# An overview of informed audio source separation

Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, Gaël Richard

# AN OVERVIEW OF INFORMED AUDIO SOURCE SEPARATION

*Antoine Liutkus[1], Jean-Louis Durrieu[2], Laurent Daudet[1], Gaël Richard[3]*

[1] Institut Langevin, ESPCI ParisTech, Paris Diderot University, CNRS UMR 7587, Paris, France
[2] Signal Processing Laboratory (LTS5), EPFL, Lausanne, Switzerland
[3] Télécom ParisTech, Institut Mines-Télécom, Paris, France

## ABSTRACT

Audio source separation consists in recovering different unknown signals called sources by filtering their observed mixtures. In music processing, most mixtures are stereophonic songs and the sources are the individual signals played by the instruments, e.g. bass, vocals, guitar, etc. Source separation is often achieved through a classical generalized Wiener filtering, which is controlled by parameters such as the power spectrograms and the spatial locations of the sources. For an efficient filtering, those parameters need to be available and their estimation is the main challenge faced by separation algorithms. In the blind scenario, only the mixtures are available and performance strongly depends on the mixtures considered. In recent years, much research has focused on informed separation, which consists in using additional available information about the sources to improve the separation quality. In this paper, we review some recent trends in this direction.

## 1. INTRODUCTION

Audio source separation is the signal processing task which consists in recovering the constitutive sounds, called *sources*, of an observed *mixture*, which can be multichannel. In music signal processing, the sources to recover coincide with the different instrumental sounds and the mixture is most often stereophonic. This particular topic has many interesting applications, such as audio editing, extraction of sound samples, respatialization or upmixing. Hence, it has attracted much research in the last 20 years [2, 32] and is particularly difficult due to the complexities of both the mixing process [5] and the spatio-temporal behaviour of the sources [4].

When the sources are to be recovered using the mixtures only, i.e. without any other information, the separation is called *blind* and relies on a few very general assumptions. Among them, the most classical are to assume a linear mixing process and independent and identically distributed (i.i.d.) non-Gaussian sources. Indeed, the DARMOIS theorem guarantees that separation is feasible if more linear mixtures than sources are available [2], provided all sources are i.i.d and at most one of them is Gaussian. Hence, Independent Component Analysis (ICA) has focused on this scenario. If fewer mixtures than sources are available as is the case in music processing, other blind procedures proposed further assumptions, for instance a time-invariant spatial distribution of the sources over the mixture channels (e.g. the left-right location of the sources) [5]. In any case, all these approaches suppose that the number of sources is known.

Still, performance of blind source separation has proved to be highly dependent of the mixtures considered. In many cases, it is not acceptable for further professional use or for broad-audience remixing applications. This is mainly due to the high sensitivity of most algorithms to initialization and to the fact that real-world sources and mixing processes do not perfectly obey desired assumptions. In order to address these limitations, several authors have worked on using any available information about the sources for the separation. In this paper, we review some of these so-called *informed* source separation ideas and argue that they can be divided into model-based and signal-based side-information.

This paper is organized as follows. In section 2, we give some notations as well as the most common separation procedure, based on Wiener filtering of locally stationary Gaussian processes. We show that its parameters are the sources power spectral densities (PSD) and their spatial positions within the mixtures. In section 3, we show how some prior knowledge allowed many researchers to consider that the sources PSD lie in some low-dimensional subspace, permitting easier estimation of its parameter. In section 4, we show how some further information about the source signals can be used so as to strongly improve separation performance.

## 2. GENERALIZED WIENER FILTERING

### 2.1. Notations

In a musical processing context, the $J$ sources $s_j$ are assumed to be monophonic signals, corresponding to the instruments playing in the song, e.g. voice, bass and guitar. A common strategy [5, 22] to model the production of an $I$-channel mixture, e.g. stereophonic ($I = 2$), is to suppose that each monophonic source $s_j$, e.g. the voice, is spatialized so as to yield a corresponding $I$-channel *image* $y_j$. The mixture is then modeled as the sum of the images. In that context, the objective of source separation is either to recover the original source $s_j$ or their images $y_j$, given the mixtures.

