From blind to guided audio source separation: How models and side information can improve the separation of sound

Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, Frédéric Bimbot

From Blind to Guided Audio Source Separation

Emmanuel Vincent[1], Nancy Bertin[2], Rémi Gribonval[3], and Frédéric Bimbot[2]

[1]Inria
54600 Villers-lès-Nancy, France
emmanuel.vincent@inria.fr

[2]CNRS, IRISA – UMR 6074
35042 Rennes Cedex, France
firstname.name@irisa.fr

[3]Inria
35042 Rennes Cedex, France
remi.gribonval@inria.fr

Audio is a domain where signal separation has long been considered as a fascinating objective, potentially offering a wide range of new possibilities and experiences in professional and personal contexts, by better taking advantage of audio material and finely analyzing complex acoustic scenes. It has thus always been a major area for research in signal separation and an exciting challenge for industrial applications.

Starting with blind separation of toy mixtures in the mid 90's, research has progressed up to real-world scenarios today, with applications to speech enhancement and recognition, music editing, 3D sound rendering, and audio information retrieval, among others. This has mostly been made possible by the development of increasingly informed separation techniques incorporating knowledge about the sources and/or the mixtures at hand. For instance, speech source separation for remote conferencing can benefit from prior knowledge of the room geometry and/or the names of the speakers, while music remastering will exploit instrument characteristics and knowledge of sound engineers mixing habits.

After a brief historical account, we provide an overview of recent and ongoing research in this field, illustrating a variety of models and techniques designed so as to guide the audio source separation process towards efficient and robust solutions.

1 Audio source separation: basic concepts

Initially, audio source separation was formulated as a standard source separation problem, i.e., as a linear system identification and inversion problem. In the following, we assume that the sources do not move and we denote the number of sources and microphones by J and I, respectively, which are assumed to be known. We adopt the following notation: scalars are represented by plain letters, vectors by bold lowercase letters, and matrices by bold uppercase letters. The mixture signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ observed at time t when recording the source signals $\mathbf{s}(t) = [s_1(t), \ldots, s_J(t)]^T$ can be modeled by the convolution process

$$\mathbf{x}(t) = (\mathbf{A} \star \mathbf{s})(t) \tag{1}$$

where $\mathbf{A}(t) = [\mathbf{a}_1(t), \ldots, \mathbf{a}_J(t)]$ is the matrix of room impulse responses or *mixing filters* associated with sound propagation from each source to each microphone, ^{T} denotes matrix transposition, and $\star$ is the convolution operator, i.e., $x_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{\infty} a_{ij}(\tau) s_j(t-\tau)$.

1.1 Spatial images and time-frequency processing

It soon became clear that this formulation had intrinsic limitations, especially with respect to audio specificities. Firstly, the modeling of the system as impulse responses between each source location and each microphone location implicitly assumes that each source emits sound from a single point in space, preventing the modeling of spatially diffuse sources [1]. Secondly, unless extra information is available, the sources may be recovered at best up to indetermined permutation and filtering. Thirdly, the linear

system $\mathbf{A}(t)$ may be inverted only in *determined* scenarios involving fewer sources than the number of microphones ($J \leq I$).

In 1998, Cardoso [2] proposed to reformulate the mixing process as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{2}$$

so that source separation became the problem of extracting the contribution $\mathbf{c}_j(t) = [c_{j1}(t), \ldots, c_{jI}(t)]^T$ of each source to the mixture. The quantity $\mathbf{c}_j(t)$ was later called the *spatial source image* of the j-th source [3]. This reformulation circumvented the filtering indeterminacy by joining $\mathbf{a}_j(t)$ and $s_j(t)$ into a single quantity

$$\mathbf{c}_j(t) = (\mathbf{a}_j \star s_j)(t) \tag{3}$$

and the general model (2) became applicable to spatially diffuse sources which cannot be expressed as (3).

At the same time, several researchers proposed to switch to the time-frequency domain by means of the complex-valued short time Fourier transform (STFT). By rewriting the mixing process in each time frame n and each frequency bin f as

$$\mathbf{x}(n, f) = \sum_{j=1}^{J} \mathbf{c}_j(n, f), \tag{4}$$

source separation was recast as a problem akin to clustering, whereby sound in a given time-frequency bin must be allocated to the one or few active sources in that bin, and separation became achievable in *under-determined* scenarios with more sources than microphones ($J > I$) [4]. In the following, $\mathbf{x}$, $\mathbf{s}$, $\mathbf{A}$, $\mathbf{c}_j$, s_j, and $\mathbf{a}_j$ refer to time-domain variables when used with the time index t and to their time-frequency domain counterparts when used with the frame and frequency bin indices n and f.

While early source separation techniques relied on *spatial diversity*, that is the assumption that the sources have different directions of arrival, the move to time-frequency domain processing enabled the exploitation of *spectral diversity*, that is the assumption that their short-term spectra follow distinct distributions. This made it possible to handle single-channel mixtures and mixtures of sources sharing the same direction of arrival, such as vocals and drums which are often both mixed to the center in pop music.

