorithms is reported in Table 3.1 for woodwind music. An estimated note is deemed to be correct if its pitch is within one quarter-note of the ground truth and performance in assessed in terms of the classical F-measure for information retrieval [vR79]. On average, harmonic NMF increases the F-measure by 3% absolute compared to unconstrained NMF and by 4% absolute compared to the harmonic sum algorithm in [Kla06], but it performs 2% worse than the Bayesian harmonic sinusoidal estimation algorithm introduced in Section 2.4.1. It must be however be reminded that the latter algorithm requires training on isolated notes from each instrument, while harmonic

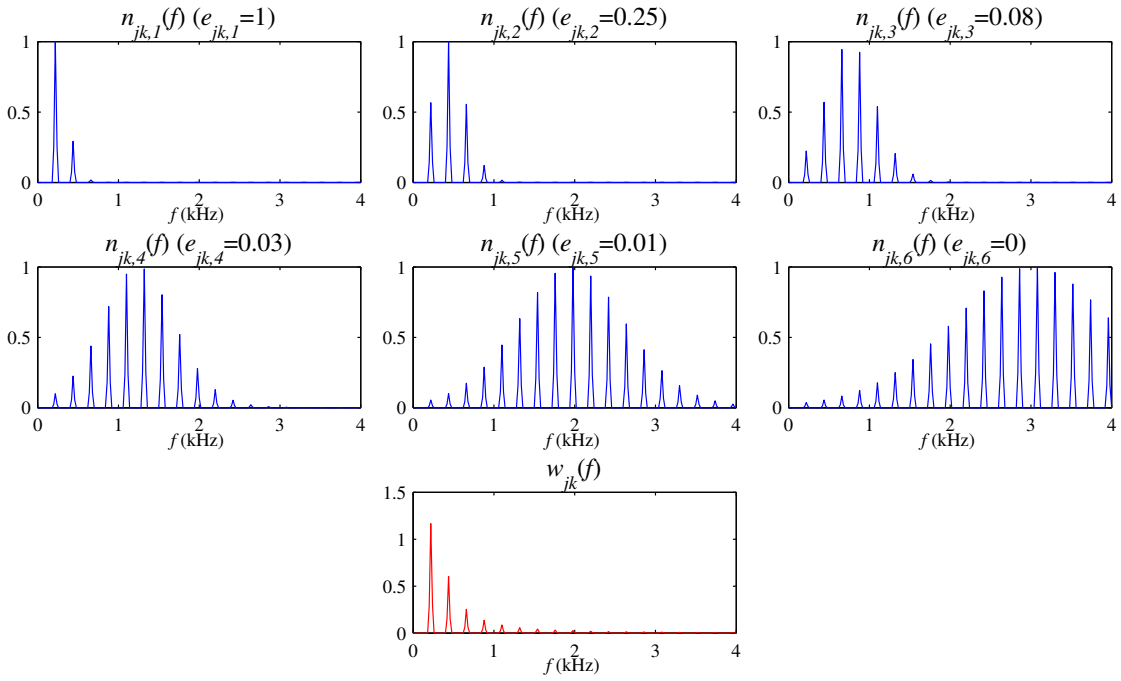


Figure 3.4: Representation of a basis spectrum with a fundamental frequency of 220 Hz (bottom) as the weighted sum of six narrowband harmonic spectra (top).

Algorithm	Number of instruments 2	3	4	5	Average
Bayesian [20]	**89.3**	**70.9**	63.1	56.5	**70.0**
Harmonic NMF [17]	76.5	64.7	**67.5**	**62.5**	67.8
Unconstrained NMF	79.9	56.3	62.1	61.9	65.1
Harmonic sum [Kla06]	73.4	59.1	63.5	59.9	64.0
Spectral peak clustering [PI08]	27.8	24.7	33.5	34.0	30.0

Table 3.1: F-measure (%) achieved by the proposed Bayesian harmonic sinusoidal estimation and harmonic NMF algorithms compared to other algorithms for multiple pitch estimation on woodwind data [17, 20].

NMF is unsupervised and applicable to any instrument. Also, the computational load of NMF is also several orders of magnitude lower than that of Bayesian estimation.

The proposed harmonic NMF algorithm ranked second after the entry of [RK05] on average over all data for the note tracking subtask of the "Multiple fundamental frequency estimation & tracking" task of the 2007 *Music Information Retrieval Evaluation eXchange* (MIREX) [78].

3.4.2 Flexible spectral model

After addressing harmonicity constraints, we pursued the goal of constraining the parameters of NMF. We defined narrowband "noise" spectra to complement the harmonic spectra [12] and set temporal smoothness priors over the activation coefficients [14]. Other authors merged harmonic NMF with the *excitation-filter* model of speech production for melody tracking and source separation instead [FCC08, DRDF10]. In [8], we generalized these and other ideas into a multilevel NMF model. The observed spectrum is assumed to be the product of an excitation spectrum and a filter spectrum; the excitation and filter spectra are then separately expressed as the sum of basis spectra multiplied by time-varying activation coefficients; finally, the basis spectra are decomposed as the sum of narrowband (not necessarily harmonic) spectral patterns multiplied by spectral envelope coefficients, while the series of activation coefficients are represented as the sum of time-localized (not necessarily smooth) patterns multiplied by temporal envelope coefficients. Overall, this results in a product of eight nonnegative matrix variables in addition to the spatial covariance matrices, which may be either fixed or jointly estimated in the ML sense via an algorithm combining multiplicative updates and EM updates.

We systematically studied the source separation performance achieved by this algorithm over speech and music mixtures by either fixing or adapting certain variables [8]. The results showed that the best performance is often achieved when both the basis spectra and the activation coefficients are constrained. Furthermore, this algorithm was one of the few algorithms submitted to the "Professionally produced music recordings" task of SiSEC 2011 which was able to separate all sources and it performed best among those algorithms (see Table 1.3).

More generally speaking, this algorithm is quite flexible, in the sense that it encompasses many existing algorithms as shown in [8] and that it allows quick development of new algorithms by incorporating the prior knowledge available about the sources at hand. As a matter of fact, we implemented it as a toolbox called FASST, which has become the basis for a majority of research developments and industrial transfers in the field of source separation within the METISS team. For instance, we recently used it in the context of online source separation [42].

3.5 Artifact reduction

As a complement to all the algorithms presented in this section and motivated by ongoing industrial transfers requiring high audio quality, we started investigating post-processing techniques for the reduction of source separation artifacts. The reduction of artifacts is typically achieved at the cost of increased interference. Many techniques have been studied in the context of single-channel or multichannel denoising [EM84, Coh04, DM05, CBHD06, YMB08], which we adapted and tested in the context of under-determined source separation [67]. The best tradeoff between SAR increase and SIR decrease was achieved by temporal smoothing of the estimated source covariance matrices $\mathbf{R}_{\mathbf{c}_j}(n, f)$ before the derivation of the Wiener filter. In another series of studies, we introduced a new method for the design of the Wiener filter which accounts for the overcompleteness of the STFT, so as to favor smoother estimates of the source STFT coefficients and to ensure that these estimates actually correspond to the STFTs of time-domain signals [3, 63]. This method led to the filing of a patent [115].

Chapter 4

Description of multisource and multilayer contents

Supervision: Kamil Adiloğlu, Stanisław Raczyński (postdocs), Gabriel Sargent (PhD student), Christopher Sutton, Antoine Movschin, Ricardo Scholz, Christophe Hauser (MSc interns)
Main collaborations: IRCAM - STMS (France), University of Tokyo (Japan)

Beyond the low-level or mid-level signal processing techniques for source separation and multiple pitch estimation considered in the two previous chapters, we recently started investigating the application of these techniques to higher-level content description tasks in multisource conditions. In the same vein, we began conducting research on symbolic multilayer "language" models of music. The current chapter is devoted to these more exploratory studies.

4.1 Towards robust description of multisource contents

While the robustness of speech recognition systems has greatly increased in the last few years, accurate speech recognition remains a challenging task in real-world nonstationary noise environments [DA08]. Errors in the transcription output may be problematical for certain applications, e.g., spoken language understanding for handheld personal assistants, and prohibitive for others, e.g., dictation. The task of identifying the singer or the individual musical instruments within a polyphonic music recording is conceptually similar [MV07, FGKO10]. These tasks are typically addressed in two steps: in the first step, the input signal is transformed into a sequence of *feature vectors* such as mel frequency cepstral coefficients (MFCCs), and in the second step ML classification or decoding is performed based on *acoustic models* of the classes such as GMMs or hidden Markov models (HMMs).

In addition to the design of more robust features, three categories of approaches have been proposed to compensate for the effect of background noise or competing sound sources on a given set of features [Den11]. Front-end approaches, which employ denoising or source separation as a pre-processing step, often yield limited improvement because the distortions introduced over the features reduce or even outweigh the effect of noise reduction. Back-end approaches,

which modify the parameters of the acoustic models by applying static linear transformations or by training them from noisy data, perform better but they require significant computational power or training data matching the noise conditions of the test data. Hybrid approaches coupling front-end and back-end compensation appear most promising.

Within the last category, the emerging paradigm of *uncertainty propagation* offers a robust way of integrating source separation, feature extraction and classification [AK11]. Its principle is to estimate the uncertainty or equivalently the confidence about the separated source signals and to propagate it through the subsequent processing steps. This uncertainty can be encoded via the mean and covariance matrix of a multivariate Gaussian distribution representing the posterior distribution of the signals or the features. Efficient techniques exist to propagate the uncertainty from the source signals to the features [AK11] and to decode acoustic models from uncertain features [DDA05]. We hence focus on the remaining challenges of estimating the uncertainty about the source signals and training acoustic models from uncertain features.

4.1.1 Bayesian uncertainty estimation

Regarding the initial estimation of the uncertainty about the source signals, a heuristic approach is to assume that the uncertainty in a given time-frequency bin is proportional to the squared difference between the separated sources and the mixture [DNW09, KAHO10]. A more principled approach is to consider the uncertainty stemming from the Wiener filter [AK11]

$$p(\mathbf{c}|\mathbf{x}) \approx \prod_{j,n,f} p(\mathbf{c}_j(n,f)|\mathbf{x}(n,f), \widehat{\boldsymbol{\theta}}) \tag{4.1}$$

$$= \prod_{j,n,f} \mathcal{N}(\mathbf{c}_j(n,f)|\mathbf{W}_j(n,f)\mathbf{x}(n,f), (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f)) \tag{4.2}$$

where $\mathbf{c}$ and $\mathbf{x}$ denote the set of all STFT coefficients of the sources and the mixture, $\widehat{\boldsymbol{\theta}}$ the ML value of the parameters of the chosen variance model and $\mathbf{W}_j(n,f)$ the associated Wiener filter defined in (3.5). This approach remains nevertheless mathematically inaccurate since the model parameters are fixed instead of being marginalized over.

In [1, 39], we defined the exact Bayesian estimator of uncertainty as

$$p(\mathbf{c}|\mathbf{x}) = \int p(\mathbf{c}, \boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{4.3}$$

This integral has no closed form expression. Numerical integration via MCMC, which we experimented in [44], is also practically intractable due to the thousands of dimensions involved. If a factored approximation of the joint posterior can be found, however, as the product of smaller dimensional distributions over the source STFT coefficients $\mathbf{c}(n,f)$ in each time-frequency bin and over subsets of parameters $\boldsymbol{\theta}_i$

$$p(\mathbf{c}, \boldsymbol{\theta}|\mathbf{x}) \approx q(\mathbf{c}, \boldsymbol{\theta}) = \prod_{n,f} q(\mathbf{c}(n,f)) \prod_i q(\boldsymbol{\theta}_i) \tag{4.4}$$

then the marginal posterior is simply obtained as $p(\mathbf{c}|\mathbf{x}) = \prod_{n,f} q(\mathbf{c}(n,f))$.

Such an approximation may be obtained via variational Bayesian (VB) inference [Bis06]. The principle of VB inference is to minimize the KL divergence between the approximation and the true posterior. It can be shown that this is equivalent to maximizing the *free energy*

$$\mathscr{L}(q) = \int q(\mathbf{c}, \boldsymbol{\theta}) \log \frac{p(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta})}{q(\mathbf{c}, \boldsymbol{\theta})} \, d\mathbf{c} \, d\boldsymbol{\theta}. \tag{4.5}$$

In our setting, $\mathscr{L}(q)$ is not maximizable in closed form, so we resort to further minorization using auxiliary variables $\boldsymbol{\omega}$. Considering an appropriate lower bound of the joint likelihood $f(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\omega}) \leq p(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta})$, we have

$$\mathscr{L}(q) \geq \mathscr{B}(q, \boldsymbol{\omega}) = \int q(\mathbf{c}, \boldsymbol{\theta}) \log \frac{f(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\omega})}{q(\mathbf{c}, \boldsymbol{\theta})} d\mathbf{c} \, d\boldsymbol{\theta}. \tag{4.6}$$

This bound is iteratively maximized with respect to the auxiliary variables $\boldsymbol{\omega}$ and with respect to the parameters of the approximating distributions $q(\mathbf{c}(n, f))$ and $q(\boldsymbol{\theta}_i)$ as

$$q(\mathbf{c}(n, f)) \propto \exp(\mathbb{E}_{(n', f') \neq (n, f), i}[\log f(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\omega})]) \tag{4.7}$$

$$q(\boldsymbol{\theta}_i) \propto \exp(\mathbb{E}_{(n, f), i' \neq i}[\log f(\mathbf{x}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\omega})]). \tag{4.8}$$

In practice, one iteration of the algorithm consists of computing the sufficient statistics of the variables and updating the parameters of their distributions. This is an extension of the classical EM algorithm for ML or MAP inference where both the model parameters and the hidden variables are now treated as random. VB inference is an emerging topic in audio signal processing and it has been rarely used so far, with the notable exception of [CFG07] for local sparse modeling or [HBC10] for single-level NMF.