It is common to process audio signals in the Short Term Fourier Transform (STFT) domain. In this domain, $s_j(f,n) \in \mathbb{C}$ is the STFT of source $j$ (e.g. guitar) at Time-Frequency (TF) bin $(f,n)$, whereas $\mathbf{x}(f,n)$ and $\mathbf{y}_j(f,n)$ are the $I \times 1$ vectors gathering the STFT coefficients of the mixtures (e.g. left and right channels) and of the image of source $j$, respectively, at TF bin $(f,n)$.

This work is partly supported by LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under reference ANR-10-IDEX-0001-02 PSL* and by the MetaMedia Center, EPFL. LD is on a joint affiliation with Institut Universitaire de France.

### 2.2. Gaussian model

When considering audio signals, a reasonable assumption is to assume local stationarity. More precisely, the sources images and mixtures are often assumed to be locally stationary Gaussian processes [16], which boils down to splitting them in overlapping *frames* of length of approximately $50$ms, assuming that all these frames are independent and Gaussian stationary processes.

Under these assumptions, it can be shown that all coefficients $s_j(f,n)$ of the sources STFTs are independent, circular and centered Gaussian random variables. We thus assume that $s_j(f,n) \sim \mathcal{N}_c(0, P(f,n,j))$ where $\mathcal{N}_c$ denotes the complex centered and circular distribution [6] and where the variance $P(f,n,j)$, under the stationarity assumption, is the Power Spectral Density (PSD). The model therefore states that, even if the phase of each complex STFT coefficient is random, its power $P(f,n,j)$ comes as a meaningful parameter indicating the strength of some source $j$ at some particular TF bin $(f,n)$. Strictly speaking, the STFT coefficients of each source are assumed independent *given the PSD* $P$.

The relationship between each source $j$ and its image is tackled by *mixing modeling*. One of the most powerful tractable linear model in this line of thought is the *diffuse* —or full rank— model, proposed by DUONG *et al.* in [5]. It assumes that the $I$ channels of one image are correlated and Gaussian for one given TF bin. This can be written:

$$\mathbf{y}_j(f,n) \sim \mathcal{N}_c(0, P(f,n,j)\,\mathbf{R}_j(f)),$$

where $\mathbf{R}_j(f)$, of dimension $I \times I$, is called the *spatial covariance matrix*, and encodes the correlation between the different channels of one image at frequency band $f$. Particular cases of this general model include instantaneous and narrowband-convolutive mixing, for which $R_j(f)$ is a rank-1 matrix [5]. This model is also called the Local Gaussian Model (LGM) by some authors [32, 22].

### 2.3. Separation procedure

Given the parameters $P(f,n,j)$ and $R_j(f)$ of the LGM, it can be shown that the minimum mean-squared estimate (MMSE) of one image $j_0$ given the mixture is given by [5, 16, 22]:

$$\hat{\mathbf{y}}_{j_0}(f,n) = P(f,n,j_0)\,\mathbf{R}_{j_0}(f)\left(\sum_{j=1}^{J} P(f,n,j)\,\mathbf{R}_j(f)\right)^{-1}\mathbf{x}(f,n). \quad (1)$$

The corresponding waveforms are then easily obtained through an inverse STFT transform. Estimation (1) is also called a multichannel generalized WIENER filter, incorporating information of both the sources PSD and spatial locations. It has many interesting special cases, such as the celebrated TF masking [1], or the Degenerate Unmixing Estimation Technique (DUET) [33].

Given the true PSD $P(f,n,j)$ of the sources and spatial covariances $\mathbf{R}_j(f)$, the performance of the MMSE estimate (1) is very good [31]. The main challenge faced by source separation methods then mainly lies in the estimation of these parameters.

