

Music Source Separation and its Applications to MIR

Emmanuel Vincent and Nobutaka Ono

INRIA Rennes - Bretagne Atlantique, France
The University of Tokyo, Japan

Tutorial supported by the VERSAMUS project
<http://versamus.inria.fr/>

Contributions from Alexey Ozerov, Ngoc Duong, Simon Arberet, Martin Klein-Hennig and Volker Hohmann.



Part I: General principles of music source separation

- ① Source separation and music
- ② Computational auditory scene analysis
- ③ Probabilistic linear modeling
- ④ Probabilistic variance modeling
- ⑤ Summary and future challenges

Audio source separation

Many sound scenes are **mixtures** of several concurrent sound sources.

When facing such scenes, humans are able to perceive and focus on individual sources.

Source separation is the problem of **recovering the source signals** underlying a given mixture.

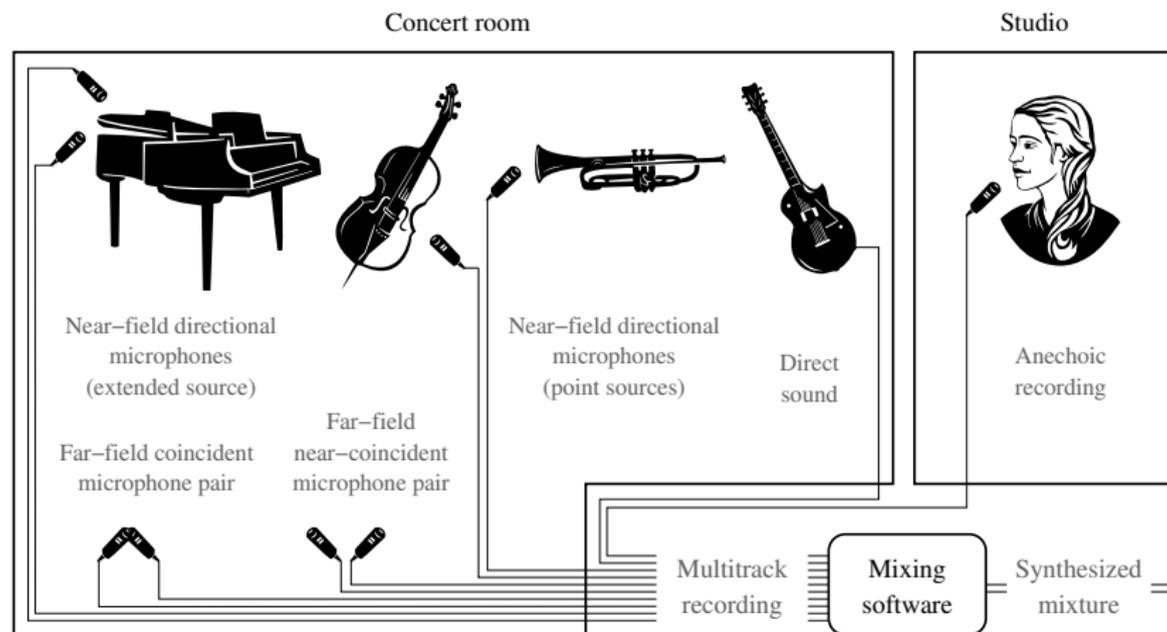
It is a core problem of audio signal processing, with applications such as:

- hearing aids,
- post-production, remixing and 3D upmixing,
- spoken/multimedia document retrieval,
- MIR.

The data at hand

As an inverse problem, source separation requires some **knowledge**.

Music is among the most difficult application areas of source separation because of the wide variety of sources and mixing processes.



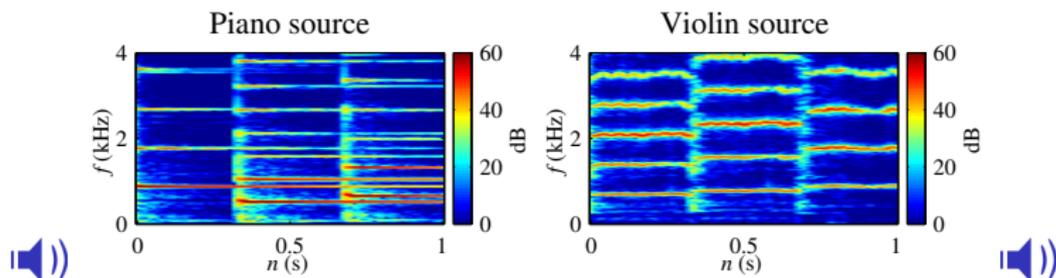
Music sources

Music sources include acoustical or virtual instruments and singing voice.