In [1], we derived a VB inference algorithm for our flexible variance model presented in Section 3.4.2. This led to generalized inverse Gaussian distributions [Jor82] for the multilevel NMF parameters and Gaussian approximating distributions for the spatial parameters $\widetilde{\mathbf{A}}_j(f)$ and the source STFT coefficients $\mathbf{c}(n, f)$, which we propagated to the features via moment

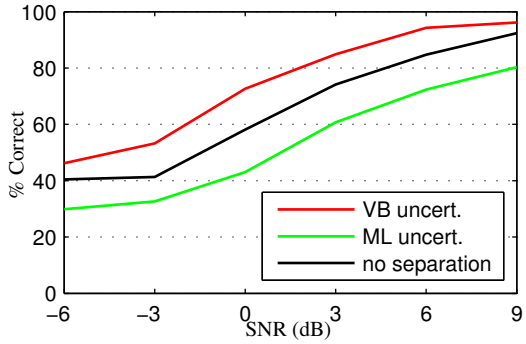


Figure 4.1: Speaker identification accuracy achieved on mixtures of speech and real-world background noise with or without source separation, as a function of the input SNR [1].

matching [AK11]. The accuracy of the resulting feature means and covariances was assessed for a GMM-based speaker recognition task on speech data corrupted by real-world background noise. The results are displayed in Figure 4.1 as a function of the input signal-to-noise ratio (SNR). For this particular data and source model, the ML uncertainty estimator (4.1) does not overcome the distortions introduced by source separation and performs worse than the baseline approach involving no source separation. By comparison, the proposed VB uncertainty estimator improves the recognition accuracy by 9% absolute with respect to the baseline.

4.1.2 Uncertainty training

Let us move up to the classification stage and denote by $\boldsymbol{\mu}_{\mathbf{f}}(n)$ and $\boldsymbol{\Sigma}_{\mathbf{f}}(n)$ the mean and covariance of the feature vector $\mathbf{f}(n)$ in time frame n estimated by uncertainty propagation. The established way of exploiting this uncertain data is called *uncertainty decoding* [DDA05]. Assuming a GMM acoustic model for simplicity and denoting by $\boldsymbol{\mu}_q$, $\boldsymbol{\Sigma}_q$, ω_q the mean, diagonal covariance and weight of the q-th Gaussian mixture for a given speaker class C, the likelihood of the data being produced by that class is given by integrating the likelihood of the GMM over the distribution of the data, which yields

$$p(\boldsymbol{\mu}_{\mathbf{f}}, \boldsymbol{\Sigma}_{\mathbf{f}}|C) = \prod_{n=1}^{N} \sum_{q} \omega_q \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}}(n)|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_{\mathbf{f}}(n)). \tag{4.9}$$

In this equation, the covariance of the model and the uncertainty add up, which can be seen as a form of dynamic model compensation. This modified likelihood boils down to the usual GMM likelihood in the case when $\boldsymbol{\Sigma}_{\mathbf{f}}(n)$ is zero.

In practice, the acoustic models used for decoding are often trained on clean data [DDA05]. This strategy is unfortunately not applicable to singer identification because singer voices are not available in isolation. Also, the estimated uncertainty never perfectly compensates the distortion over the features, so the residual distortion must be compensated by training from noisy data. The simplest strategy is to train the acoustic models in a conventional manner from separated noisy data [DKN$^+$11, KAA$^+$11], but this accounts for the noise twice: the noise in the training data is accounted for by the acoustic model parameters and the noise in the test data is accounted for by uncertainty decoding. Obviously, only the latter should be accounted for at decoding time and the acoustic models should be as invariant as possible to the training noise conditions.

In [6], we designed an EM algorithm for the training of GMMs or HMMs from noisy data. This algorithm exploits the dynamic uncertainty about the training data by maximizing the modified likelihood (4.9) over the model parameters and it operates similarly to the algorithm proposed earlier in [LG07] for static model compensation. By analogy with uncertainty decoding, we refer to this training objective as *uncertainty training*. The EM updates for GMMs are shown in Algorithm 2. It can be seen that only the E-step differs from conventional EM-based GMM training. In particular, the moments of the clean underlying features are estimated by Wiener filtering and used in place of those of the noisy input features in the M-step. This algorithm is fairly general and can be seen as a fundamental tool for GMM-based or HMM-based classification in other fields than audio.

Algorithm 2: Uncertainty training for GMMs. Changes compared to conventional training are shown in red. The diag operator zeroes non-diagonal elements out.

E-step: estimate the underlying clean feature moments by Wiener filtering

$$\gamma_q(n) = \frac{\omega_q \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}}(n)|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_{\mathbf{f}}(n))}{\sum_{q'} \omega_{q'} \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}}(n)|\boldsymbol{\mu}_{q'}, \boldsymbol{\Sigma}_{q'} + \boldsymbol{\Sigma}_{\mathbf{f}}(n))}$$

$$\mathbf{W}_q(n) = \boldsymbol{\Sigma}_q(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_{\mathbf{f}}(n))^{-1}$$

$$\widehat{\mathbf{f}}_q(n) = \boldsymbol{\mu}_q + \mathbf{W}_q(n)\,(\boldsymbol{\mu}_{\mathbf{f}}(n) - \boldsymbol{\mu}_q)$$

$$\widehat{\mathbf{R}}_{\mathbf{ff}_q}(n) = \widehat{\mathbf{f}}_q(n)\widehat{\mathbf{f}}_q(n)^T + (\mathbf{I} - \mathbf{W}_q(n))\boldsymbol{\Sigma}_q$$

M-step: update the GMM parameters

$$\omega_q = \frac{1}{N}\sum_{n=1}^{N} \gamma_q(n)$$

$$\boldsymbol{\mu}_q = \frac{1}{\sum_{n=1}^{N} \gamma_q(n)} \sum_{n=1}^{N} \gamma_q(n)\widehat{\mathbf{f}}_q(n)$$

$$\boldsymbol{\Sigma}_q = \mathrm{diag}\left(\frac{1}{\sum_{n=1}^{N} \gamma_q(n)} \sum_{n=1}^{N} \gamma_q(n)\widehat{\mathbf{R}}_{\mathbf{ff}_q}(n) - \boldsymbol{\mu}_q\boldsymbol{\mu}_q^T\right)$$

We evaluated this algorithm over the same speaker recognition task as above using a different uncertainty propagation technique based on vector Taylor series [MRS96]. We reported an improvement of recognition accuracy on the order of 7 to 11% absolute compared to conventional training on clean data and 3 to 4% absolute compared to conventional training on noisy data [6]. Interestingly, this improvement was not only observed in matched training and test noise conditions, but also with multi-condition or unmatched training data. More recently, we applied the same approach to singer identification in polyphonic music recordings and achieved a promising accuracy of 94% for 10 classes, compared to 64% without source separation [41].

4.1.3 Evaluation campaigns

Numerous evaluation campaigns have been held in the field of robust speech processing in the last twenty years. The effect of reverberation, background noise or competing talkers has been evaluated in [PH00, CHR10] for instance. Real-world data have also been recorded and publicly released, especially for meeting environments [JBE+03, RHB07]. In order to pursue this effort and assess recent progress, I co-organized the *1st CHiME Speech Separation and Recognition Challenge* in 2011. This challenge aimed to recognize spoken commands consisting of one letter and one digit binaurally recorded in a real-world domestic noise environment with SNRs ranging from -6 to +9 dB. We provided an analysis of the results in [5]. The best algorithm [DKN+11] achieved an average keyword error rate of 12% at -6 dB SNR and 4% at 9 dB SNR. This is

much lower than the baseline but still twice as much as the error rate of a human listener, which indicates that room is left for progress. A second edition of the challenge is currently being run, which considers more difficult situations involving moving sources or larger vocabulary.

4.2 Towards multilayer modeling of musical language

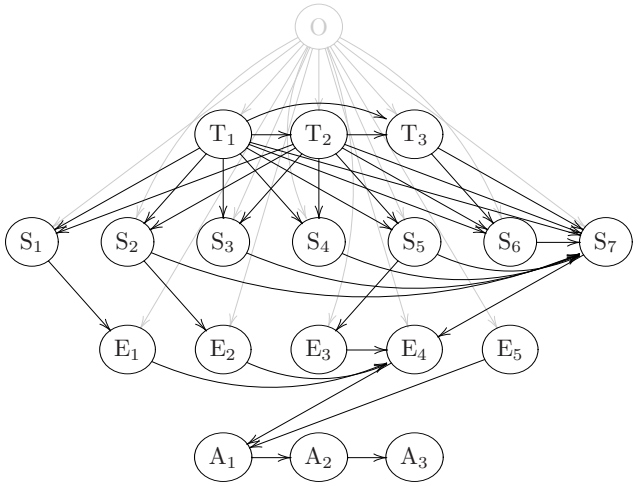
While natural language processing is a mature topic, few studies have attempted to model the "language" of music in its full complexity. Most music information retrieval systems are designed to estimate a single piece of information. For instance, polyphonic pitch transcription and chord identification are typically performed independently of each other without accounting for the strong dependencies between these two pieces of information. In addition, most systems rely on general pattern recognition techniques applied onto unordered bags of features or on first-order HMMs representing the short-term evolution of the considered variable. It is admitted that the use of such low-level models bounds the accuracy of these systems to a glass ceiling [AP04] and that a system integrating multiple layers and time scales of information would be more versatile and accurate and enable more complex interaction with the data [DBC09].

In [66], we attempted to lay down the scientific challenges raised by the development of such a complete system and presented a tentative roadmap. Figure 4.2 illustrates the dependencies between some of the variables underlying a music piece in the form of a graphical model. In this graph, each node represents a variable or a temporal sequence of variables and the statistical dependencies between these variables are indicated by arrows such that the conditional distribution of a variable given its ancestors depends on its parents only. It is worth pointing that there is no agreed-upon definition of variables such as structure or rhythm and that this figure is provided merely to illustrate the complexity of the problem. This complexity can be seen in particular in the fact that the "overall features" affect all other variables, while the "quantized notes" depend on many other variables. Dependencies also arise between variables at multiple time scales, ranging from beats, bars, or structural blocks to the whole piece.

Altogether, the curse of dimensionality resulting from these complex dependencies makes it impossible to learn the joint distribution of all variables directly. Low-dimensional parametric approximations of this distribution and model smoothing techniques must be found to avoid overfitting. The limited application range of musicological expertise and the existence of different music cultures also raise the issues of unsupervised learning and model selection. Finally, learning will require large annotated music corpora which may be acquired in the future only by developing new semi-automatic annotation procedures exploiting the wealth of data available in physical and online music archives.

4.2.1 Polyphonic pitch and chord modeling

In a preliminary study [74], we considered the modeling of medium-term dependencies in chord sequences via probabilistic N-grams [MDH$^+$07]. A chord is an abstraction of a set of concurrent pitches which encodes the musically relevant pitches within this set by a symbol, e.g., "C major". We explored model smoothing techniques introduced for the reduction of overfitting in spoken language modeling [CG98] and demonstrated their applicability to this new context.



Overall features

O *Tags*: set of tags in $\mathcal{V}_{\text{tags}}$ covering all or part of the piece, *e.g.* genre, mood, composer, performer, place and user preference

Temporal organization features

T_1 *Structure*: set of possibly overlapping/nested sections, each defined by its quantized duration in bars and by a section symbol in $\mathcal{V}_{\text{sect}}$

T_2 *Meter*: sequence of bars, each defined by its reference beat and time signature in $\mathcal{V}_{\text{meter}}$ and by the associated metrical accentuation level of each beat and beat subdivision

T_3 *Rhythm*: sequence of *events* associated to one or more simultaneous note onsets, each defined by its quantized duration in beats and by the associated number of onsets

Symbolic features

S_1 *Notated tempo*: beat-synchronous sequence of tempo and tempo variation symbols in $\mathcal{V}_{\text{tempo}}$

S_2 *Notated loudness*: beat-synchronous sequence of loudness and loudness variation symbols in $\mathcal{V}_{\text{loud}}$

S_3 *Key/mode*: beat-synchronous sequence of key/mode symbols in $\mathcal{V}_{\text{key}}$

S_4 *Harmony*: beat-synchronous sequence of chord symbols in $\mathcal{V}_{\text{chord}}$

S_5 *Instrumentation*: beat-synchronous sequence of sets of active voices, each defined by a voice symbol in $\mathcal{V}_{\text{inst}}$ (including instruments, orchestra sections, singer identities, and sample IDs, with various playing or singing styles)

S_6 *Lyrics*: event-synchronous sequence(s) of syllables in $\mathcal{V}_{\text{syll}}$

S_7 *Quantized notes*: set of notes (including pitched/drum notes, voices or samples), each defined by quantized onset and duration in beats, its articulation symbol in $\mathcal{V}_{\text{artic}}$, its loudness and loudness variation symbol in $\mathcal{V}_{\text{loud}}$, its quantized pitch and pitch variation symbol in $\mathcal{V}_{\text{pitch}}$, its voice symbol in $\mathcal{V}_{\text{inst}}$ and its syllable in $\mathcal{V}_{\text{syll}}$

Expressive performance features

E_1 *Expressive tempo*: beat-synchronous sequence of actual tempo values in bpm

E_2 *Expressive loudness*: beat-synchronous sequence of actual global loudness values in sones

E_3 *Instrumental timbre*: beat-synchronous sequence of vectors of parameters modeling the timbre space of each voice

E_4 *Expressive notes*: set of notes, each defined by its actual onset time and duration in s, its loudness curve in sones, its pitch curve in Hz, and its trajectory in the timbre space

E_5 *Rendering*: time-synchronous sequence of vectors of parameters characterizing the recording setup (e.g., reverberation time, mic spacing) or the software mixing effects and the spatial position and spatial width of each voice

Acoustic features

A_1 *Tracks*: rendered acoustic signal of each voice

A_2 *Mix*: overall acoustic signal

A_3 *Classical low-level features*: MFCCs, chroma, etc

Figure 4.2: Draft model of a music piece, from [66]. Dependencies upon overall features are shown in light gray for legibility. The alphabets $\mathcal{V}_{\text{sect}}$, $\mathcal{V}_{\text{meter}}$, $\mathcal{V}_{\text{tempo}}$, $\mathcal{V}_{\text{loud}}$, $\mathcal{V}_{\text{chord}}$, $\mathcal{V}_{\text{inst}}$, $\mathcal{V}_{\text{artic}}$, $\mathcal{V}_{\text{pitch}}$ may depend on the music culture or style, e.g., Western vs. Indian.