## 3. MODEL-BASED SIDE-INFORMATION

When the sources are unknown, estimation of the LGM parameters needs to be achieved given available information only. In the typical *blind* scenario, this information reduces to the mixture only. Even if a Maximum Likelihood (ML) estimation is possible [5], it faces an issue: convergence to good estimates is hindered by too many local minima and degrees of freedom. In this section, we review some ways that were proposed to constraint this difficult process so as to yield meaningful estimates.

### 3.1. Source-specific PSD models

Considering musical sources, an important observation with far-reaching consequences is that instrumental signals most often exhibit a large amount of redundancies, both in time and frequency. More precisely, depending on the considered source, important prior knowledge may be available that permits to model the PSD $P(f,n,j)$ as depending only on few parameters.

A first widely studied model [1, 30, 19] assumes that the PSD of any given source $j$ can be understood as the modulation over time of one single spectral template $W(f,j)$, hence having: $P(f,n,j) = W(f,j)\,H(n,j)$, where both $W$ and $H$ are understood as nonnegative energy quantities. This Nonnegative Matrix Factorization (NMF) approach permits to strongly reduce the number of unknown parameters to estimate.

Notwithstanding its expressive power, even the NMF model may appear under-restrictive or inadequate to model some sound sources. In particular, some further prior information may be available, stating that a source exhibits a strong harmonic structure. For example, the vocal signal is known to be correctly modeled as a source-filter model. This prior knowledge may lead to another modeling of the source PSD $P(f,n,j)$ as a sum of harmonic templates, activated over time by some *piano-roll* latent representation [6, 11]. This kind of further *source-specific* model-based knowledge leads to improved source separation.

More generally, the unknown PSD $P(f,n,j)$ of the sources may exhibit some known latent structure, allowing them to be correctly modeled as a combination of low-rank nonnegative tensors. The general Probabilistic Latent Tensor Factorisation (PLTF) framework [34] permits to handle such models.

### 3.2. Bayesian prior distributions

Apart from pertaining to the particular model to use for the PSD, such as NMF, prior knowledge may be available about the values of its parameters. For instance, a musical source $j$ will usually not abruptly change several times per second. The NMF time activation parameter $H(n,j)$, ought to be *smooth* rather than completely unconstrained.

In a Bayesian setting, this kind of prior knowledge can be taken into account through *prior distributions* over the parameters [32]. This interesting line of research was proved to yield NMF decompositions that are more meaningful than their unconstrained counterparts [3]. Such ideas can be further generalized to incorporate many kinds of prior knowledge apart from mere smoothness and may easily account for a known periodicity, or time-varying lengthscales [27].

### 3.3. Sparse and dense sources

While some source PSDs present some kind of redundancy, other signals such as singing voice may be *sparse* in the Fourier domain, meaning that only a few of their coefficients are nonzero [24] and lack a low-rank structure. Recent studies [14] demonstrated impressive performance in the separation of vocals using such a *sparse+low-rank* model. In the same vein, the REpeating Pattern Extraction Technique (REPET) was proposed [25] that separates repeating background from sparse voice signals.

## 4. SIGNAL-BASED SIDE-INFORMATION

We previously discussed a number of strategies to exploit a specific model or prior distributions on its parameters to improve source separation. However, in some cases, extra data which is known to be related to the sources may be available and could be used to guide the separation process. In the so-called *signal-based* informed source separation scheme, the main challenge lies in correctly modeling the *interaction* between observed side-information and unknown sources.

### 4.1. Score-informed source separation

A first approach that was undertaken to improve the performance of audio source separation is to make use of available score-sheets, or MIDI files, that describe the musical content of the songs. Such *score-informed* source separation recently gathered some attention [26, 13, 9, 28].