Sound is produced by transmission of one or more excitation movements/signals through a resonant body/filter.

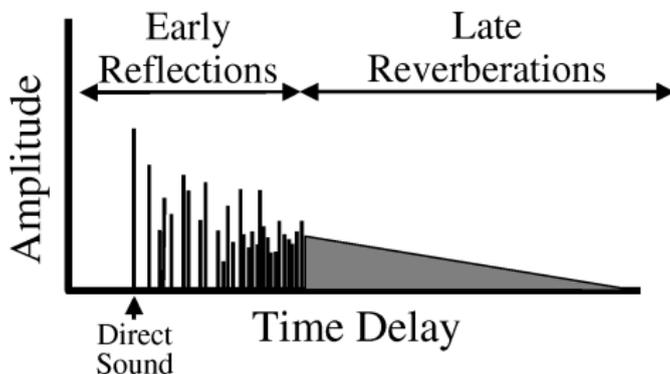
This results in a wide variety of sounds characterized by their:

- polyphony (monophonic or polyphonic)
- temporal shape (transitory, constant or variable)
- spectral fine structure (random or pitched)
- spectral envelope



Effects of microphone recording

For point sources, room acoustics result in **filtering** of the source signal



where the **intensity** and **delay** of direct sound are functions of the **source position** relative to the microphone.

Diffuse sources (piano, drums) amount to (infinitely) many point sources.

The mixture signal is equal to the sum of the contributions of all sources at each microphone.

Software mixing effects

Usual software mixing effects include:

- compression and equalization
- **panning**, *i.e.* channel-dependent intensity scaling
- **reverb**
- **polarity** and autopan

The latter are widely employed to achieve perceptual envelopment, whereby even point sources are mixed diffusely.

Again, the intensity of direct sound is a function of the source position and the mixture signal is equal to the sum of the contributions of all sources in each channel.

Overview

Hundreds of source separation systems were designed in the last 20 years. . .

. . . but few are yet applicable to real-world music, as illustrated by the 2008 and 2010 Signal Separation Evaluation Campaigns (SiSEC).

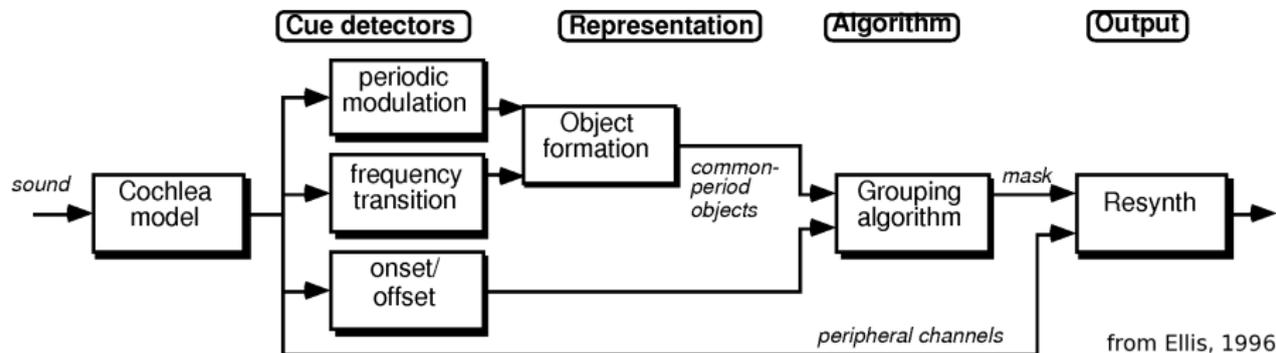
The wide variety of techniques boils down to three modeling paradigms:

- **computational auditory scene analysis (CASA)**,
- **probabilistic linear modeling**, including independent component analysis (ICA) and sparse component analysis (SCA),
- **probabilistic variance modeling**, including hidden Markov models (HMM) and nonnegative matrix factorization (NMF).

- ① Source separation and music
- ② Computational auditory scene analysis
- ③ Probabilistic linear modeling
- ④ Probabilistic variance modeling
- ⑤ Summary and future challenges

Computational auditory scene analysis (CASA)

CASA aims to emulate the human auditory system.