1.2 Levels of guidance

Over the past years, successive breakthroughs have resulted from the development of audio source separation techniques increasingly suited to the properties of audio sources and to the specificities of the acoustic mixing conditions : more and more sophisticated models and algorithms have been developed to incorporate available side information (or to estimate it on the fly) about the sources and the mixing environment so as to *guide* the separation process. Today, some of the most advanced source separation systems integrate a fair number of spatial and spectral models into a single framework [5, 6]. Figure 1 summarizes visually this evolution.

According to conventional terminology, *blind* source separation does not exploit any information about the sources nor about the mixing process. Its application domain is essentially restricted to dealing with determined instantaneous mixtures, which practically never arise in audio.

Conversely, various terms such as *semi-blind* or *informed* have been used to characterize separation techniques based on some level of informedness. For instance the use

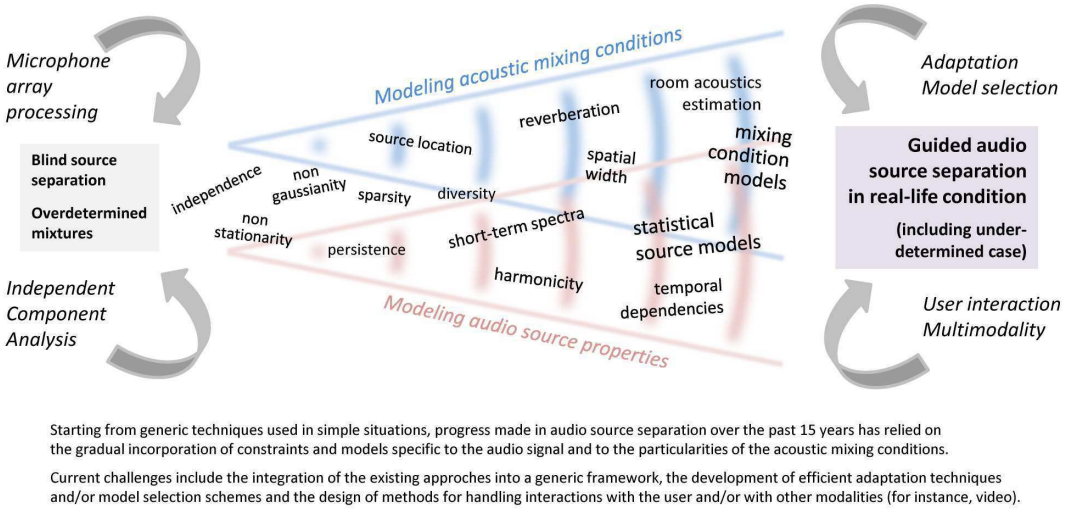

Figure 1: Audio source separation: a general overview of the evolution in the field.

of the adjective *informed* is restricted to separation techniques relying on highly precise side information coded and transmitted along with the audio, e.g., the mixing filters and the short-term power spectra of the sources, which can be seen as a form of audio coding and is not covered hereafter (see [7] for a review). As these terms happen to be used either quite specifically or rather inconsistently, we will use in the present article the term *guided* source separation.

In that sense, algorithms employing information about the general behaviour of audio sources and/or of the acoustic mixing process in general, e.g., "the sources are sparsely distributed" or "the mixture was recorded outdoors", can be described as *weakly guided*. By contrast, algorithms taking advantage of specific information about the mixture to be separated, e.g., the source positions, the names of the speakers or the musical score, may be coined as *strongly guided*.

2 Modeling paradigms

Before we focus on specific types of guidance, let us introduce the common foundations of blind and guided algorithms. It was proved early on that separation is unfeasible if more than one source has a stationary white Gaussian distribution [8]. Separation hence relies on two alternative modeling paradigms: nongaussianity or nonstationarity, where nonstationarity may manifest itself over time, over frequency, or over both [8]. These two paradigms are essentially interchangeable: choosing one of them does not restrict the type of information that may be included as guidance or the practical scenarios that can be considered.

2.1 Sparse nongaussian modeling

In the time-frequency domain, the convolutive mixing model (3) may be approximated under a *narrowband assumption* by complex-valued multiplication in each frequency bin

$$\mathbf{c}_j(n,f) = \mathbf{a}_j(f) s_j(n,f) \tag{5}$$

where the Fourier transform $\mathbf{a}_j(f)$ of $\mathbf{a}_j(t)$ is the so-called *mixing vector* for the j-th source or, in matrix form, $\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f)$ where $\mathbf{A}(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_J(f)]$ is the so-called *mixing matrix*.