The main principle behind most score-informed separation techniques is to make use of the onset/offset information found in MIDI files to correctly initialize the parameters of a parametric model. In the case of the NMF model described above in section 3.1, the score sheets permit to set to 0 the activation parameters $H(n, j)$ for one given source when it is known to be inactive. Such a simple procedure is shown in [9] to dramatically increase separation performance, by initializing the parameters to a sensible value, hence much closer to the global minimum sought for during optimization. Pitch information can also be used to initialize the spectral templates $W$, or to adequately drive comb-filters as in [26]. In the case of more flexible parametric models such as a NMF with time-varying spectral templates [13], score information may also be used with a noticeable gain in separation quality.

The main issue with such score-initialized audio decompositions is the requirement that MIDI files be *synchronized* with the audio mixtures. Even if efficient alignment techniques do exist to this purpose, mismatch in the alignment may lead to wrongly initialized decompositions, yielding poor separated sources. In a recent study, SIMSEKLI *et al.* [28] showed that MIDI information can actually be used without assuming such an alignment. The main fact underlying their technique is that apart from their temporal position, the score also contains information about co-occurrences of the notes as well as their pitch information. Even in case of misalignment, these may be supposed to be the same in the actual audio mixture. In practice, such co-occurrences are modeled as common factors in a Generalized Tensor Factorization framework [34], where both scores and audio are *jointly* analyzed.

At last, several works have also involved the user to guide the separation, by manually assigning the activations of the sources [20] or by choosing the source to separate thanks to a score-like representation estimated from the mixture [7, 10]. These works mainly demonstrate that these cues are beneficial to the separation, and therefore motivate further investigations into the challenging task of their automatic estimation.

### 4.2. Exemplar-based source separation

Whereas MIDI files may provide *symbolic* data that can help separation, there are cases where additional *audio recordings* are available, which are known to be related to the mixture to separate.

For example, *common signal separation* [17, 15] was proposed as a way to separate the music+effect track from surrounding music in a movie soundtrack. To this purpose, the music is assumed to be the same in several international versions of the same movie.

Another example of such signal-based informed source separation lies in *cover-informed* source separation [12], which gathered some attention recently. Its objective is to separate a stereo song into its constitutive instruments with the use of *cover versions*, for which the sources are available. In the method proposed in [12], the imitative sources are used to correctly initialize an NMF parametric model. We expect this kind of *exemplar-based* separation to benefit from recent advances of coupled factorization methods, just as score-informed separation in [28]. Alternatively, instead of a given cover of the target song, a user could also provide his own *cover*, for instance by singing the desired source to separate [29].

### 4.3. Oracle source separation and spatial object coding

Provided the right parameters are used for the separation procedure (1), separated signals are very good estimations of the original sources. Actually, the sources recovered through source separation may actually sound much better than those obtained through a conventional low-bitrate audio coder. This fact led some researchers to wonder whether source separation could actually be used in an audio coding framework.

In this framework, the sources to transmit are first analyzed jointly with their mixture in an *encoding* stage. This mixture may either be obtained automatically or produced by a professional sound engineer. During this joint analysis, a *small* side-information is computed and made available at the *decoder*, along with the mixtures. Decoding is then performed simply through separation of the mixture, using the pre-computed side-info parameters.

As can be seen, this workflow can either be envisioned as a particularly informed source separation procedure where the parameters are learned on the true source signals and ought to be encoded concisely, or as a spatial audio coder, where multichannel signals are recovered through respatialization of their downmix. In fact, it appears that two distinct communities worked on the same precise framework. On the one hand, Spatial Audio Object Coding (SAOC) emerged from the audio coding side [8], whereas Informed Source Separation (ISS) was studied by source separation researchers [23, 18]. Theoretical connections between ISS and source coding were made recently by OZEROV *et al.* that set ISS on information theoretic grounds [21].

## 5. CONCLUSION

Audio source separation is an extremely challenging task, especially when considering real-world stereophonic full-tracks. For this reason, it is clear that even if blind separation techniques do exist, their performance may be greatly

improved by making use of any available information apart from the mere mixture, leading to *informed source separation*.