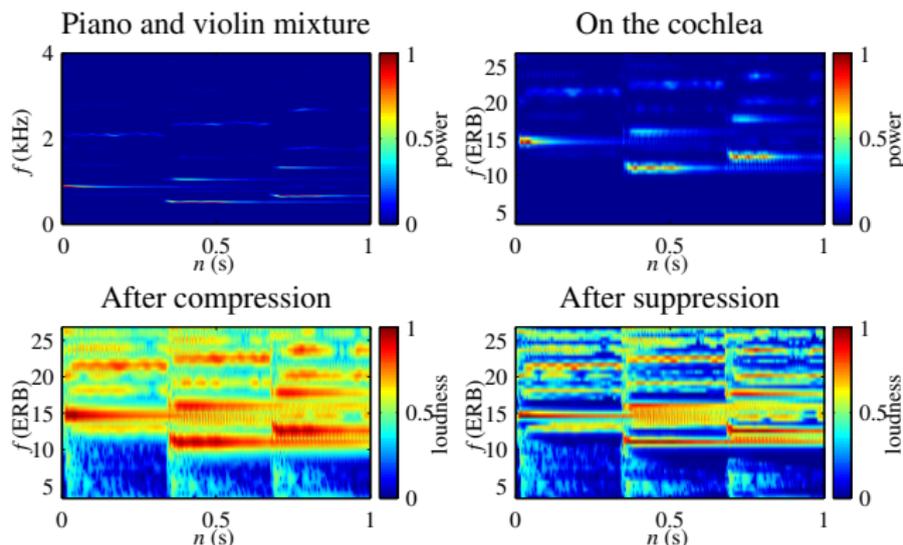
Source formation relies on the *Gestalt* rules of cognition:

- proximity,
- similarity,
- continuity,
- closure,
- common fate.

Auditory front-end

The sound signal is first converted into an **auditory nerve representation** via a series of processing steps:

- outer- and middle-ear: filter
- cochlear traveling wave model: filterbank
- haircell model: halfwave rectification + bandwise compression + cross-band suppression

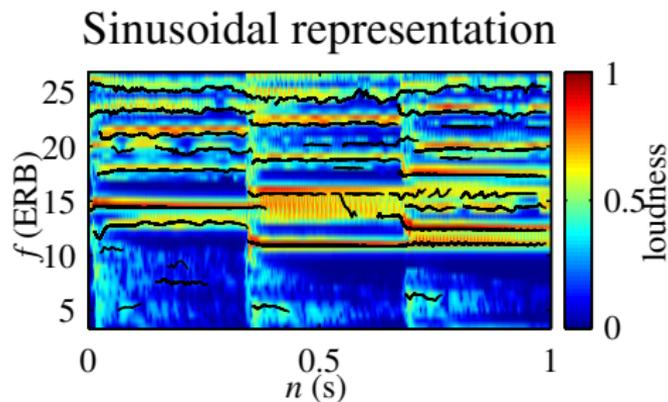


Sinusoidal+noise decomposition

Many systems further decompose the signal into a collection of **sinusoidal tracks** plus residual noise.

This decomposition is useful to

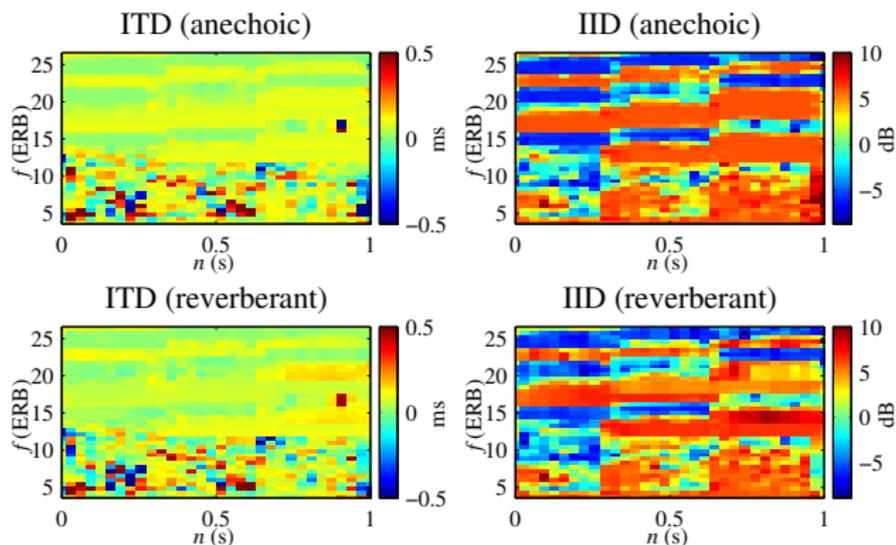
- reduce the number of sound atoms to be grouped into sources,
- enable the exploitation of advanced cues, e.g. amplitude and frequency modulation.