We then focused on the parameterization and learning of higher-dimensional distributions over symbolic music variables, epitomized by the problem of polyphonic pitch modeling. Assuming a discrete pitch scale such as the semitone scale and denoting by N_{pn} the Boolean variable indicating whether a note of pitch p is active in time frame n or not, the problem consists of learning the prior distribution $p(\mathbf{N})$ of all pitches on all time frames. In [RK05], a "horizontal" key-dependent model was proposed which represents the pitch durations and the intervals between successive pitches in a given instrument line, but not those between concurrent pitches. In [RS03], a "vertical" model was proposed instead which accounts for the dependency of concurrent pitches on the underlying chord, but chromatic pitch classes were considered instead of absolute pitches and temporal dependencies were only present between chords.

In [4, 65], we designed the first polyphonic pitch model to our knowledge that accounts both for horizontal and vertical structure. This was achieved by training and smoothing a number of low-dimensional *submodels* representing the probability of the current pitch N_{pn} being active conditionally to subsets of other variables, e.g., the previous pitch in the same line, the lower pitches in the same time frame or the underlying chord. These submodels were then merged via linear or log-linear interpolation [Kla98] and the interpolation weights were optimized on development data. The joint model was compared to individual submodels and shown to slightly improve the prediction of the test data as measured in terms of *cross-entropy*, that is a normalized version of the log-likelihood. We also analyzed the prediction performance in different contexts, e.g., note onsets or offsets, using a new cross-entropy measure designed for this purpose.

Note that the problem of joint horizontal and vertical modeling was separately tackled in [MD10] in the context of chord sequences. The authors multiplied submodels without normalizing the resulting distribution [MD10, eq. 12], though, which is mathematically erroneous and may lead to wrong estimation of the most probable chord sequence. This multiplication operation may also excessively sparsify the distribution and it lacks tunable factors to account for the greater importance of certain submodels compared to others. The above interpolation techniques address all these issues and we expect them to play an important role in music language processing in the future beyond the problem of polyphonic pitch modeling alone.

4.2.2 Music structure estimation

On a complementary line, we leveraged our work on the formalization of the concept of music structure (see Section 1.2) to build a family of algorithms for automatic structure estimation [53]. The input audio is first segmented into structural blocks via a Viterbi algorithm accounting for the paradigmatic (repetition) and syntagmatic (rupture) properties of the blocks, as well as for their duration. The estimated blocks are then grouped into equivalence classes and the optimal number of classes is found via a data-adaptive model selection criterion. More recently, we also showed how to exploit the morphological properties of the blocks for this task [103]. These algorithms are too complex to be detailed here, but let us just say that the algorithm in [53] ranked first for the "Audio Structural Segmentation" task of MIREX 2011 in terms of segment boundary F-measure, both with 0.5 s and 3 s tolerance [104].

Chapter 5

Conclusion and perspectives

5.1 Achievements

To sum up, I have been targeting the long-term goal of addressing the challenges raised by the multisource and multilayer structure of audio signals in the fields of signal processing and information retrieval. As a researcher, a supervisor or a collaborator, I have gradually sought to cover the various areas relevant to this goal, from low-level signal processing tasks to higher-level content description tasks. My contributions range from the formalization of the studied problems and the design of models to the derivation of estimation algorithms and their experimental evaluation. The heterogeneity of the data and information involved led me to rely on multidisciplinary theoretical foundations from Bayesian inference and convex optimization to acoustics and computational musicology. When needed, I also contributed to these foundations by developing new theoretical tools which are applicable outside of the field of audio.

In the last eight years since the end of my PhD, audio source separation has become a mature research topic. Commercial services and products have grown, especially for speech or music remixing applications in which artifacts over the separated signals disappear to a great extent in the remixed output signal. Most of my work on variance model-based source separation has been or is currently being transferred to companies via research collaborations and patent filing. Reverberant, dynamic, or noisy mixtures have remained nevertheless difficult to separate to the level of quality required by other potential applications and further research is required.

As an ubiquitous application of speech processing, I have become interested in noise-robust speech recognition. While a huge body of work has been published in this area, many current systems rely on traditional beamforming or denoising techniques and on somewhat heuristic integration with automatic speech recognition. The use of modern source separation techniques and uncertainty propagation has just started and the proposed Bayesian uncertainty estimation framework appears very promising in this context. Room is still left for progress towards human performance and I expect that this work will especially benefit from my change of affiliation to the PAROLE team of Inria Nancy - Grand Est on January 1, 2013.

In addition to the above problems, I have devoted some time to the computational modeling of music. The potential applications are huge, but so are the challenges and we barely scratched the surface so far.

5.2 Directions

My research program aims to pursue these efforts and eventually come up with audio processing systems able to model the complexity of the data and provide a unified approach to various application needs from signal processing and information retrieval to content manipulation. The theory of Bayesian inference offers a suitable framework, which makes it possible to integrate individual model pieces in a principled fashion and to derive estimation algorithms which are intrinsically modular and robust to missing or uncertain data.

Regarding the acoustic modeling of audio sources, the studies made up to now remain to be merged into a unified modeling framework combining the advantages of linear modeling for point sources or periodic sources with those of variance modeling for diffuse or noisy sources. The modeling of the spectral envelope coefficients within the proposed multilevel NMF model also appears necessary in order to allow two-way interaction between acoustic modeling on the one hand and speech recognition on the other hand. Finally, computationally efficient modeling of the source movements remains a relatively open issue. In order to obtain a complete system in the longer term, I am also interested in developing Bayesian language models of speech and music. The PhD of Alex Mesnil, which just started, will be focusing on this direction. Sparse regularization is a promising way to overcome the overfitting issues limiting the performance of conventional probabilistic language models and come closer to the performance of state-of-the-art deep neural networks. Together, these studies will contribute to filling the library of models which are necessary to represent audio contents.

Contrary to MAP estimation which is preferred for its simplicity today, full Bayesian estimation is not only robust to missing data and noise but it also yields a confidence measure in terms of the posterior distribution of the model parameters. The estimation of this posterior distribution is the main theoretical basis behind the robust integration of multiple processing blocks such as source separation, feature extraction, speech recognition and machine translation. The design of computationally scalable Bayesian inference algorithms is thus crucial. The theory of VB inference provides a solid algorithmic foundation, but its implementation raises several challenges in a context involving thousands or millions of dependent variables. One of these issues is to automatically find an approximation of the target distribution which is at the same time close enough and easy to optimize. This may be achieved by analyzing the posterior dependencies between the variables and developing the use of complex approximations such as structured mean field [Wie00] or mixture mean field [JJ98] which have not been used in audio so far. The PhD of Tran Dung, which will start at the end of this year, will be dedicated to this issue in the specific context of source separation and uncertainty propagation. Another issue is to suitably initialize the variables so as to avoid convergence to bad local optima. The joint use of multiple models estimated from different initial values [Die00] is an interesting direction that will be considered. Finally, the issue of model selection has been very little explored in audio.

The above models and algorithms will mainly be applied to speech enhancement and recognition in noisy environments and to derived applications such as noise-robust indexing of spoken documents. Focused music applications will also be explored. The benefit of the Bayesian approach will be validated in terms of its robustness to the presence of several sources but also by its ability to infer and exploit complex models from small amounts of data. The Inria Technolog-

ical Development Action “FASST”, which will develop our flexible source separation toolbox into efficient software, is expected to play a strong role in the dissemination of our source separation technology towards companies or other areas of audio signal processing research.

Appendix A

Detailed CV

Office address:
Inria
Campus de Beaulieu
F-35042 Rennes Cedex, France

Phone: +33 2 9984 2269
Fax: +33 2 9984 7171
Email: emmanuel.vincent@inria.fr
Web: http://www.irisa.fr/metiss/members/evincent/

A.1 Positions held

Experienced Research Scientist (CR1), Inria, Rennes (France)	2009–
Junior Research Scientist (CR2), Inria, Rennes (France)	2006–2009
Post-doctoral Research Assistant, Queen Mary, University of London (UK)	2004–2006
Graduate teaching assistant, University Paris 6 (France)	2003–2004
Teaching assistant, Lycée Louis-le-Grand, Paris (France)	2000–2002

A.2 Degrees

PhD, IRCAM and Dept. of Computer Science, University of Paris 6 (France) 2004
Dissertation: *Instrument models for source separation and transcription of musical audio*
Advisor: Xavier Rodet – Jury: C. Jutten, B. Torrésani, P. Flandrin, J.-D. Polack, M. Sandler

MSc, IRCAM, Paris (France), ranked first 2001
Thesis: *Separation of audio signals : statistical principles of independent component analysis and application to monophonic signals*
Advisor: Xavier Rodet

BSc, Department of Mathematics, École Normale Supérieure, Paris (France) 2000
Thesis: *Smoluchowski's equation: existence of solutions and gelation phenomenon*
Advisor: Benoît Perthame

A.3 Distinctions

Recipient of the **SPIE ICA Unsupervised Learning Pioneer Award** for contributions to signal separation (2012)

2nd Prize of Rennes 1 Foundation given to Ngoc Duong for his PhD co-supervised with R. Gribonval (2012)

Co-author of the **best source separation system** for instantaneous mixtures at SASSEC 2007, SiSEC 2010 and 2011, and for professionally produced music recordings (all sources) at SiSEC 2011

Co-author of the **best structural music segmentation system** (segment boundary F-measure) at MIREX 2011

IEEE Senior Member (2009–)

A.4 Research supervision

Post-doctoral research assistants

Joachim Thiemann (Canon Research contract, 50% with N. Bertin)	2012–2013
Stanisław Raczyński (INRIA grant, 100%)	2011–2013
Laurent Simon (i3DMusic project grant, 100%)	2011–2013
Kamil Adiloğlu (Quaero project grant, 100%)	2010–2012
Alexey Ozerov (Quaero project grant, 100%), postdoc at Technicolor	2009–2011
Valentin Emiya (INRIA grant, 50% with R. Gribonval), lecturer at U. Provence	2008–2009

PhD students

Tran Dung (INRIA grant, 50% with D. Jouvet)	2012–2015
Alex Mesnil (ENS Lyon grant, 50% with G. Obozinski)	2012–2015
Alexis Benichoux (MENRT grant, 50% with R. Gribonval)	2010–2013
Gabriel Sargent (MENRT grant, 50% with F. Bimbot)	2009–2012
Nobutaka Ito (JSPS grant jointly with U. Tokyo, 50% with N. Ono), now at NTT	2009–2012
Ngoc Duong (INRIA grant, 75% with R. Gribonval), postdoc at Technicolor	2008–2011

Research engineers

Dimitris Moreau (100%)	2012–2013

MSc students: Christopher Sutton, Antoine Movschin, Pascal Bado, Ricardo Scholz, Christophe Hauser, Gabriel Sargent, Charles Blandin

Undergraduates: Arthur Vimond

Visitors (1+ months): Pierre Leveau (ENST, France), Satoru Fukayama, Hideyuki Tachibana, Jun Wu and Kosuke Suzuki (University of Tokyo, Japan), Andrés Coca Salazar (University of São Paulo)

A.5 Research and technology transfer projects

Inria Technological Development Action "FASST" (PI) 2012–2014
Partners: Inria teams PAROLE (Nancy) and TEXMEX (Rennes)
Goal: development of a reference software toolbox for audio source separation
Funding: 1 software developer for 2 years

NII Grand Challenge Project (PI) 2012–2013
Partners: National Institute of Informatics, U. of Tsukuba, Tokyo Inst. of Technology (Japan)
Goal: research on source localization and separation by ad-hoc recording devices
Funding: 50,000 JPY

Contract with Musiciens Artistes Interprètes Associés SARL (co-PI) 2011–2014
Goal: application of my research to the sound engineering services offered by the company
Funding: 216,837 EUR

Research contract with Canon Research Centre France SAS (PI) 2011–2013
Goal: application of my research to some of the products sold by Canon Inc.
Funding: 148,184 EUR

EUREKA Eurostars project "i3DMusic" (WP manager) 2010–2012
Partners: Audionamix SA (France), Sonic Emotion AG (Switzerland), EPFL (Switzerland)
Goal: enabling real-time interactive respatialization of stereo music content
Funding: 164,561 EUR

Inria Associate Team VERSAMUS (PI) 2010–2012
Partner: University of Tokyo (Japan)
Goal: integrating multiple features for music information retrieval
Funding: 37,700 EUR

Contract with Nippon Telegraph and Telephone Corp. (PI) 2010–2011
Goal: joint research on consistent Wiener filtering