Assuming that the source STFT coefficients have a stationary nongaussian distribution $P(.)$, separation may be achieved in the maximum likelihood (ML) sense as [9][1]

$$\min_{\mathbf{A},\mathbf{s}} \sum_{j,n,f} -\log P(s_j(n,f)) \text{ subject to } \mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f). \tag{6}$$

A similar objective may be derived from a deterministic inverse problem perspective [9]:

$$\min_{\mathbf{A},\mathbf{s}} \frac{1}{2} \sum_{n,f} \|\mathbf{x}(n,f) - \mathbf{A}(f)\mathbf{s}(n,f)\|_2^2 + \lambda \sum_{n,f} \mathcal{P}(\mathbf{s}(n,f)) \tag{7}$$

where $\mathcal{P}(.)$ (in calligraphic font) is a penalty term. The choice of the tradeoff parameter λ is not a trivial task. When the constraint $\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f)$ holds, the minimum of $\sum_{n,f} \mathcal{P}(\mathbf{s}(n,f))$ subject to this constraint is obtained in the limit when $\lambda \to 0$.

For typical STFT window lengths on the order of 50 to 100 ms [4], the STFT coefficients of audio signals exhibit a *sparse* distribution, with a sharp peak at zero and heavy tails compared to a Gaussian. The generalized Gaussian distribution $P(s_j(n,f)) \propto \exp(-\lambda|s_j(n,f)|^p)$ and the associated ℓ_p sparsity inducing norm $\mathcal{P}(\mathbf{s}(n,f)) = \|\mathbf{s}(n,f)\|_p^p = \sum_{j=1}^J |s_j(n,f)|^p$ with $0 < p < 2$ are popular choices to model this behavior [9, 10].

In the determined case, the objective(6) has been shown to maximize the statistical independence of the sources, hence the name *independent component analysis* (ICA). In the under-determined case, both objectives are called *sparse component analysis* (SCA) and they are typically addressed by first estimating $\mathbf{A}(f)$ and then deriving $\mathbf{s}(n,f)$ using greedy algorithms such as matching pursuit, convex optimization algorithms such as iterative soft thresholding, or nonconvex optimization algorithms depending on the chosen distribution $P(.)$ or penalty $\mathcal{P}(.)$.

If the sources are sufficiently sparse, there is a good chance that each time-frequency bin is dominated by a single source, i.e., $\mathbf{x}(n,f) \approx \mathbf{a}_j(f)s_j(n,f)$ for one source j. This leads to approximate SCA as a clustering problem. The mixing vectors $\mathbf{a}_j(f)$ are first estimated by clustering the observations $\mathbf{x}(n,f)$ and the sources $\mathbf{s}(n,f)$ are derived by grouping the time-frequency bins dominated by the same source, an operation known as *time-frequency masking*. For a more detailed introduction to ICA and SCA, see [11].

2.2 Gaussian nonstationary modeling

An alternative paradigm is to assume that the vectors of STFT coefficients of the source spatial images have a zero-mean nonstationary Gaussian distribution

$$P(\mathbf{c}_j(n,f)|\mathbf{\Sigma}_{\mathbf{c}_j(n,f)}) = \frac{1}{\det(\pi\mathbf{\Sigma}_{\mathbf{c}_j}(n,f))} e^{-\mathbf{c}_j(n,f)^H \mathbf{\Sigma}_{\mathbf{c}_j(n,f)}^{-1} \mathbf{c}_j(n,f)} \tag{8}$$

where ^{H} denotes conjugate transposition. The covariance $\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)$ depends on time and frequency. It can be factored into the product of a scalar spectro-temporal power $v_j(n,f)$ and a spatial covariance matrix $\mathbf{R}_j(f)$ [1]:

$$\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) = v_j(n,f)\mathbf{R}_j(f). \tag{9}$$

[1] In the absence of specific information over $\mathbf{A}$ or $\mathbf{s}$, minimization is typically achieved under a scaling constraint to avoid divergence of $\mathbf{A}$ and $\mathbf{s}$ to infinitely large or small values.

Separation is typically achieved by estimating the model parameters in the ML sense

$$\min_{\mathbf{R},v} \sum_{j,n,f} -\log P(\mathbf{c}_j(n,f)|\mathbf{R},v) \text{ subject to } \mathbf{x}(n,f) = \sum_{j=1}^{J} \mathbf{c}_j(n,f) \tag{10}$$

using an expectation-maximization (EM) algorithm. Once $\mathbf{R}$ and v have been estimated, $\mathbf{c}_j(n,f)$ can be derived in the minimum mean square error (MMSE) sense by multichannel Wiener filtering:

$$\widehat{\mathbf{c}}_j(n,f) = \mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \left(\sum_{j=1}^{J} \mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \right)^{-1} \mathbf{x}(n,f). \tag{11}$$

For more detailed presentation of this paradigm, see [1].

2.3 Introducing information about the model parameters

Equations (6), (7) and (10) form the basis for all guided algorithms presented hereafter. Without any further information about $\mathbf{A}$, $\mathbf{s}$, $\mathbf{R}$, or v, the spatial source images $\mathbf{c}_j(n,f)$ may be recovered at best up to indetermined permutation in each frequency bin f. This so-called *permutation problem* was historically the first reason to investigate the incorporation of more information into the models. However, guiding separation does not only address this problem, but also improves the accuracy of the parameter estimates, which in turn improves separation.