Spatial cues

Spatial proximity is assessed by comparing the observed

- interchannel time difference (ITD),
- interchannel intensity difference (IID).



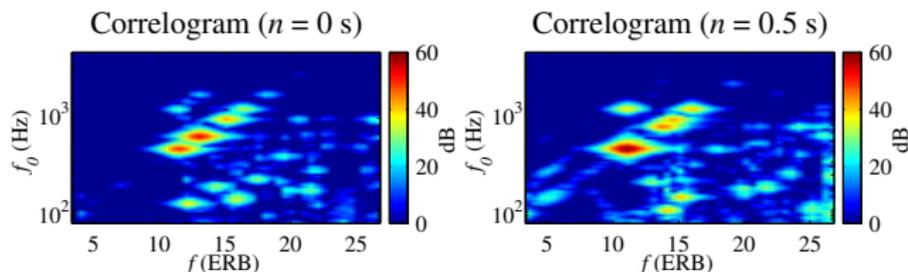
Note: in practice, most systems consider only binaural data, *i.e.* recorded by in-ear microphones.

Spectral cues

The *Gestalt* rules also translate into *e.g.*

- common pitch and onset time,
- similar spectral envelope,
- spectral and temporal smoothness,
- lack of silent time intervals,
- correlated amplitude and frequency modulation.

Most effort has been devoted to the estimation of **pitch** by cross-correlation of the auditory nerve representation in each band.



Learned cues

In addition to the above **primitive cues**, the auditory system relies on a range of **learned cues** to focus on a given source:

- veridical expectation (episodic memory): "I know the lyrics"
- schematic expectation (semantic memory): "The inaudible word after *love you* must be *babe*"
- dynamic adaptive expectation (short-term memory): "This melody already occurred in the song"
- conscious expectation

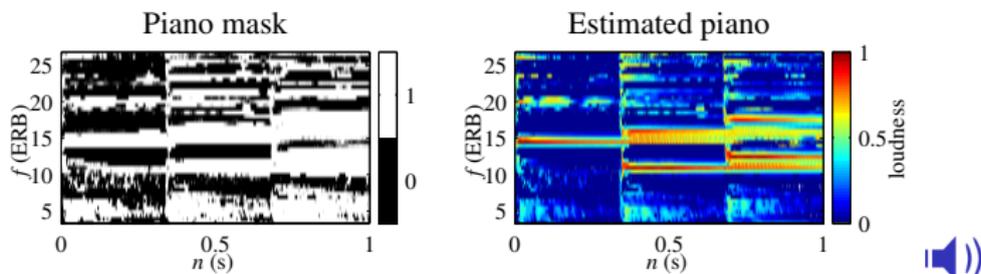
Source formation and signal extraction

Each time-frequency bin or each sinusoidal track is associated to a single source according to the above cues: this is known as **binary masking**.

Individual cues are ambiguous, *e.g.*

- the observed IID/ITD may be due to a single source in the associated direction or to several concurrent sources around that direction,
- a given sinusoidal track may be a harmonic of different sources.

Most systems exploit several cues with some **precedence order** or **weighting factors** determined by psycho-acousticians.



Summary of CASA

Advantages:

- wide range of spectral, spatial and learned cues
- robustness thanks to joint exploitation of several cues

Limitations:

- musical noise artifacts due to binary masking
- suboptimal cues, designed for auditory scene analysis instead of machine source separation
- practical limitation to a few spectral and/or spatial cues, with no general framework for the integration of additional cues
- (historically) bottom-up approach, prone to error propagation, and limitation to pitched sources
- no results within recent evaluation campaigns

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

Model-based audio source separation

The alternative top-down approach consists of finding the source signals that best fit the mixture and the expected properties of audio sources.

In a probabilistic framework, this translates into

- building **generative models** of the source and mixture signals,
- inferring latent variables in a **maximum a posteriori (MAP)** sense.