Bilateral French-German project PHC Procope (PI) 2009–2010
Partner: Carl von Ossietzky Universität Oldenburg (Germany)
Goal: designing new subjective and objective evaluation metrics for audio source separation
Funding: 6,000 EUR

A.6 Collective responsibilities

Titular member of the National Council of Universities (CNU section 61, 2012–2015)

Member of the **PhD jury** of Pierre Leveau (ENST and University Paris 6, 2007), Valentin Emiya (ENST, 2008), Jonathan Le Roux (ENS and University Paris 5, 2009) and Jean-Louis Durrieu (Télécom ParisTech, 2010)

Associate editor, *IEEE Transactions on Audio, Speech, and Language Processing* (2011–2014)

Guest editor:

- *Computer Speech and Language*, special issue on Speech Separation and Recognition in Multisource Environments (2012)
- *Signal Processing*, special issue on Latent Variable Analysis and Signal Separation (2011)

Steering committee member, International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA, formerly the International Conference on Independent Component Analysis and Signal Separation) (2010–)

General chair:

- $2^{\mathbf{nd}}$ International Workshop on Machine Listening in Multisource Environments (CHiME) (2013)
- $9^{\mathbf{th}}$ International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA) (120 attendees, 2010)
- $1^{\mathbf{st}}$ International Workshop on Machine Listening in Multisource Environments (CHiME) (70 attendees, 2011)
- $15^{\mathbf{th}}$ French Symposium on Computer Music (JIM) (50 attendees, 2010)
- $1^{\mathbf{st}}$ French Young Researchers Conference on Audition, Musical acoustics and Audio signal processing (JJCAAS) (40 attendees, 2003)

Evaluation challenges:

- Founding chair, Signal Separation Evaluation Campaign (SiSEC) (3 editions totaling 114 entries, 2008–)
- Founding chair, CHiME Speech Separation and Recognition Challenge (1 edition with 13 entries in 2011, second edition ongoing)
- Vice-chair, $1^{\mathbf{st}}$ Annual Music Information Retrieval Evaluation eXchange (MIREX) (80 entries, 2005)
- General chair, Stereo Audio Source Separation Evaluation Campaign (SASSEC) (15 entries, 2007)

Mailing list founding manager: machinelistening@googlegroups.com (1087 members)

Member of the **IEEE Audio and Acoustic Signal Processing Technical Committee** (2012–2015)

Program committee: LVA/ICA 2012, ISMIR 2011 and 2012

Special sessions: LVA/ICA 2012, ICA 2007, ICA 2006

A.7 Keynotes and tutorials

R.F. ASTUDILLO, E. VINCENT AND L. DENG. Uncertainty handling for environment-robust speech recognition. *Interspeech*, 2012.

E. VINCENT. Advances in audio source separation and multisource audio content retrieval. *SPIE Defense, Security, and Sensing*, 2012.

E. VINCENT. Music source separation. *14th Int. Conf. on Digital Audio Effects (DAFx)*, 2011.

E. VINCENT AND N. ONO. Music source separation and its applications to MIR. *2010 Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2010.

E. VINCENT. Audio source separation using hierarchical phase-invariant models. *2009 ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, 2009.

E. VINCENT. Blind audio source separation: A review of state-of-the-art techniques. *UK ICA Research Network Workshop*, 2005.

A.8 Teaching

Audio coding, rendering and source separation, MSc in Computer Science, University of Rennes 1, and ESAT (16h/year)	2006–2012
Polyphonic music processing, MSc in Digital Music Processing, Queen Mary, University of London (9h)	2005–2006
Audio source separation, MSc in Digital Music Processing, Queen Mary, University of London (2h)	2004–2005
Mathematics, 1st year BSc, University of Paris 6 (60h)	2003–2004
Mathematics, 1st and 2nd year Preparatory Engineering Class, Lycée Louis-le-Grand	2000–2002

Appendix B

List of publications

Full-text versions of all my publications are available by clicking on the Publications link on my homepage http://www.irisa.fr/metiss/members/evincent/. Citation counts are provided for the 22 most cited papers according to Google Scholar on August 27, 2012. The total citation count is 1763.

B.1 Papers in international peer-reviewed journals

[1] K. ADILOĞLU AND E. VINCENT. Variational Bayesian inference for source separation and robust feature extraction. To be submitted (preprint: http://hal.inria.fr/hal-00726146/PDF/RT-428.pdf).

[2] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. Spatial location priors for Gaussian model-based reverberant audio source separation. To be submitted (preprint: http://hal.inria.fr/hal-00727781/PDF/RR-8057.pdf).

[3] J. LE ROUX AND E. VINCENT. Consistent Wiener filtering for audio source separation. Under review (preprint: http://hal.inria.fr/hal-00725350/PDF/RR-8049.pdf).

[4] S.A. RACZYŃSKI, E. VINCENT AND S. SAGAYAMA. Dynamic Bayesian networks for symbolic polyphonic pitch modeling. To be submitted (preprint: http://hal.inria.fr/hal-00728771/PDF/RT-430.pdf).

[5] J. BARKER, E. VINCENT, N. MA, H. CHRISTENSEN AND P. GREEN. The PASCAL CHiME Speech Separation and Recognition Challenge. *Computer Speech and Language*, to appear.

[6] A. OZEROV, M. LAGRANGE AND E. VINCENT. Uncertainty-based learning of acoustic models from noisy data. *Computer Speech and Language*, to appear.

[7] C. BLANDIN, A. OZEROV AND E. VINCENT. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[8] A. OZEROV, E. VINCENT AND F. BIMBOT. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Lan-*

guage Processing, vol. 20, no. 4, pp. 1118–1133, 2012.

[9] V. VIGNERON, V. ZARZOSO, R. GRIBONVAL AND E. VINCENT. Latent variable analysis and signal separation (Guest editorial). *Signal Processing*, vol. 92, pp. 1765–1766, 2012.

[10] E. VINCENT, S. ARAKI, F. THEIS, G. NOLTE, P. BOFILL ET AL.. The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[11] V. EMIYA, E. VINCENT, N. HARLANDER AND V. HOHMANN. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[12] J. WU, E. VINCENT, S.A. RACZYŃSKI, T. NISHIMOTO, N. ONO AND S. SAGAYAMA. Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds. *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no.6, pp. 1124–1132, 2011.

[13] R. BADEAU, N. BERTIN AND E. VINCENT. Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, vol. 21, no. 11, pp. 1869–1881, 2010.

[14] N. BERTIN, R. BADEAU AND E. VINCENT. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, p. 538–549, 2010 (**45 citations**).

[15] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010 (**26 citations**).

[16] M. KOWALSKI, E. VINCENT AND R. GRIBONVAL. Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[17] E. VINCENT, N. BERTIN AND R. BADEAU. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010 (**46 citations**).

[18] M.G. JAFARI, E. VINCENT, S.A. ABDALLAH, M.D. PLUMBLEY AND M.E. DAVIES. An adaptive stereo basis method for convolutive blind audio source separation. *Neurocomputing*, vol. 71, pp. 2087–2097, 2008.

[19] P. LEVEAU, E. VINCENT, G. RICHARD AND L. DAUDET. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116–128, 2008 (**52 citations**).

[20] E. VINCENT AND M.D. PLUMBLEY. Efficient Bayesian inference for harmonic models via adaptive posterior factorization. *Neurocomputing*, vol. 72, pp. 79–87, 2008.

[21] E. VINCENT AND M.D. PLUMBLEY. Low bit-rate object coding of musical audio using Bayesian harmonic models. *IEEE Transactions on Audio, Speech and Language Processing*,

vol. 15, no. 4, pp. 1273–1282, 2007 (**25 citations**).

[22] E. VINCENT, R. GRIBONVAL AND M.D. PLUMBLEY. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007 (**48 citations**).

[23] E. VINCENT. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, p. 91–98, 2006 (**82 citations**).

[24] E. VINCENT, R. GRIBONVAL AND C. FÉVOTTE. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006 (**344 citations**).

B.2 Book chapters

[25] G. EVANGELISTA, S. MARCHAND, M. PLUMBLEY AND E. VINCENT. Sound source separation, In *Digital Audio Effects, 2nd Edition*, U. Zölzer (ed.), Wiley, pp. 551–588, 2011.

[26] A. NESBIT, M.G. JAFARI, E. VINCENT AND M.D. PLUMBLEY. Audio source separation using sparse representations. In *Machine Audition: Priciples, Algorithms and Systems*, W. Wang (ed.), IGI Global, pp. 246–265, 2010.

[27] E. VINCENT AND Y. DEVILLE. Audio applications. In *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, P. Comon and C. Jutten (eds.), Academic Press, pp. 779–819, 2010.

[28] E. VINCENT, M.G. JAFARI, S.A. ABDALLAH, M.D. PLUMBLEY AND M.E. DAVIES. Probabilistic modeling paradigms for audio source separation. In *Machine Audition: Priciples, Algorithms and Systems*, W. Wang (ed.), IGI Global, pp. 162–185, 2010 (**22 citations**).

[29] M.E. DAVIES, M.G. JAFARI, S.A. ABDALLAH, E. VINCENT AND M.D. PLUMBLEY. Blind source separation using space-time independent component analysis. In *Blind speech separation*, S. Makino, T.-W. Lee and H. Sawada (eds.), Springer, pp. 79–99, 2007.

B.3 Invited papers in international conferences

[30] S. ARAKI, F. NESTA, E. VINCENT, Z. KOLDOVSKY, G. NOLTE ET AL.. The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio source separation -. In *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 414–422, 2012.

[31] G. NOLTE, D. LUTTER, A. ZIEHE, F. NESTA, E. VINCENT ET AL.. The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Biomedical data analysis -. In *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 423–429, 2012.

[32] E. VINCENT. Advances in audio source separation and multisource audio content retrieval.

In *Proc. SPIE Defense, Security, and Sensing*, 2012.

[33] E. VINCENT. Audio source separation using hierarchical phase-invariant models. In *Proc. 2009 ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, pp. 20–24, 2009.

[34] E. VINCENT, S. ARAKI AND P. BOFILL. The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In *Proc. 8^{th} Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 734–741, 2009 (**44 citations**).

[35] E. VINCENT, H. SAWADA, P. BOFILL, S. MAKINO AND J.P. ROSCA. First Stereo Audio Source Separation Evaluation Campaign: data, algorithms and results. In *Proc. 7^{th} Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 552–559, 2007 (**67 citations**).

[36] E. VINCENT AND M.D. PLUMBLEY. Single-channel mixture decomposition with Bayesian harmonic models. In *Proc. 6^{th} Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 722–730, 2006.

[37] S. DOWNIE, K. WEST, A. EHMANN AND E. VINCENT. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proc. 6^{th} Int. Conf. on Music Information Retrieval (ISMIR)*, p. 320–323, 2005 (**84 citations**).

B.4 Invited papers in national conferences

[38] E. VINCENT, C. FÉVOTTE AND R. GRIBONVAL. Comment évaluer les algorithmes de séparation de sources audio ? In *Actes du 19^{e} colloque GRETSI*, vol. 1, p. 27–30, 2003.

B.5 Peer-reviewed papers in international conferences

[39] K. ADILOĞLU AND E. VINCENT. A general variational Bayesian framework for robust feature extraction in multisource recordings. In *Proc. 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 273–276, 2012.

[40] F. BIMBOT, E. DERUTY, G. SARGENT AND E. VINCENT. Semiotic structure labeling of music pieces: concepts, methods and annotation conventions. In *Proc. 13^{th} Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012.

[41] M. LAGRANGE, A. OZEROV AND E. VINCENT. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. In *Proc. 13^{th} Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012.

[42] L.S.R. SIMON AND E. VINCENT. A general framework for online audio source separation. In *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 397–404, 2012.

[43] E. VINCENT. Improved perceptual metrics for the evaluation of audio source separation. In *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 430–437, 2012.

[44] K. ADILOĞLU AND E. VINCENT. An uncertainty estimation approach for the extraction of source features in multisource recordings. In *Proc. 19th European Signal Processing Conference (EUSIPCO)*, pp. 1663–1667, 2011.

[45] R. BADEAU, N. BERTIN AND E. VINCENT. Stability analysis of multiplicative update algorithms for non-negative matrix factorization. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2148–2151, 2011.

[46] A. BENICHOUX, E. VINCENT AND R. GRIBONVAL. A compressed sensing approach to the simultaneous recording of multiple room impulse responses. In *Proc. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 285–288, 2011.

[47] F. BIMBOT, E. DERUTY, G. SARGENT AND E. VINCENT. Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks. In *Proc. 2011 Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 287–292, 2011.

[48] C. BLANDIN, E. VINCENT AND A. OZEROV. Multi-source TDOA estimation using SNR-based angular spectra. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2616–2619, 2011.

[49] N.Q.K. DUONG, H. TACHIBANA, E. VINCENT, N. ONO, R. GRIBONVAL AND S. SAGAYAMA. Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 205–208, 2011.

[50] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. An acoustically-motivated spatial prior for under-determined reverberant source separation. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9–12, 2011.

[51] A. OZEROV AND E. VINCENT. Using the FASST source separation toolbox for noise robust speech recognition. In *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pp. 86–87, 2011.

[52] A. OZEROV, M. LAGRANGE AND E. VINCENT. GMM-based classification from noisy features. In *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pp. 30–35, 2011.

[53] G. SARGENT, F. BIMBOT AND E. VINCENT. A regularity-constrained Viterbi algorithm and its applicatioin to the structural segmentation of songs. In *Proc. 2011 Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 483–488, 2011.