Information may be introduced either in the form of *deterministic constraints* over $\mathbf{A}$, $\mathbf{s}$, $\mathbf{R}$, or v, which restrict the values that these parameters may take, or in the form of *penalty functions* or *probabilistic priors*, which are added to the objective functions in (6), (7) and (10) and used to estimate $\mathbf{A}$, $\mathbf{s}$, $\mathbf{R}$, and v in the maximum a posteriori (MAP) sense. These contraints, penalties and priors involve their own parameters, which we call hyper-parameters. The key difference between weakly guided and strongly guided separation is that the values of the hyper-parameters must be estimated from the mixture in the former case, while they are fixed using expert knowledge or training in the latter case.

3 Modeling and exploiting spatial information

A first way to introduce information in audio source separation is to account for the fact that the mixing vectors $\mathbf{a}_j(f)$ and the spatial covariance matrices $\mathbf{R}_j(f)$ are not independent across frequency, but that they are linked by the spatial properties of the source and the recording room. We review a number of increasingly complex properties that may be used in this context, from the spatial location of the source to the full acoustics of the room. Each presented model embeds the information carried by the previous model plus some new information.

3.1 Spatial location

In the free field, the mixing vectors $\mathbf{a}_j(f)$ would be collinear with

$$\mathbf{d}_j(f) = \left[\frac{1}{r_{1j}} e^{-2i\pi f r_{1j}/c}, \ldots, \frac{1}{r_{Ij}} e^{-2i\pi f r_{Ij}/c} \right]^T \tag{12}$$

that is the *steering vector* modeling the sound attenuation and delay from the source to the microphones, with c the sound velocity and r_{ij} the distance from the j-th source to the i-th microphone. In practical recording conditions, $\mathbf{a}_j(f)$ deviates from $\mathbf{d}_j(f)$ due to reflections on the boundaries of the room, which include *early echoes* and dense late echoes known as *reverberation*. Figure 2 shows the amount of deviation as a function of the *reverberation time* RT_{60}, that is the time taken by late echoes to decay by 60 decibels (dB).

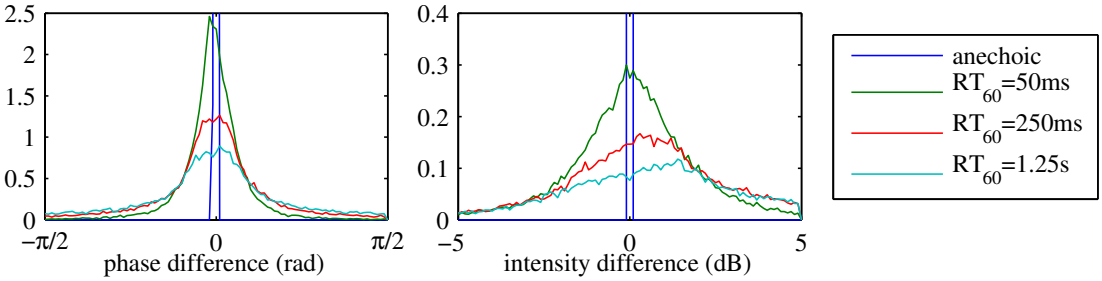

Figure 2: Example distribution over the whole frequency range of the phase and intensity differences between $\mathbf{a}_j(f)$ and $\mathbf{d}_j(f)$ as a function of RT_{60} for two microphones spaced by 20 cm recording a source at 1 m distance at a sampling frequency of 8 kHz.

Parra and Alvino [12] were the first to exploit the proximity of $\mathbf{a}_j(f)$ to $\mathbf{d}_j(f)$ by defining a penalty term $\mathcal{P}(\mathbf{A}(f))$ over the mixing matrix. Many other penalties and priors were then suggested, including Euclidean distances and Gaussian priors on the interchannel phase and intensity differences by Yılmaz et al. [4] and Mandel et al. [13]. One of the simplest is the squared Euclidean distance between $\mathbf{a}_j(f)$ and $\mathbf{d}_j(f)$

$$\mathcal{P}(\mathbf{a}_j(f)) = \|\mathbf{a}_j(f) - \mathbf{d}_j(f)\|_2^2. \tag{13}$$

Sawada et al. [14] showed that minimizing (13) w. r. t. r_{ij} is equivalent to source localization via the generalized cross-correlation (GCC) technique. This led to a joint iterative approach to source localization and separation where the source signals and the source locations are alternately updated.

3.2 Spatial width

Duong et al. [1] later observed that the narrowband approximation (5) is invalid for reverberated and/or spatially diffuse sources: the sound emitted by each source reaches the microphones from many directions at once at each frequency instead of a single apparent direction $\mathbf{a}_j(f)$, so that the channels of $\mathbf{c}_j(n, f)$ are partly uncorrelated. The spread of the distribution of incoming directions governs the perceived *spatial width* of the source at that frequency. They introduced the concept of full-rank spatial covariance matrices $\mathbf{R}_j(f)$ which, in comparison with the rank-1 spatial covariance matrices $\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f)$ resulting from (5), account not only for the spatial location of the sources but also for their width.