In this paper, we have reviewed some of the most prominent research trends in this direction and have argued that they can roughly be divided into two main categories. First, *model-based* informed source separation permits to specialize the source and mixing models as well as adequately set priors over their parameters. Such an approach allows handling many kinds of prior information pertaining to *generative models* or to handle specific musicological knowledge. Second, *signal-based* informed source separation recently appeared as a desirable framework whenever some signals are available, such as score-sheets or cover version, which are related to the unknown sources to estimate. In that case, the main challenge is generally to adequately perform a joint analysis of both the mixture and these helper signals so as to yield meaningful decompositions.

## 6. REFERENCES


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

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

[3] O. Dikmen and A. Cemgil. Unsupervised single-channel source separation using Bayesian NMF. In *Proc. of the 2009 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA)*, pages 93–96, NY, USA, Oct. 2009.

[4] O. Dikmen and A. T. Cemgil. Gamma Markov random fields for audio source modelling. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):589–601, Mar. 2010.

[5] N. 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*, 18(7):1830 –1840, sept. 2010.

[6] J.-L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180 –1191, oct. 2011.

[7] J.-L. Durrieu and J.-P. Thiran. Musical audio source separation based on user-selected F0 track. In *Proc. of International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, March 12-15 2012.

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

[9] S. Ewert and M. Müller. Score informed source separation. In M. G. Meinard Müller and M. Schedl, editors, *Multimodal Music Processing*, Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

[10] B. Fuentes, R. Badeau, and G. Richard. Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012.

[11] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard. Probabilistic model for main melody extraction using constant-Q transform. In *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, 2012.

[12] T. Gerber, M. Dutasta, L. Girin, and C. Févotte. Professionally-produced music separation guided by covers. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, page to appear, 2012.

[13] R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, may 2011.

[14] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[15] P. Leveau, S. Maller, J. Burred, and X. Jaureguiberry. Convolutive common audio signal extraction. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 165–168, New Paltz, NY, USA, Oct. 2011. IEEE.

[16] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[17] A. Liutkus and P. Leveau. Separation of music+effects sound track from several international versions of the same movie. In *Audio Engineering Society Convention 128*, May 2010.

[18] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937 – 1949, 2012.

[19] 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, March 2010.

[20] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, pages 257–260, Prague, May 2011.

[21] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct. 2011.

[22] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, PP(99):1, 2011.

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

[24] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. Davies. Sparse Representations in Audio and Music: from Coding to Source Separation. *Proceedings of the IEEE.*, 98:995–1005, 06 2010.

[25] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech & Language Processing*, 21(1):71–82, 2013.

[26] C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *Comput. Music J.*, 32(1):51–59, Mar. 2008.

[27] M. N. Schmidt. Function factorization using warped Gaussian processes. In *Int. Conf. on Machine Learning (ICML'09)*, volume 382 of *ACM International Conference Proceeding Series*, page 116. ACM, 2009.

[28] U. Simsekli and A. Cemgil. Score guided musical source separation using generalized coupled tensor factorization. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[29] P. Smaragdis and G. J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *Proceedings IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72, 2009.

[30] E. Vincent, S. Arberet, and R. Gribonval. Underdetermined instantaneous audio source separation via local Gaussian modeling. In *Independent Component Analysis and Signal Separation. Lecture Notes in Computer Science.*, volume 5441/2009, pages pp 775–782, Paraty, Brésil, 2009. Springer-Verlag Berlin Heidelberg 2009.

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

[32] E. Vincent, G. Jafari, A. Abdallah, D. Plumbley, and E. Davies. Probabilistic modeling paradigms for audio source separation. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, pages 162–185. IGI Global, 2010.

[33] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52(7):1830–1847, 2004.

[34] Y. Yilmaz and A. Cemgil. Algorithms for probabilistic latent tensor factorization. *Signal Processing*, 92(8):1853–1863, 2012.