[54] J. WU, E. VINCENT, S.A. RACZYŃSKI, T. NISHIMOTO, N. ONO AND S. SAGAYAMA. Multipitch estimation by joint modeling of harmonic and transient sounds. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 25–28, 2011.

[55] J. WU, E. VINCENT, S.A. RACZYŃSKI, T. NISHIMOTO, N. ONO AND S. SAGAYAMA. Musical instrument identification based on new boosting algorithm with probabilistic decisions. In *Proc. 2011 Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, 2011.

[56] S. ARBERET, A. OZEROV, N.Q.K. DUONG, E. VINCENT, R. GRIBONVAL ET AL.. Non-negative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *Proc. 2010 IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1–4, 2010.

[57] F. BIMBOT, O. LE BLOUCH, G. SARGENT AND E. VINCENT. Decomposition into autonomous and comparable blocks: A structural description of music pieces. In *Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 189–194, 2010.

[58] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. Under-determined convolutive blind source separation using spatial covariance models. In *Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9–12, 2010.

[59] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 73–80, 2010.

[60] V. EMIYA, E. VINCENT, N. HARLANDER AND V. HOHMANN. Multi-criteria subjective and objective evaluation of audio source separation. In *Proc. AES 38th Conf. on Sound Quality Evaluation*, p. 251–259, 2010.

[61] N. ITO, N. ONO, E. VINCENT AND S. SAGAYAMA. Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra. In *Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2818–2821, 2010.

[62] N. ITO, E. VINCENT, N. ONO, R. GRIBONVAL AND S. SAGAYAMA. Crystal-MUSIC: accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, p. 81–88, 2010.

[63] J. LE ROUX, E. VINCENT, Y. MIZUNO, H. KAMEOKA, N. ONO AND S. SAGAYAMA. Consistent Wiener filtering: generalized time-frequency masking respecting spectrogram consistency. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 89–96, 2010.

[64] A. OZEROV, E. VINCENT AND F. BIMBOT. A general modular framework for audio source separation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, p. 33–40, 2010.

[65] S.A. RACZYŃSKI, E. VINCENT, F. BIMBOT AND S. SAGAYAMA. Multiple pitch transcription using DBN-based musicological models. In *Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 363–368, 2010.

[66] E. VINCENT, S.A. RACZYŃSKI, N. ONO AND S. SAGAYAMA. A roadmap towards versatile MIR. In *Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 662–664, 2010.

[67] E. VINCENT. An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 157–164, 2010.

[68] N. BERTIN, R. BADEAU AND E. VINCENT. Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 29–32, 2009.

[69] N.Q.K. DUONG, E. VINCENT AND R. GRIBONVAL. Spatial covariance models for underdetermined reverberant audio source separation. In *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 129–132, 2009.

[70] V. EMIYA, E. VINCENT AND R. GRIBONVAL. An investigation of discrete-state discriminant approaches to single-sensor source separation. In *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 97–100, 2009.

[71] A. NESBIT, E. VINCENT AND M.D. PLUMBLEY. Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation. In *Proc. 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 37–40, 2009.

[72] A. NESBIT, E. VINCENT AND M.D. PLUMBLEY. Extension of sparse, adaptive signal decompositions to semi-blind audio source separation. In *Proc. 8^{th} Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 605–612, 2009.

[73] M. PUIGT, E. VINCENT AND Y. DEVILLE. Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources. In *Proc. 8^{th} Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 613–620, 2009.

[74] R. SCHOLZ, E. VINCENT AND F. BIMBOT. Robust modeling of musical chord sequences using probabilistic N-grams. In *Proc. 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 53–56, 2009.

[75] E. VINCENT, S. ARBERET AND R. GRIBONVAL. Underdetermined instantaneous audio source separation via local Gaussian modeling. In *Proc. 8^{th} Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 775–782, 2009.

[76] M. KOWALSKI, E. VINCENT AND R. GRIBONVAL. Under-determined source separation via mixed-norm regularized minimization. In *Proc. 16^{th} European Signal Processing Conference (EUSIPCO)*, 2008.

[77] A. NESBIT, M.D. PLUMBLEY AND E. VINCENT. Oracle evaluation of flexible adaptive transforms for underdetermined audio source separation. In *Proc. UK ICA Research Network International Workshop*, 2008.

[78] E. VINCENT, N. BERTIN AND R. BADEAU. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proc. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 109–112, 2008 (**55 citations**).

[79] E. VINCENT AND R. GRIBONVAL. Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 110–113, 2007.

[80] E. VINCENT. Complex nonconvex ℓ_p norm minimization for underdetermined source separation. In *Proc. 7^{th} Int. Conf. on Independent Component Analysis and Signal Separation*

(ICA), pp. 430–437, 2007 (**29 citations**).

[81] S. WELBURN, M.D. PLUMBLEY AND E. VINCENT. Object-coding for resolution-free musical audio. In *Proc. 31st AES Int. Conf. on New Directions in High Resolution Audio*, 2007.

[82] P. LEVEAU, E. VINCENT, G. RICHARD AND L. DAUDET. Mid-level sparse representations for timbre identification: design of an instrument-specific harmonic dictionary. In *Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS)*, pp. 1–11, 2006.

[83] M.D. PLUMBLEY, S.A. ABDALLAH, T. BLUMENSATH, M.G. JAFARI, A. NESBIT, E. VINCENT AND B. WANG. Musical audio analysis using sparse representations. In *Proc. 17th Symposium on Computational Statistics (COMPSTAT)*, pp. 104–117, 2006.

[84] E. VINCENT AND M.D. PLUMBLEY. Fast factorization-based inference for bayesian harmonic models. In *Proc. 2006 IEEE Int. Workshop on Machine Learning for Signal Processing*, pp. 117–122, 2006.

[85] E. VINCENT AND M.D. PLUMBLEY. A prototype system for object coding of musical audio. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 239–242, 2005.

[86] E. VINCENT AND X. RODET. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proc. 5th Int. Conf. on Music Information Retrieval (ISMIR)*, p. 576–581, 2004 (**37 citations**).

[87] E. VINCENT AND X. RODET. Underdetermined source separation with structured source priors. In *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, p. 327–332, 2004 (**26 citations**).

[88] E. VINCENT AND X. RODET. Music transcription with ISA and HMM. In *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, p. 1197–1204, 2004 (**33 citations**).

[89] E. VINCENT, C. FÉVOTTE, R. GRIBONVAL, L. BENAROYA, X. RODET ET AL.. A tentative typology of audio source separation tasks. In *Proc. 4th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, p. 715–720, 2003 (**29 citations**).

[90] R. GRIBONVAL, L. BENAROYA, E. VINCENT AND C. FÉVOTTE. Proposals for performance measurement in source separation. In *Proc. 4th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, p. 763–768, 2003 (**121 citations**).

B.6 Peer-reviewed papers in national conferences

[91] A. BENICHOUX, E. VINCENT AND R. GRIBONVAL. Optimisation convexe pour l'estimation simultanée de réponses acoustiques. In *Actes du 23e Colloque GRETSI*, article ID 132, 2011.

[92] N. ITO, E. VINCENT, N. ONO, R. GRIBONVAL AND S. SAGAYAMA. Diffuse noise robust multiple source localization based on matrix completion via trace norm minimization. In *Proc. ASJ Spring Meeting*, 2011.

[93] S.A. RACZYŃSKI, E. VINCENT AND S. SAGAYAMA. Dynamic Bayesian networks for

symbolic polyphonic pitch modeling. In *Proc. 91st IPSJ Special Interest Group on MUSic and computer (SIGMUS) Meeting*, article ID 2011-MUS-91, 2011.

[94] F. BIMBOT, O. LE BLOUCH, G. SARGENT, E. VINCENT. Décomposition en blocs autonomes comparables - Une proposition de description et d'annotation de structure pour le traitement automatique des morceaux de musique. In *Actes des Journées d'Informatique Musicale (JIM)*, pp. 187–196, 2010.

[95] N. ITO, E. VINCENT, N. ONO, R. GRIBONVAL AND S. SAGAYAMA. Diffuse noise robust multiple source localization based on noise reduction in covariance matrix domain. *IEICE Technical Report*, vol. 110, p. 31–36, 2010.

[96] J. LE ROUX, E. VINCENT, Y. MIZUNO, H. KAMEOKA, N. ONO AND S. SAGAYAMA. Consistent Wiener filtering: designing generalized time-frequency masks respecting spectrogram consistency. In *Proc. ASJ Spring Meeting*, 2010.

[97] G. SARGENT, F. BIMBOT AND E. VINCENT. Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique. In *Actes des Journées d'Informatique Musicale (JIM)*, pp. 177–186, 2010.

[98] V. EMIYA, E. VINCENT AND R. GRIBONVAL. Estimateurs oracles pour la séparation de sources monocapteur par approches spectrales à états discrets. In *Actes du 22^e colloque GRETSI sur le traitement du signal et des images*, 2009.

[99] J. LE ROUX, H. KAMEOKA, E. VINCENT, N. ONO, K. KASHINO AND S. SAGAYAMA. Complex NMF under spectrogram consistency constraints. In *Proc. ASJ Autumn Meeting*, 2009.

[100] E. VINCENT, M.G. JAFARI AND M.D. PLUMBLEY. Preliminary guidelines for subjective evalutation of audio source separation algorithms. In *Proc. UK ICA Research Network Workshop*, 2006.

[101] M.G. JAFARI, E. VINCENT, S.A. ABDALLAH, M.D. PLUMBLEY AND M.E. DAVIES. Blind source separation of convolutive audio using an adaptive stereo basis. In *Proc. UK ICA Research Network Workshop*, 2006.

[102] E. VINCENT AND R. GRIBONVAL. Construction d'estimateurs oracles pour la séparation de sources. In *Actes du 20^e colloque GRETSI sur le traitement du signal et des images*, p. 1245–1248, 2005.

B.7 Extended abstracts

[103] G. SARGENT, F. BIMBOT AND E. VINCENT. A music structure inference algorithm based on morphological analysis. In *Proc. 8th Music Information Retrieval Evaluation eXchange (MIREX)*, 2012.

[104] G. SARGENT, S.A. RACZYŃSKI, F. BIMBOT, E. VINCENT AND S. SAGAYAMA. A music structure inference algorithm based on symbolic data analysis. In *Proc. 7th Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

[105] G. SARGENT, F. BIMBOT AND E. VINCENT. A structural segmentation of songs using

generalized likelihood ratio under regularity assumptions. In *Proc. 6^{th} Music Information Retrieval Evaluation eXchange (MIREX)*, 2010.

[106] N. BERTIN, E. VINCENT AND R. BADEAU. Fast Bayesian constrained NMF for polyphonic pitch transcription. In *Proc. 5^{th} Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[107] E. VINCENT, N. BERTIN AND R. BADEAU. Two nonnegative matrix factorization methods for polyphonic pitch transcription. In *Proc. 3^{rd} Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[108] C. SUTTON, E. VINCENT, M.D. PLUMBLEY AND J.P. BELLO. Transcription of vocal melodies using voice characteristics and algorithm fusion. In *Proc. 2^{nd} Music Information Retrieval Evaluation eXchange (MIREX)*, 2006.

[109] E. VINCENT AND M.D. PLUMBLEY. Predominant-F0 estimation using Bayesian harmonic waveform models. In *Proc. 1^{st} Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.

B.8 Theses

[110] E. VINCENT. Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux. PhD thesis, Université Paris 6, 2004 (**21 citations**).

[111] E. VINCENT. Séparation de signaux audio: principes de l'analyse en composantes indépendantes et applications au signal monophonique. MSc thesis, IRCAM, 2001.

[112] R. JOLY AND E. VINCENT. L'équation de Smoluchowski: existence de solutions et phénomène de gélation. BSc thesis, École Normale Supérieure, 2000.

B.9 Technical reports

[113] E. VINCENT, M.G. JAFARI, S.A. ABDALLAH, M.D. PLUMBLEY AND M.E. DAVIES. Blind Audio Source Separation. Technical report C4DM-TR-05-01, Queen Mary, University of London, 2005 (**21 citations**).

[114] C. FÉVOTTE, R. GRIBONVAL AND E. VINCENT. BSS_EVAL Toolbox User Guide. Publication Interne 1706, IRISA, 2005 (**94 citations**).

B.10 Patents

[115] J. LE ROUX, H. KAMEOKA, E. VINCENT, Y. MIZUNO, N. ONO AND S. SAGAYAMA. Apparatus, method, and program for signal separation, patent JP 2010-35052, filed 19/02/2010.

B.11 Software

C. BLANDIN, E. VINCENT AND A. OZEROV. BSS Locate: Matlab toolbox for source localization, http://www.irisa.fr/metiss/bss_locate/, GPL, 2011.

A. OZEROV, E. VINCENT AND F. BIMBOT. FASST: Matlab toolbox for flexible audio source separation, http://www.irisa.fr/metiss/fasst/, GPL, 2011.

E. VINCENT. Quadratic ERB-scale time-frequency transform and multichannel filtering in the ERB-scale time-frequency domain, http://www.irisa.fr/metiss/members/evincent/software, GPL, 2010.

E. VINCENT AND N.Q.K. DUONG. Under-determined audio source separation by local Gaussian modeling for instantaneous and convolutive mixtures, http://www.irisa.fr/metiss/members/evincent/software, GPL, 2010.

K. KOWALSKI, E. VINCENT AND R. GRIBONVAL. URSS: under-determined convolutive audio source separation by wideband convex optimization, http://www.lss.supelec.fr/perso/kowalski/downloads/URSS.zip, CeCILL, 2010.

E. VINCENT. Multiple pitch estimation and note tracking using NMF under harmonicity and spectral smoothness constraints, http://www.irisa.fr/metiss/members/evincent/software, GPL, 2009.