Assuming that the distances from the sources to the microphones are known but that their absolute location in the room is unknown, the mean of $\mathbf{R}_j(f)$ over these unknown absolute locations is approximately equal to [15]

$$\boldsymbol{\mu}_{\mathbf{R}_j}(f) = \mathbf{d}_j(f)\mathbf{d}_j^H(f) + \sigma_{\mathrm{ech}}^2\boldsymbol{\Omega}(f). \tag{14}$$

The first term accounts for direct sound, as modeled by the steering vector $\mathbf{d}_j(f)$ in (12), and the second term for echoes and reverberation, as modeled by the power of echoes

and reverberation σ^2_{ech} and by the covariance matrix of an isotropic sound field $\mathbf{\Omega}(f)$. For omni-directional microphones, the entries of $\mathbf{\Omega}(f)$ are given by the sinc function

$$\Omega_{ii'}(f) = \frac{\sin(2\pi f d_{ii'}/c)}{2\pi f d_{ii'}/c} \tag{15}$$

with $d_{ii'}$ the distance between microphones i and i'. Theoretical expressions are also available for σ^2_{ech} depending on the room dimensions and reflection coefficients. Duong et al. [15] exploited this fact to estimate $\mathbf{R}_j(f)$ in the MAP sense under an inverse-Wishart prior $P(\mathbf{R}_j(f))$.

3.3 Early echoes and reverberation

Although the full-rank model (9) improved upon the narrowband model (5), it remains an approximation of the true mixing process (3). Figure 3 illustrates the shape of a room impulse response $a_{ij}(t)$ over time. In typical reverberation conditions, these responses are several hundred milliseconds long, so that they extend over several time frames. This prompted authors to generalize (9) in the single-channel case as the convolution of $v_j(n, f)$ and a nonnegative exponentially decaying filter $q_j(l, f)$ representing the power of $a_j(t)$ for a delay of l time frames [16]. This model has been used for single-source dereverberation given knowledge of RT_{60} and it is making its way into source separation.

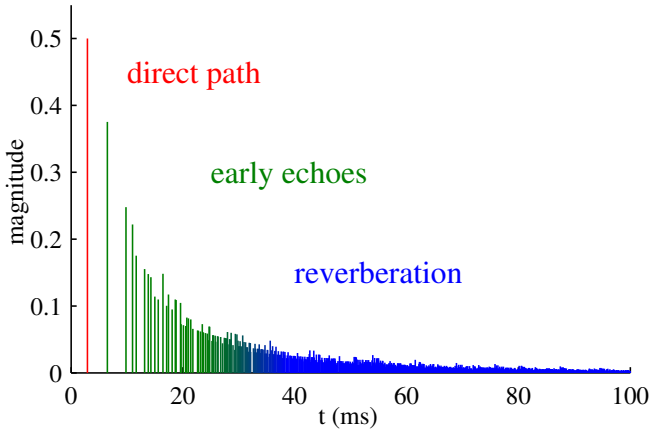

Figure 3: Schematic illustration of the magnitude of a room impulse response between a source and a microphone for a reverberation time $\text{RT}_{60} = 250$ ms.

Going one step further, Kowalski et al. [17] argued for a move back to time-domain modeling of the mixing filters, while still exploiting the sparsity of the sources in the time-frequency domain. This was achieved by replacing the narrowband loss term in (7) by the exact wideband loss term

$$\min_{\mathbf{A},\mathbf{s}} \frac{1}{2} \sum_t \|\mathbf{x}(t) - (\mathbf{A} \star \mathbf{s})(t)\|_2^2 + \lambda \sum_{n,f} \mathcal{P}(\mathbf{s}(n, f)) \tag{16}$$

and by deriving an iterative soft thresholding algorithm that effectively alternates between the time domain and the time-frequency domain at each iteration, assuming that $\mathcal{P}(\mathbf{s}(n, f))$ is a convex penalty.

This study was the starting point for subsequent studies aiming to define penalties over the mixing filters in the time domain. Benefiting from the fact that early echoes are sparsely distributed over time, as can be seen from Figure 3, Benichoux et al. [18]

exploited an ℓ_p penalty over the filters

$$\mathcal{P}(\mathbf{a}_j) = \sum_{i,t} |a_{ij}(t)|^p \tag{17}$$

with $0 < p \leq 2$. The exponential decaying shape of reverberation was later included by time-dependent rescaling of (17). The key difference with previous models is that the deviations of $\mathbf{a}_j(f)$ from $\mathbf{d}_j(f)$ are not modeled as random anymore but they must result in sparse early echoes.

3.4 Full room acoustics

Lately, in a major departure from conventional audio source separation, a number of researchers proposed to stop modeling the room impulse responses between individual sources and microphones but to learn them between all possible pairs of points in the room instead, under the constraint that the source separation system is always to be used in that room. The rationale is that room impulse responses span a manifold (said differently, a small movement in the room results in a small deviation of the impulse response), so that measuring impulse responses for a few points may suffice to predict them for other points. This accounts for all possibly available spatial information, including the direct path, the delays and amplitudes of early echoes and the shape of reverberation. Asaei et al. [19] consider each point in the room as a source and constrain most sources to be inactive by means of a group sparsity penalty (see below). More recently, Deleforge et al. [20] attempted to learn a smaller-dimensional representation of the manifold by probabilistic local linear embedding. The latter approach achieved impressive source separation results given thousands of room impulse response measurements, and its extension to practical setups with a smaller number of measurements constitutes a great avenue for research.

4 Modeling and exploiting spectro-temporal information

Besides spatial information, the source spectra and their evolution across time are the second main supply of information for audio source separation. We review increasingly complex properties of $s_j(n, f)$ and $v_j(n, f)$ that may be used to guide separation, from local persistence to long-term dependencies.

4.1 Time-frequency persistence

In audio signals, significant STFT coefficients are not randomly distributed in the time-frequency plane but they tend to cluster together. This is illustrated on Fig. 4, where vertical and horizontal lines appear, corresponding to transient and tonal parts of musical notes, respectively. Similar and more complex structures can be found in speech.

This *persistence* over time or over frequency can be promoted by the use of *group sparsity* or other *structured sparsity* penalties on $s_j(n, f)$ [21]. For instance, the $\ell_{1,2}$ norm

$$\mathcal{P}(s_j) = \sum_n \sqrt{\sum_f |s_j(n, f)|^2}, \tag{18}$$

imposes sparsity over time but no constraint over frequency. An alternative technique is to set a hidden Markov model (HMM) prior on sequences of STFT coefficients. Févotte et al. [22] showed that the latter approach outperforms unstructured priors in a denoising task.

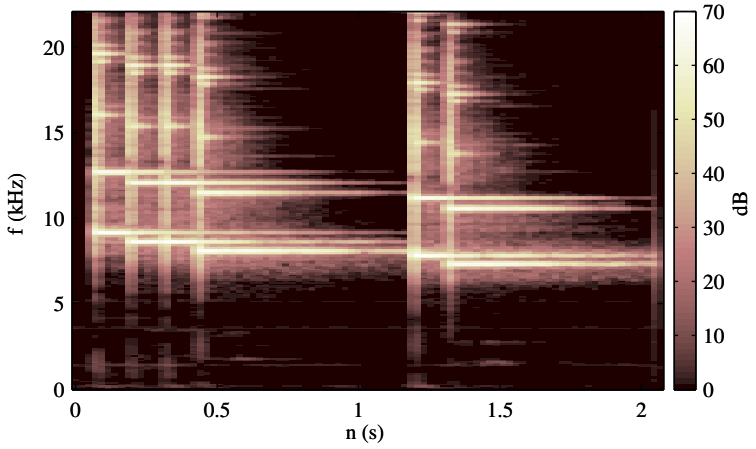

Figure 4: Spectrogram of a xylophone melody.

4.2 Short-term spectra

Beyond frequency persistence, sound sources are characterized by their short-term spectra, that is the dependencies between $v_j(n, f)$ over the whole frequency range f. A popular approach is to represent the source short-term spectra $v_j(n, f)$ as the sum of nonnegative basis spectra $w_{jk}(f)$, scaled by nonnegative time-varying activation coefficients $h_{jk}(n)$ [23, 24][2]

$$v_j(n, f) = \sum_{k=1}^{K} w_{jk}(f) h_{jk}(n). \tag{19}$$

Each basis spectrum may represent, e.g., part of a speech phoneme or a musical note, as illustrated in the top left part of Figure 5. Due to its equivalent matrix form $\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j$, this model is better known as *nonnegative matrix factorization* (NMF). Considering the fact that only one speech phoneme or few musical notes may be active at once, sparsity was enforced by reducing the sum to a single component k [25] or by adding penalties such as the ℓ_1 norm $\mathcal{P}(\mathbf{H}_j) = \sum_{k,n} |h_{jk}(n)|$ [23]. Group sparsity penalties and priors were also introduced to favor simultaneous activity of basis spectra associated with the same phoneme or note, or to select the correct speaker or instrument among a collection of basis spectra trained on different speakers or instruments [26].

4.3 Fine spectral structure and spectral envelope

Several extensions were brought to NMF to further constrain the basis spectra. A first idea is to decompose the basis spectra themselves by NMF as the sum of narrowband spectral patterns $b_{jkm}(f)$ weighted by spectral envelope coefficients e_{jkm}:

$$w_{jk}(f) = \sum_{m=1}^{M_k} b_{jkm}(f) e_{jkm}. \tag{20}$$

The narrowband spectra may be fixed so as to enforce harmonicity (i.e., spectral peaks at integer multiples of a given fundamental frequency) or smoothness, which are common structures to many sound sources, and to adapt the spectral envelope coefficients to the mixture, which are specific to each source. These structures are suitable for sustained

[2]This model has been indifferently applied to magnitude spectra or to power spectra in the single-channel case, however only the latter easily generalizes to the multichannel case.