E. VINCENT. Roomsimove: Matlab toolbox forthe computation of simulated room impulse reponses for moving sources, http://www.irisa.fr/metiss/members/evincent/software, GPL, 2008.

E. VINCENT, S. ARAKI AND P. BOFILL. SiSEC Reference Software: Matlab toolbox including reference algorithms for audio source separation, http://sisec.wiki.irisa.fr/, GPL, 2008 (**200+ downloads**).

E. VINCENT. BSS Eval 3.0: Matlab toolbox for the evaluation of source separation algorithms, http://members/evincent/bss_eval/, GPL, 2008 (**250+ downloads**).

E. VINCENT, R. GRIBONVAL AND M.D. PLUMBLEY. BSS Oracle 2.1: Matlab toolbox for the computation of theoretical performance bounds for audio source separation, http://www.irisa.fr/metiss/bss_oracle/, GPL, 2007.

E. VINCENT AND M.E.P. DAVIES. Music object-based coding and decoding: software for very low bit-rate coding of musical audio based on Bayesian harmonic models, http://www.elec.qmul.ac.uk/digitalmusic/objectcoding/obc-matlab.zip, 2006.

E. VINCENT. MUSHRAM 1.0: Matlab interface for the conduction of listening tests according to the MUSHRA standard, http://www.elec.qmul.ac.uk/digitalmusic/downloads/#mushram, GPL, 2005 (selected for the *Digital Audio eFfects Transformation Rating* (DAFxTRa 2008) initiative).

B.12 Data

M. DESNOUES, J.-L. DURRIEU, T. FILLON, O. LE BLOUCH, G. RICHARD AND E. VINCENT. QUASI Database: Corpus of 11 multitrack songs, each mixed in several different ways by a professional sound engineer, for the evaluation of audio source separation, http://www.tsi.telecom-paristech.fr/aao/?p=605, Creative Commons, 2012.

E. VINCENT, H. SAWADA, S. ARAKI AND P. BOFILL. Stereo Audio Source Separation Database: Corpus of 112 synthetic or recorded mixtures of audio sources for the evaluation of audio source separation, http://sisec.wiki.irisa.fr/, Creative Commons, 2008 (selected for the *1^{st} Signal Separation Evaluation Campaign* (SiSEC 2008)).

Appendix C

Résumé des contributions

Les données audio occupent une place centrale dans notre vie: communication parlée, vidéos personnelles, radio-télévision, musique, cinéma, jeux, spectacle vivant. Cela engendre une multitude de besoins applicatifs, allant du rehaussement du signal à la recherche d'information en passant par la réditorialisation et la manipulation interactive des contenus. Ces données ont souvent une structure complexe due à la présence simultanée de plusieurs sources sonores et/ou niveaux d'information. Mes travaux concernent la conception de modèles et d'algorithmes pour la séparation des signaux sources et l'extraction d'information. Ces travaux reposent sur les outils de modélisation et d'estimation bayésienne d'une part et de représentation parcimonieuse et d'optimisation convexe d'autre part.

Ce document résume les travaux effectués depuis la fin de ma thèse et indépendamment de celle-ci selon quatre axes: la formalisation et l'évaluation diagnostique des problèmes étudiés, la modélisation linéaire des signaux audio et les algorithmes associés, la modélisation de la variance des signaux audio et les algorithmes associés, et la description des contenus multi-sources et multi-niveaux. J'ai choisi d'adopter un style peu conventionnel mélangeant le "je" pour les travaux en tant que chercheur unique ou principal au "nous" pour les travaux en tant qu'encadrant ou collaborateur.

C.1 Problèmes, évaluation et diagnostic

Encadrement: Valentin Emiya (post-doctorant), Gabriel Sargent (doctorant)
Collaborations principales: Carl von Ossietzky Universität Oldenburg (Allemagne), NTT Communication Science Labs (Japon), Queen Mary University of London (Royaume-Uni)

C.1.1 Séparation de sources

La séparation de sources audio consiste à extraire les signaux d'une ou plusieurs sources sonores dans un enregistrement. Jusque récemment, la comparaison des résultats de différents algorithmes de séparation restait difficile car les signaux de test et (parfois involontairement) la définition du problème variaient d'un auteur à l'autre. C'est en organisant une des premières campagnes d'évaluation sur ce problème appelée *Stereo Audio Source Separation Evaluation*

Campaign (SASSEC) [35] que j'ai proposé une méthodologie de référence pour l'évaluation basée sur un ensemble de tâches à résoudre, de critères d'évaluation et de bornes de performance. Cette méthodologie a par la suite été reprise et complétée par la série des *Signal Separation Evaluation Campaigns* (SiSEC) que j'ai fondée puis co-organisée [10, 30].

En général, le signal multicanal de mélange $\mathbf{x}(t)$ peut toujours s'exprimer comme la somme des *images spatiales* $\mathbf{c}_j(t)$ des sources

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{C.1}$$

où $\mathbf{c}_j(t)$ est la partie du signal de mélange engendrée par la source j. Dans le cas particulier d'une source ponctuelle, cette image spatiale résulte du processus de convolution

$$\mathbf{c}_j(t) = \mathbf{a}_j \star s_j(t) \tag{C.2}$$

où $s_j(t)$ est le signal source monocanal correspondant et les coefficients de $\mathbf{a}_j(\tau)$ sont appelés *filtres de mélange*. Mais ce processus ne s'applique pas aux sources spatialement diffuses qu'il est impossible de représenter par un signal monocanal sans perte d'information. Le problème de séparation de sources se traduit donc par deux tâches différentes selon la nature ponctuelle ou non des sources et selon l'application: l'estimation des signaux sources $s_j(t)$ ou bien celle de leurs images spatiales $\mathbf{c}_j(t)$ [10, 35].

En m'inspirant des critères définis pendant ma thèse pour l'évaluation des signaux sources [24], j'ai défini un ensemble de critères d'évaluation de l'image spatiale estimée $\widehat{\mathbf{c}}_j(t)$ d'une source par rapport à un signal de référence $\mathbf{c}_j(t)$ supposé connu [10, 35]. Ces critères appelés rapport signal-à-distorsion (RSD), rapport image-à-distorsion spatiale, rapport signal-à-interférences et rapport signal-à-artefacts quantifient respectivement la distorsion totale ainsi que trois types de distorsion: la présence de filtrage fréquentiel ou spatial de la référence, la présence résiduelle d'autres sources et la présence d'artefacts de "bruit musical" introduits par l'algorithme de séparation. Leur calcul fait appel à la décomposition du signal estimé en quatre composantes correspondant à la référence et aux trois types de distorsion par projection orthogonale sur les sous-espaces engendrés par les références du signal cible et des autres sources.

Plus récemment, nous avons proposé un ensemble de critères perceptuels d'évaluation [11, 43]. Pour cela, nous avons conçu un protocole de test d'écoute dédié à la séparation de sources et collecté les scores attribués par 20 auditeurs à 80 sons de la campagne SiSEC en terme de qualité globale et de qualité associée à chaque type de distorsion. L'image spatiale estimée d'une source est décomposée en un ensemble de signaux localisés en temps et en fréquence avec une résolution comparable à celle de l'oreille et chacun de ces signaux est à son tour décomposé en quatre composantes comme précédemment. La saillance perceptuelle de chaque composante est quantifiée à l'aide de la mesure de similarité perceptuelle PEMO-Q, puis les quatre valeurs ainsi obtenues sont combinées par un réseau de neurones appris sur les scores subjectifs collectés. Les nouveaux critères ainsi obtenus accroissent la corrélation avec les scores subjectifs par rapport aux anciens critères.

La performance d'un algorithme de séparation dépend de plusieurs facteurs: la difficulté intrinsèque de séparation du signal traité, le choix d'un modèle sous-jacent, les contraintes sur

ce modèle (taille de fenêtre, nombre de paramètres, *etc*) et enfin l'algorithme d'estimation des paramètres du modèle. Afin de diagnostiquer l'importance des trois premiers facteurs par rapport au dernier, j'ai proposé un ensemble d'estimateurs *oracles* et d'algorithmes associés permettant de calculer les bornes supérieures de performance de certaines classes d'algorithmes sur un signal donné [22] puis nous avons étendu ces estimateurs à d'autres classes d'algorithmes [98]. L'utilisation de ces estimateurs a permis de montrer par exemple que le filtrage monocanal adaptatif (ou masquage temps-fréquence) induit une borne de performance de séparation très inférieure à celle du filtrage multicanal adaptatif et que, si la performance des algorithmes actuels en est proche dans le cas de mélanges dits instantanés, des progrès importants restent possibles dans le cas de mélanges réverbérants.

C.1.2 Estimation de la structure musicale

Un autre problème mal posé auquel nous nous sommes intéressés est l'estimation de la structure musicale. En musicologie, la structure d'un morceau découle d'un ensemble de règles de composition variant d'un genre musical à l'autre. En *music information retrieval*, la structure est au contraire considérée comme un concept purement subjectif variant d'un auditeur à l'autre. Cette multiplicité de points de vue empêche la comparaison entre algorithmes d'estimation reposant sur des points de vue différents. Nous avons proposé une définition opérationnelle du concept de structure musicale basée sur des axiomes issus du structuralisme, qui permet une annotation quasi-univoque de la structure pour un éventail de genres musicaux sans faire appel à aucune expérience musicologique. Cette approche baptisée décomposition en blocs autonomes et comparables [40, 47, 57] considère un morceau comme l'agencement régi par un processus d'assemblage dit *syntagmatique* d'un ensemble de blocs structurels comparables entre eux par des relations d'équivalence dites *paradigmatiques*. L'ensemble forme un *système* au sens structuraliste. Le morceau est une observation issue de ce système sous-jacent qu'il s'agit de retrouver. Nous avons validé sur une base de 20 morceaux que cette approche mène à une concordance entre annotateurs de 91% sur les frontières temporelles des blocs.

C.2 Modèles linéaires des signaux audio et algorithmes associés

Encadrement: Alexis Benichoux (doctorant), Pascal Bado (stagiaire master 2)
Collaborations principales: Université Paris 6 - LAM, Supélec - L2S (France), Queen Mary University of London (Royaume-Uni)

C.2.1 Principe général

Le paradigme classique de modélisation linéaire consiste à représenter les signaux dans une base (parfois sur- ou sous-déterminée) de signaux $\mathbf{\Phi}$ et à spécifier une certaine distribution *a priori* ou une certaine fonction de coût sur leurs coefficients dans cette base. Lorsque $\mathbf{\Phi}$ est elle-même définie par un paramétrage non-linéaire, on parle de modélisation linéaire généralisée.

C.2.2 Modélisation parcimonieuse locale

Dans le cadre de la séparation de sources, la base $\mathbf{\Phi}$ choisie est typiquement une Transformée de Fourier à Court Terme (TFCT) avec une fenêtre fixée. Sous une approximation de bande étroite valable pour une source ponctuelle faiblement réverbérée, la TFCT $\widetilde{\mathbf{c}}_j(n,f)$ de l'image spatiale d'une source est alors approximée par

$$\widetilde{\mathbf{c}}_j(n,f) \simeq \widetilde{s}_j(n,f)\,\widetilde{\mathbf{a}}_j(f) \tag{C.3}$$

où $\widetilde{\mathbf{a}}_j(f)$ est la transformée de Fourier des filtres de mélange et $\widetilde{s}_j(n,f)$ la TFCT du signal source correspondant. Dans le cas d'un mélange instantané, une Transformée en Cosinus Discrète Modifiée (TCDM) peut aussi être utilisée.

Les algorithmes classiques de séparation exploitent la parcimonie des coefficients de TFCT des sources en appliquant un masquage binaire ou en minimisant leur norme ℓ_1 [25, 27]. J'ai étendu cette approche à la minimisation de la norme ℓ_p des coefficients de TFCT des sources pour $p < 2$ [80]. Il s'agit d'un problème de minimisation non-convexe pour lequel les résultats théoriques préexistant pour des données à valeurs réelles ne s'appliquent pas aux données complexes. J'ai caractérisé les minima locaux de la norme ℓ_p dans un cas particulier et déduit un algorithme pour l'estimation du minimum global. Cet algorithme a obtenu la meilleure performance sur les mélanges instantanés de la campagne d'évaluation SASSEC [35].

En parallèle, j'ai proposé un algorithme de sélection de la meilleure base $\mathbf{\Phi}$ parmi une bibliothèque de bases de paquets de cosinus dyadique [79]. Chaque base a une structure similaire à la TCDM, mais où la taille de fenêtre varie au cours du temps en suivant une partition dyadique de l'axe temporel. L'algorithme proposé trouve la partition qui minimise un critère de recouvrement des sources, ce qui permet de mieux les séparer. Nous avons généralisé cette idée en relâchant la contrainte de partition dyadique d'une part [26] et en apprenant une base de signaux multicanaux pour la représentation parcimonieuse du signal de mélange d'autre part [18]. Ces algorithmes ont permis d'obtenir une amélioration modeste de la qualité de séparation, tout en suggérant que la modélisation séparée du processus de mélange et des signaux sources est nécessaire pour progresser plus avant.