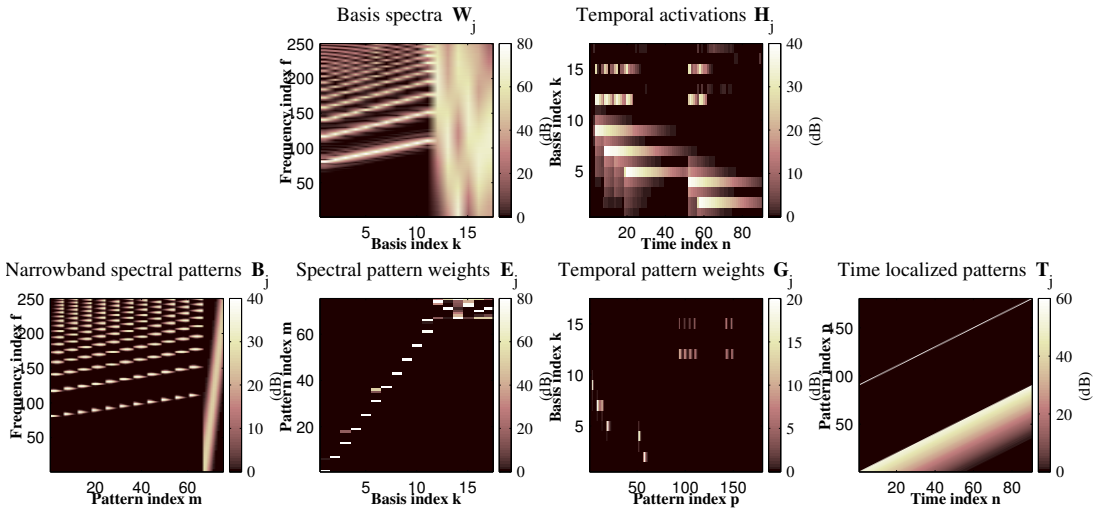

Figure 5: Multilevel NMF decomposition of the spectrogram in Fig. 4: $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j = \mathbf{B}_j\mathbf{E}_j\mathbf{G}_j\mathbf{T}_j$. Top: decomposition as the product of basis spectra $\mathbf{W}_j$ and temporal activations $\mathbf{H}_j$. Bottom: second level decomposition of $\mathbf{W}_j$ as the product of harmonic and noisy narrowband spectral patterns $\mathbf{B}_j$ and associated spectral envelopes $\mathbf{E}_j$, and of $\mathbf{H}_j$ as the product of time-localized patterns $\mathbf{T}_j$ activated at some time weights $\mathbf{G}_j$.

and transient musical sounds for instance, as shown in the bottom left part of Figure 5.

Another refinement complying with the physical production of many natural sounds is to decompose the source short-term spectra via the *excitation-filter* model

$$v_j(n,f) = v_j^{\text{ex}}(n,f)v_j^{\text{ft}}(n,f) \tag{21}$$

where $v_j^{\text{ex}}(n,f)$ and $v_j^{\text{ft}}(n,f)$ represent the excitation signal (e.g., the glottal source) and the response of the filter (e.g., the vocal tract) and they are modeled by NMF [27]. This constraint enforces similar spectra for different fundamental frequencies, in a similar way as the *shift-invariance* constraint in [28], that is the constraint that all basis spectra are spectrally translated versions of a single spectrum.

Ozerov et al. [5] lately proposed a comprehensive multilevel NMF framework integrating (19)–(21) by multiplication of up to eight matrices, each of them capable of embodying specific knowledge or constraints in a flexible way. All these extensions can be compactly formalized as *non-negative tensor factorization* (NTF), an extension of NMF to multi-dimensional arrays.

4.4 Temporal evolution

The aforementioned models do not directly model the temporal evolution of the spectra. At a short time scale, Virtanen [23] enforced the continuity of NMF activation coefficients by adding the penalty $\mathcal{P}(\mathbf{H}_j) = \sum_n |h_{jk}(n+1) - h_{jk}(n)|^2$ while Ozerov et al. [5] modeled them in a similar fashion as (20) as the product of time-localized patterns and sparse temporal envelopes, as depicted in the bottom right part of Figure 5. Continuous or HMM priors on $h_{jk}(n)$ were also used to this aim.

At a medium time scale, Smaragdis [29] generalized (19) into the *convolutive NMF* model

$$v_j(n,f) = \sum_{k=1}^{K} \sum_{l} w_{jk}(l,f)h_{jk}(n-l) \tag{22}$$

where the basis elements $w_{jk}(l, f)$ are now spectro-temporal patches rather than single-frame spectra, thus explicitly encoding the temporal evolution of sound events at each frequency. Musicological models and spoken language models were also exploited to favor certain note and chord progressions or certain sequences of words using longer-term HMM priors on $h_{jk}(n)$. Mysore and Sahani [26] provided an efficient algorithm to separate multiple sources, each modeled by an HMM.

In another major departure from conventional audio source separation, several researchers recently proposed to exploit the information encoded by redundancy and repetitive patterns at very long time scales, so as to optimize the use of available information over the whole signal duration. Robust principal component analysis (RPCA), which decomposes an input spectrogram as the sum of a low-rank matrix and a sparse matrix, was used by Huang et al. [30] to separate (sparse) drum and melody sources from a (low-rank) repetitive tonal accompaniment. The search for repeating patterns in music was also exploited by Rafii et al. [31] through the identification of repeating segments (of up to 40 seconds duration), their modeling, and their extraction via time-frequency masking. In the future, such ideas may be applied to automatic learning of fine-grained models from larger and larger amounts of audio data eventually covering the sounds arising in the mixture to be separated.

5 Impact and perspectives

Over the past fifteen years, audio source separation has recorded constant progress and today it has reached a level of maturity which enables its integration in real-life application contexts. For instance, multichannel NMF and NTF have improved performance by 3 to 4 dB signal-to-distortion ratio (SDR) compared to SCA in certain scenarios and they have made it possible to separate real-world music recordings using weakly guided models for typical instruments (vocals, drums, bass) and for the remaining instruments [3]. Joint spatial and spectral modeling [5, 6] and convolutive NMF have contributed to the reduction of the keyword error rate for small-vocabulary automatic speech recognition (ASR) from 44% down to as little as 8% in a strongly guided real-world domestic scenario involving knowledge of the speaker and his/her spatial position [32]. Finally, weakly guided separation of percussive and harmonic content in music has helped several music information retrieval (MIR) tasks, reducing for instance the relative error rate for chord recognition by 28% [33].

These and other results show that improved separation performance in many scenarios can be obtained by modeling and exploiting spatial and spectral properties of sounds, i.e., by designing models and constraints which account for the specificities of audio sources and acoustic mixing conditions. Two trends can be seen: developing complex, hierarchical models with little training so as to adapt to unknown situations with little amounts of data, or training simpler models on huge amounts of data, e.g., thousands of room impulse responses and dozens of hours of speech, so as to benefit from the power of big data and turn parameter estimation into a model selection problem.

In either case, the design of clever, computationally efficient convex relaxations and nonconvex optimization algorithms is given increasing attention in order to handle the optimization of all model parameters and hyper-parameters at once and to escape extra local optima which may hinder the benefit of such models. In certain scenarios, some hyper-parameters can be set using expert knowledge or training on separate data, and only the remaining hyper-parameters need to be estimated from the mixture.

With few exceptions [5, 6], most separation systems currently exploit only a limited set of constraints, penalties, or priors. Research is ongoing on the improvement of the above models, as well as on the incorporation of side-information that has little been

exploited so far, e.g., visual information about the source movements. Ultimately, the integration of the variety of developed models and schemes into a complete, fully versatile system constitutes a challenge in itself.

References

[1] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[2] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 4, pp. 1941–1944.

[3] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Process.*, vol. 92, pp. 1928–1936, 2012.

[4] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[6] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[7] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation : a comparative study," in *Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 2397–2401.

[8] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. 3rd Int. Conf. Independent Component Analysis and Blind Signal Separation*, 2001, pp. 1–6.

[9] R. Gribonval and M. Zibulevsky, "Sparse component analysis," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pp. 367–420. Academic Press, 2010.

[10] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. 7th Int. Conf. Independent Component Analysis and Signal Separation*, 2007, pp. 430–437.

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

[12] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, 2002.

[13] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[14] 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 Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1592–1604, 2007.

[15] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP J. Adv. Signal Process.*, vol. 2013, pp. 149, 2013.

[16] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, 2009.

[17] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[18] A. Benichoux, P. Sudhakar, F. Bimbot, and R. Gribonval, "Well-posedness of the permutation problem in sparse filter estimation with ℓ^p minimization," *Applied Computat. Harm. Anal.*, vol. 35, no. 3, pp. 394–406, 2013.

[19] A. Asaei, M. E. Davies, H. Bourlard, and V. Cevher, "Computational methods for structured sparse component analysis of convolutive speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 2425–2428.

[20] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for binaural sound-source separation and localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 76–80.

[21] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image, Video Process.*, vol. 3, no. 3, pp. 251–264, 2009.

[22] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani, "Sparse regression with structured priors: application to audio denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 3, pp. 57–60.

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

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

[25] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, 2006.

[26] G. Mysore and M. Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *Proc. 29th Int. Conf. Machine Learning*, 2012, pp. 1887–1894.

[27] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, 2010.

[28] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Comput. Intell. Neurosci.*, vol. 2008, 2008, Article ID 872425.

[29] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, 2007.

[30] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 57–60.

[31] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 73–84, 2013.

[32] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 621–633, 2013.

[33] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Advances in Music Information Retrieval*. Springer, 2010.