C.2.3 Modélisation à large bande des filtres de mélange

En ce qui concerne la modélisation du processus de mélange, nous avons proposé de remplacer l'approximation de bande étroite (C.3) par le modèle exact à large bande (C.2) et l'égalité (C.1) par un terme d'attache aux données quadratique. Nous avons conçu un algorithme de séparation de sources basé sur ce modèle par minimisation de la norme ℓ_1 des coefficients de TFCT des sources par seuillage itératif doux [16]. Cet algorithme alterne entre le calcul du gradient du terme d'attache aux données dans le domaine temporel et le seuillage des sources dans le domaine temps-fréquence. Il a permis une amélioration du RSD de 3 à 4 décibels (dB) par rapport à l'état de l'art sur des mélanges réverbérants de quatre sources de parole en supposant les filtres de mélange connus. Afin d'évoluer vers un scénario aveugle où les filtres de mélange sont inconnus, nous avons récemment proposé d'ajouter divers termes de pénalité sur les filtres dans le domaine temporel et évalué leur apport pour l'estimation des filtres de mélange en supposant les signaux sources connus [46].

C.2.4 Modélisation sinusoïdale harmonique des signaux sources

Concernant la modélisation des signaux sources, je me suis focalisé sur la modélisation sinusoïdale harmonique des mélanges de signaux périodiques. Cette approche consiste à approximer le signal au sein d'une trame temporelle par une somme de signaux périodiques, chacun constitué de sinusoïdes à des fréquences harmoniques paramétrées par leur amplitude et leur phase et par la fréquence fondamentale. Dans une série de travaux, j'ai adopté un point de vue bayésien en proposant des distributions *a priori* adaptées pour ces paramètres et en concevant un algorithme efficace pour l'estimation de la probabilité *a posteriori* des fréquences fondamentales qui requiert une intégration de grande dimension sur les paramètres d'amplitude et de phase [20]. J'ai appliqué ces travaux au codage des signaux musicaux sous forme d'objets sonores périodiques et montré que ce codage permet à la fois la manipulation du signal (changement de hauteur, de durée ou de timbre des notes par exemple) et un gain significatif de qualité par rapport au codage MPEG-1 Layer 3 pour la compression à très bas débit [21].

Dans une autre série de travaux, nous avons exploré l'utilisation de représentations parcimonieuses. Le signal est représenté au sein de chaque trame par une combinaison linéaire d'atomes harmoniques appris sur des notes isolées de divers instruments. Les atomes correspondants au signal de test sont sélectionnés par l'algorithme de *Matching Pursuit* puis des contraintes de structure sont utilisées pour extraire des suites d'atomes de fréquence fondamentale similaire. Ces travaux ont été appliqués à l'estimation de hauteurs multiples et à l'identification des instruments dans des signaux musicaux [19].

C.3 Modèles de variance des signaux audio et algorithmes associés

Encadrement: Alexey Ozerov, Laurent Simon, Joachim Thiemann (post-doctorants), Ngoc Duong, Nobutaka Ito (doctorants), Charles Blandin (stagiaire de master 2)
Collaborations: Télécom ParisTech - LTCI (France), University of Tokyo, NTT Communication Science Labs (Japon)

Si la modélisation linéaire des signaux audio permet une représentation fidèle des sources périodiques et ponctuelles, elle nécessite un coût de calcul important et ne s'applique pas naturellement aux sources non périodiques et/ou diffuses.

C.3.1 Principe général

Une approche classique pour y remédier consiste à modéliser non plus les signaux sources $s_j(t)$ eux-mêmes mais leur spectre de puissance à court terme $|\widetilde{s}_j(n,f)|^2$, tout en conservant l'approximation de bande étroite (C.3). En termes statistiques, cela correspond à représenter les coefficients de TFCT des signaux sources par une distribution invariante par rotation de phase et à modéliser la variance de cette distribution. J'ai généralisé cette idée aux sources réverbérées ou diffuses en proposant de modéliser non plus les signaux sources mais leurs images spatiales par une telle distribution [28]. Sous l'hypothèse d'une distribution gaussienne de moyenne nulle, la

matrice de covariance $\mathbf{R}_{\mathbf{c}_j}(n,f)$ de l'image spatiale de la source j peut se factoriser comme

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\,\mathbf{\Sigma}_j(f) \tag{C.4}$$

où $v_j(n,f)$ est la *variance spectrale* de la source représentant son contenu spectral et $\mathbf{\Sigma}_j(f)$ sa *matrice de covariance spatiale* représentant sa position et son étendue spatiale. Les paramètres $v_j(n,f)$ et $\mathbf{\Sigma}_j(f)$ sont alors estimés à partir de la matrice de covariance empirique $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ du signal observé, permettant ainsi d'exploiter la corrélation entre ses canaux [59, 75].

C.3.2 Modélisation et estimation des matrices de covariance spatiale

En ce qui concerne les matrices de covariance spatiale $\mathbf{\Sigma}_j(f)$, l'approximation classique de bande étroite équivaut à supposer que les canaux des images spatiales des sources sont parfaitement corrélés et que $\mathbf{\Sigma}_j(f)$ est une matrice de rang 1 égale à $\widetilde{\mathbf{a}}_j(f)\widetilde{\mathbf{a}}_j(f)^H$. Dans le cas de sources diffuses ou réverbérées, cette approximation n'est pas valable car les canaux des images spatiales sont partiellement ou totalement décorrélés. Dans une série de travaux, nous avons proposé de considérer $\mathbf{\Sigma}_j(f)$ comme une matrice de rang plein non contrainte [15]. Nous en avons déduit un algorithme Espérance-Maximisation (EM) pour l'estimation de $v_j(n,f)$ et $\mathbf{\Sigma}_j(f)$ au sens du Maximum de Vraisemblance (MV) et montré une amélioration du RSD de l'ordre de 1 dB pour la séparation de mélanges réverbérants de trois sources de parole par rapport à des algorithmes classiques basés sur l'approximation de bande étroite. Nous avons ensuite étendu cet algorithme à l'estimation au sens du Maximum *A Posteriori* (MAP) de $\mathbf{\Sigma}_j(f)$ avec une distribution *a priori* appropriée sachant la position spatiale des sources et obtenu une amélioration d'1 dB supplémentaire [2, 50]. En raison de la sensibilité de l'algorithme EM à la position spatiale initiale estimée des sources, nous avons aussi effectué une comparaison expérimentale à grande échelle des algorithmes de localisation de sources multiples et proposé de nouveaux algorithmes plus performants dans le cas de microphones faiblement espacés [7].

En parallèle, nous avons conçu une famille de modèles plus spécifique aux bruits diffus représentant la matrice de covariance spatiale $\mathbf{\Sigma}_j(f)$ du bruit dans une base matricielle de faible dimension. Dans le cas particulier de mélanges d'une source ponctuelle et de bruit diffus, les coefficients de $\mathbf{\Sigma}_j(f)$ dans cette base peuvent s'estimer par des techniques de complétion de matrice [95] ou de minimisation de la norme trace [92]. Cette approche a été appliquée au débruitage de signaux de parole et à la localisation de sources multiples dans un environnement bruité.

C.3.3 Modélisation par factorisation des spectres de puissance à court terme

En ce qui concerne la variance spectrale $v_j(n,f)$ des sources, une approche classique adoptée entre autres durant ma thèse [23] consiste à la représenter comme la somme de spectres de base pondérés par des coefficients d'activation variant dans le temps. Ce modèle dit de factorisation matricielle positive manque de flexibilité, dans la mesure où les spectres de base sont soit appris sur des signaux d'apprentissage séparés, avec le risque que les caractéristiques de ces signaux correspondent mal à celles du signal de test, soit adaptés au signal de test sans aucune contrainte, avec un risque de sur-apprentissage. J'ai montré comment intégrer une contrainte d'harmonicité

et de régularité des spectres en les factorisant comme la somme de spectres harmoniques à bande étroite fixés pondérés par des coefficients représentant l'enveloppe spectrale estimés sur le signal de test [17]. Ce modèle a débouché sur un algorithme d'estimation de hauteurs multiples classé deuxième pour une sous-tâche de la campagne d'évaluation *Music Information Retrieval Evaluation eXchange* (MIREX) 2007 [78] et complété par la suite par un *a priori* de régularité sur les coefficients d'activation temporelle [14].

Nous avons ensuite généralisé cette idée à la factorisation de spectres non harmoniques [12] puis à la factorisation flexible de la variance spectrale en un produit de huit variables représentant la structure fine spectrale, l'enveloppe spectrale, l'enveloppe temporelle et la structure fine temporelle du signal d'excitation et de la résonance [8]. Ce modèle englobe un certain nombre de modèles existants et permet de concevoir de nouveaux modèles incorporant les informations *a priori* disponibles sur les sources. L'estimation de ses paramètres au sens du MV repose sur des règles de mise à jour multiplicatives, dont nous avons analysé la convergence dans [13]. Nous avons aussi commencé à revisiter ces règles de mise à jour pour l'estimation en ligne des paramètres dans le cadre d'une application à la séparation de sources en temps réel [42].

C.3.4 Réduction des artefacts

Une fois les paramètres du modèle gaussien (C.4) estimés, la séparation de sources est effectuée au sens du MV par filtrage de Wiener multicanal adaptatif. Afin de réduire les artefacts générés par ce filtrage, nous avons étudié l'usage d'un lissage temporel préalable de la variance spectrale des sources [67] et proposé une méthode de prise en compte de la redondance de la TFCT dans le calcul du filtre [3, 63] qui a donné lieu au dépôt d'un brevet [115].

C.4 Description des contenus multi-sources et multi-niveaux

Encadrement: Kamil Adiloğlu, Stanisław Raczyński (post-doctorants), Gabriel Sargent (doctorant), Christopher Sutton, Antoine Movschin, Ricardo Scholz, Christophe Hauser (stagiaires de master 2)
Collaboration principale: IRCAM - STMS (France), University of Tokyo (Japon)

Au-delà des tâches de bas ou de moyen niveau comme la séparation de sources et l'estimation de hauteurs multiples, j'ai débuté un axe de recherche sur la description de plus haut niveau des contenus audio en tenant compte de la présence simultanée de plusieurs sources sonores et/ou niveaux d'information.

C.4.1 Vers une description robuste des contenus multi-sources

La robustesse à la présence de plusieurs sources sonores est indispensable par exemple pour la reconnaissance de la parole dans un environnement bruité ou celle du chanteur dans un enregistrement musical polyphonique. Si l'usage de la séparation de sources en tant que prétraitement peut améliorer la reconnaissance [51], cette approche n'est pas robuste aux distorsions des sources estimées. Une meilleure approche consiste à estimer l'incertitude sur les sig-

naux sources, représentée par leur distribution *a posteriori*, et à la propager aux descripteurs puis au classifieur. Dans ce cadre existant, nous avons formalisé l'estimateur bayésien exact de l'incertitude et conçu un algorithme pratique d'estimation pour le modèle de sources [8] basé sur une approximation variationnelle bayésienne [1, 39]. Nous avons aussi montré comment exploiter l'incertitude pour l'apprentissage des classifieurs directement sur des signaux multi-sources et montré son impact pour des tâches de reconnaissance du locuteur dans un environnement domestique bruité [6] ou de reconnaissance du chanteur dans la musique polyphonique [41]. Afin de promouvoir les travaux sur ce sujet, j'ai par ailleurs co-organisé le *PASCAL CHiME Speech Separation and Recognition Challenge* [5].

C.4.2 Vers une modélisation multi-niveaux du langage musical

Alors que la modélisation du langage parlé est aujourd'hui un domaine mûr, la modélisation du langage sous-jacent à la musique a été très peu étudiée. Plusieurs défis s'ajoutent, en particulier l'existence de dépendances entre informations à plusieurs échelles temporelles (temps, mesure, bloc structurel, morceau complet) et à plusieurs niveaux (signal, notes, accords, style musical) et le manque de corpus de taille significative regroupant des informations à plusieurs niveaux [66]. Après un travail préalable illustrant l'applicabilité à la musique des techniques de lissage de modèles issues du traitement de la parole [74], nous avons exploré la combinaison de modèles multiples par interpolation entre les probabilités correspondantes [4, 65]. Ce travail a permis d'obtenir le premier modèle de musique polyphonique à notre connaissance rendant compte à la fois de la structure "horizontale" et "verticale" de la musique.

Nous avons enfin mis à profit nos travaux sur la formalisation du concept de structure musicale pour construire une famille d'algorithmes d'estimation de la structure musicale [53]. La segmentation en blocs est effectué par un algorithme de Viterbi prenant en compte différents objectifs (répétition des blocs, rupture entre blocs, événements isolés entre blocs, structure interne des blocs) ainsi qu'une contrainte de régularité temporelle. L'étiquetage des blocs est ensuite réalisé en estimant le meilleur système sous-jacent au morceau par une technique de sélection de modèle. Cette approche a obtenu de bons résultats à la campagne MIREX 2011.

C.5 Conclusion et perspectives

C.5.1 Réalisations

En résumé, mes travaux portent sur la résolution des défis posés par la structure multi-sources et multi-niveaux des données audio. Les problèmes couverts portent à la fois sur le traitement du signal et sur la description des contenus et ils reposent sur des fondements théoriques interdisciplinaires. Lorsque nécessaire, j'ai aussi contribué à ces fondements en proposant des outils théoriques applicables en dehors de l'audio. La séparation de sources est maintenant un domaine de recherche mûr et la plupart de mes travaux sur la modélisation de variance ont été tranférés ou sont en cours de transfert vers des entreprises par le biais de collaborations ou de brevets. La séparation de mélanges réverbérants, dynamiques, ou bruités requiert cependant toujours des recherches. En tant qu'application phare du traitement de la parole, je me suis intéressé à la

reconnaissance robuste de la parole. La plupart des techniques actuelles reposent sur des heuristiques et, dans ce contexte, le cadre proposé d'estimation bayésienne de l'incertitude apparaît comme prometteur. La modélisation multi-niveaux de la musique reste quant à elle un problème difficile dont nous n'avons pour l'instant abordé que la surface.

C.5.2 Directions

Mon programme de recherche vise à poursuivre les efforts dans ces directions afin de proposer une approche unifiée à ces différents besoins applicatifs intégrant traitement du signal et description du contenu. Le formalisme bayésien offre un cadre théorique particulièrement approprié, permettant de combiner des sous-modèles de façon rigoureuse et d'obtenir des algorithmes naturellement modulaires et robustes aux données manquantes ou incertaines.

En ce qui concerne la modélisation acoustique des sources, les travaux effectués en parallèle jusqu'à présent restent à rassembler en un modèle unifié combinant les avantages des modèles linéaires pour les sources périodiques ou ponctuelles à ceux des modèles de variance pour les sources non périodiques ou diffuses. L'intégration de modèles de l'enveloppe spectrale des sources apparaît aussi nécessaire afin de permettre une interaction à double sens entre modélisation acoustique d'un côté et reconnaissance du locuteur et de la parole de l'autre. Enfin, la modélisation des mouvements des sources reste une question relativement ouverte. En vue d'une intégration totale à terme, je souhaite développer progressivement mes travaux sur la modélisation bayésienne du langage. L'utilisation d'*a priori* de parcimonie est une piste prometteuse pour atteindre la performance des modèles de l'état de l'art par réseaux de neurones.

Contrairement à l'estimation au sens du MAP préférée aujourd'hui pour sa simplicité, l'estimation bayésienne fournit une mesure d'incertitude sur les résultats par le biais de la distribution *a posteriori* des variables considérées. La conception d'algorithmes d'estimation bayésienne passant à l'échelle est donc essentielle. La théorie de l'approximation variationnelle bayésienne constitue un socle algorithmique solide, mais sa mise en œuvre pose plusieurs difficultés. Une première difficulté consiste à trouver une approximation de la distribution visée à la fois bonne et facile à estimer. L'utilisation d'approximations complexes de type *structured mean field* ou *mixture mean field* est une piste. Une deuxième difficulté concerne l'initialisation des variables pour éviter les maxima locaux. L'utilisation conjointe de plusieurs modèles estimés à partir d'initialisations aléatoires différentes sera explorée. Enfin, le problème de la sélection de modèle a été très peu étudié pour l'audio.

Les travaux ci-dessus seront appliqués principalement au rehaussement et à la reconnaissance robuste de la parole dans les enregistrements et les flux audiovisuels. L'apport de l'approche bayésienne sera validé en terme de robustesse à la présence de plusieurs sources et au choix de la dimension du modèle, mais aussi par sa capacité à inférer et exploiter des modèles complexes à partir d'une faible quantité de données.

Bibliography

[AAG07] A. Aissa-El-Bey, K.A. Abed-Meraim, and Y. Grenier. Blind separation of underdetermined convolutive mixtures using their time-frequency representation. *IEEE Transactions on Audio Speech and Language Processing*, 15(5):1540–1550, 2007.

[AK11] R. F. Astudillo and D. Kolossa. Uncertainty propagation. In D. Kolossa and R. Häb-Umbach, editors, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, chapter 3, pages 35–64. Springer, 2011.

[AN11] S. Araki and T. Nakatani. Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel Wiener filter. In *Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 225–228, 2011.

[AP04] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[BBG06] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, 2006.

[BGB01] L. Benaroya, R. Gribonval, and F. Bimbot. Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Actes du 18e Colloque GRETSI*, pages 434–437, 2001.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BMK06] M. J. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In *Proc. 7th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 198–201, 2006.

[BOS08] A. Brutti, M. Omologo, and P. Svaizer. Comparison between different sound source localization techniques based on a real data collection. In *Proc. 2nd Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pages 69–72, 2008.

[BR01] R. Balan and J. P. Rosca. Statistical properties of STFT ratios for two channel systems and applications to blind source separation. In *Proc. 3rd Int. Conf. on*

Independent Component Analysis and Blind Signal Separation (ICA), pages 429–434, 2001.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BW01] M. S. Brandstein and D. B. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[Car98] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. 1998 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 1941–1944, 1998.

[Car01] J.-F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. 3rd Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 1–6, 2001.

[CBHD06] J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1218–1234, 2006.

[CCA98] S. Choi, A. Cichocki, and S. Amari. Flexible independent component analysis. In *Neural Networks for Signal Processing (NNSP 8)*, pages 83–92, 1998.

[CFG07] A. T. Cemgil, C. Fevotte, and S. J. Godsill. Variational and stochastic inference for bayesian source separation. *Digital Signal Processing*, 17:891–913, 2007.

[CG98] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[CH96] D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 158–168, 1996.

[CHR10] M. P. Cooke, J. R. Hershey, and S. J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech and Language*, 24:1–15, 2010.

[CJ10] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[Coh04] I. Cohen. Speech enhancement using a noncausal a priori SNR estimator. *IEEE Signal Processing Letters*, 11(9):725–728, 2004.

[CR05] G. Casella and C. P. Robert. *Monte Carlo Statistical Methods, 2nd Edition*. Springer, 2005.

[DA08] J. Droppo and A. Acero. Environmental robustness. In J. Benesty, M. Sondhi, and Y. Huang, editors, *Handbook of Speech Processing*, pages 653–680. Springer, 2008.

[DBC09] J. S. Downie, D. Byrd, and T. Crawford. Ten years of ISMIR: Reflections on challenges and opportunities. In *Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 13–18, 2009.

[DDA05] L. Deng, J. Droppo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Transactions on Speech and Audio Processing*, 13(3):412–421, 2005.

[Den11] L. Deng. Front-end, back-end, and hybrid techniques to noise-robust speech recognition. In D. Kolossa and R. Häb-Umbach, editors, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, chapter 4, pages 67–99. Springer, 2011.

[DGI06] M. Davy, S. J. Godsill, and J. Idier. Bayesian analysis of western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, 2006.

[Die00] T. G. Dietterich. Ensemble methods in machine learning. In *Proc. 1st Int. Workshop on Multiple Classifier Systems (MCS)*, pages 1–15, 2000.

[DKN+11] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura. Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pages 12–17, 2011.

[DM05] S. Doclo and M. Moonen. On the output SNR of the speech-distortion weighted multichannel Wiener filter. *IEEE Signal Processing Letters*, 12(12):809–811, 2005.

[DNW09] M. Delcroix, T. Nakatani, and S. Watanabe. Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. *IEEE Transactions on Audio, Speech and Language Processing*, 17(2):324–334, 2009.

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

[dS16] F. de Saussure. *Cours de linguistique générale*. 1916.

[EM84] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, 1984.

[EPSG08] Z. El Chami, A. D.-T. Pham, C. Servière, and A. Guerin. A new model based underdetermined source separation. In *Proc. 11th Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[FC05] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 78–81, 2005.

[FCC08] D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience*, 2008, 2008. Article ID 872425.

[FGKO10] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer music information retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):638–648, 2010.

[FSPZ07] B. Fox, A. Sabin, B. Pardo, and A. Zopf. Modeling perceptual similarity of audio signals for blind source separation evaluation. In *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 454–461, 2007.

[GB03] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, January 2003.

[GCD12] R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012.

[Gri03] R. Gribonval. Piecewise linear source separation. In *Proc. SPIE*, volume 5207 Wavelets: Applications in Signal and Image Processing X, pages 297–310, 2003.

[GRT03] T. Gustafsson, B. D. Rao, and M. Trivedi. Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Transactions on on Speech and Audio Processing*, 11(6):791–803, 2003.

[HBC10] M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[HK06] R. Huber and B. Kollmeier. PEMO-Q – a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):1902–1911, 2006.

[ISO05] ISO. Information technology—coding of audio-visual objects—part 3: Audio (iso/iec 14496-3:2005), 2005.

[ITU03] ITU. ITU-R Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems, 2003.

[JBE+03] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wootersn. The ICSI meeting corpus. In *Proc. 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 364 – 367, 2003.

[JJ98] T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 163–173. MIT Press, 1998.

[Jor82] B. Jorgensen. *Statistical Properties of the Generalized Inverse-Gaussian Distribution*. Springer, 1982.

[KAA+11] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J.P. da Silva Neto, and R. Martin. CHIME challenge: approaches to robustness using beamforming and uncertainty-of-observation techniques. In *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pages 6–11, 2011.

[KAHO10] D. Kolossa, R. F. Astudillo, E. Hoffmann, and R. Orglmeister. Independent component analysis and time-frequency masking for speech recognition in multitalker conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, Article ID 651420, 2010.

[KB03] W. Kellermann and H. Buchner. Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1278–1282, 2003.

[KD06] A.P. Klapuri and M. Davy. *Signal processing methods for music transcription*. Springer, 2006.

[Kla98] D. Klakow. Log-linear interpolation of language models. In *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP)*, volume 5, pages 1695–1699, 1998.

[Kla06] A. P. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 216–221, 2006.

[KT08] M. Kowalski and B. Torrésani. Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients. *Signal, Image and Video Processing*, 3(3):251–264, 2008.

[LCKL07] Y. Lin, J. Chen, Y. Kim, and D. D. Lee. Blind channel identification for speech dereverberation using $l1$-norm sparse learning. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 921–928, 2007.

[LG07] H. Liao and M. J. F. Gales. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In *Proc. 2007 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 389–392, 2007.

[LS01] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13 (NIPS)*, pages 556–562, 2001.

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.

[Mar04] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, 2004.

[MD10] M. Mauch and S. Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1280–1289, 2010.

[MDH+07] M. Mauch, S. Dixon, C. Harte, M. Casey, and B. Fields. Discovering chord idioms through Beatles and Real Book songs. In *Proc. 8th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 255 – 258, 2007.

[MIZ01] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1–4):1–24, 2001.

[MK97] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[MK00] D. Maiwald and D. Kraus. Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices. *IEE Proceedings on Radar, Sonar and Navigation*, 147:162–168, 2000.

[MLS07] S. Makino, T.-W. Lee, and H. Sawada, editors. *Blind speech separation*. Springer, 2007.

[MN01] K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. In *Proc. 3rd Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 722–727, 2001.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

[MRS96] P. J. Moreno, B. Raj, and R. M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 733 – 736, 1996.

[MV07] A. Mesaros and T. Virtanen. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proc. 8th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 375–378, 2007.

[NBJ10] R. Niazadeh, M. Babaie-Zadeh, and C. Jutten. An alternating minimization method for sparse channel estimation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 319–327, 2010.

[NM11] F. Nesta and M. Matassoni. Robust automatic speech recognition through on-line semi-blind source extraction. In *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pages 18–23, 2011.

[NO12] F. Nesta and M. Omologo. Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal correlation. In *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 222–230, 2012.

[OF10] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[PE07] G.E. Poliner and D.P.W. Ellis. A discriminative model for polyphonic piano transcription. *Eurasip Journal of Advances in Signal Processing*, 2007, 2007. Article ID 48317.

[PH00] D. Pearce and G. Hirsch. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP)*, pages 29 – 32, 2000.

[PI08] A. Pertusa and J.M. Iñesta. Multiple fundamental frequency estimation using Gaussian smoothness. In *Proc. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 105–108, 2008.

[PMK10] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proc. 11th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pages 625–636, 2010.

[RG04] J. Rosier and Y. Grenier. Unsupervised classification techniques for multipitch estimation. In *Proc. AES 116th Convention*, 2004. Paper number 6037.

[RHB07] S. Renals, T. Hain, and H. Bourlard. Recognition and understanding of meetings - the AMI and AMIDA projects. In *Proc. 2007 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 238–247, 2007.

[RJ93] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

[RK05] M. P. Ryynänen and A. P. Klapuri. Polyphonic music transcription using note event modeling. In *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 319 – 322, 2005.

[Ros03] J. Rosier. *Méthodes d'estimation de fréquences fondamentales multiples pour la séparation de signaux de parole et de musique*. PhD thesis, École Nationale Supérieure des Télécommunications, France, 2003.

[RS03] C. Raphael and J. Stoddard. Harmonic analysis with probabilistic graphical models. In *Proc. 4th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 177–181, 2003.

[SAM11] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):516–527, 2011.

[SAMM07] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1592–1604, 2007.

[SJ03] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Proc. 20th Int. Conf. on Machine Learning (ICML)*, pages 720–727, 2003.

[STS99] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proc. 1st Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pages 261–266, 1999.

[TH11] M. Togami and K. Hori. Multichannel semi-blind source separation via local Gaussian modeling for acoustic echo reduction. In *Proc. 19th European Signal Processing Conf. (EUSIPCO)*, pages 496–500, 2011.

[TKT+12] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga. Multichannel speech dereverberation and separation with optimized combination of linear and non-linear filtering. In *Proc. 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4057–4060, 2012.

[TY10] K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6:615–640, 2010.

[Vir06] T. Virtanen. Unsupervised learning methods for source separation. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, chapter 9, pages 267–296. Springer, 2006.

[VK02] T. Virtanen and A. P. Klapuri. Separation of harmonic sounds using linear models for the overtone series. In *Proc. 2002 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1757–1760, 2002.

[vR79] C.J. van Rijsbergen. *Information retrieval, 2nd Edition*. Butterworths, 1979.

[Wie00] W. Wiegerink. Variational approximations between mean field theory and the junction tree algorithm. In *Proc. 16th Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 626–633, 2000.

[WKSM07] S. Winter, W. Kellermann, H. Sawada, and S. Makino. MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007. Article ID 24717.

[YMB08] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Transactions on Signal Processing*, 56(5):1830–1839, 2008.

[YR04] Ö. Yılmaz and S. T. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[ZPBK01] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition in a signal dictionary. In S. Roberts and R. Everson, editors, *Independent Component Analysis : Principles and Practice*, pages 181–208. Cambridge Press, Cambridge, UK, 2001.