Linear modeling

The established linear modeling paradigm relies on two assumptions:

- ① point sources
- ② low reverberation

Under assumption 1, the sources and the mixing process can be modeled as **single-channel source signals** and a **linear filtering process**.

Under assumption 2, this filtering process is equivalent to complex-valued multiplication in the **time-frequency domain** via the short-time Fourier transform (STFT).

In each time-frequency bin (n, f)

$$\mathbf{X}_{nf} = \sum_{j=1}^J S_{jnf} \mathbf{A}_{jf}$$

\mathbf{X}_{nf} : vector of mixture STFT coeff.

J : number of sources

S_{jnf} : j th source STFT coeff.

\mathbf{A}_{jf} : j th mixing vector

Priors over the mixing vectors

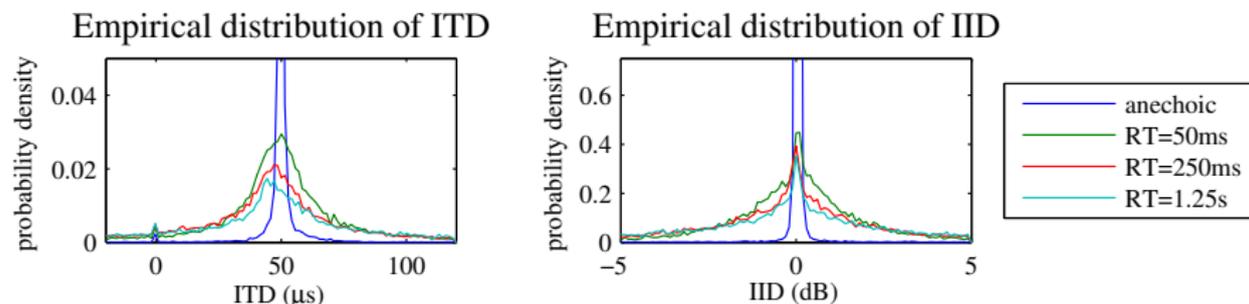
The mixing vectors \mathbf{A}_{jf} encode the apparent sound direction in terms of

- ITD τ_{jf} ,
- IID g_{jf} .

For non-echoic mixtures, ITDs and IIDs are **constant over frequency** and related to the direction of arrival (DOA) θ_j of each source

$$\mathbf{A}_{jf} \propto \begin{pmatrix} 1 \\ g_j e^{-2i\pi f \tau_j} \end{pmatrix}$$

For echoic mixtures, ITDs and IIDs follow a **smearing distribution** $P(\mathbf{A}_{jf}|\theta_j)$



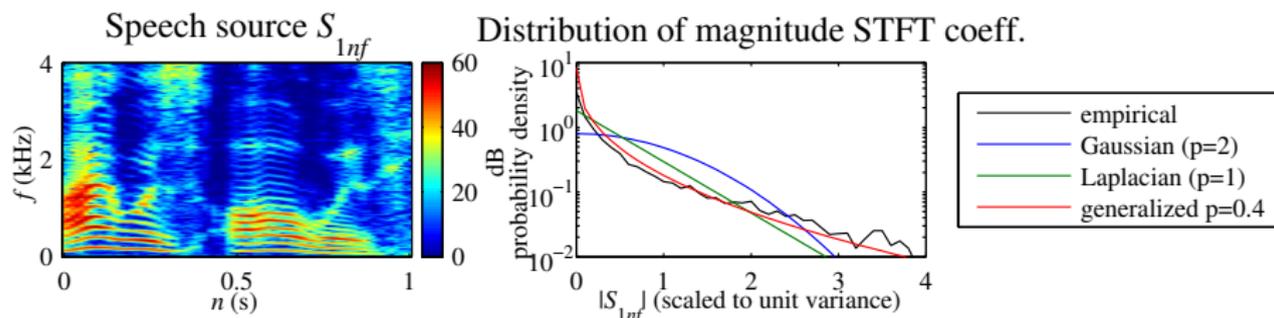
I.i.d. priors over the source STFT coefficients

Most systems assume that the sources have random spectra, *i.e.* their STFT coefficients S_{jnf} are **independent and identically distributed (i.i.d.)**.

The magnitude STFT coefficients of audio sources are **sparse**: at each frequency, few coefficients have large values while most are close to zero.

This property is well modeled by the generalized exponential distribution

$$P(|S_{jnf}| | p, \beta_f) = \frac{p}{\beta_f \Gamma(1/p)} e^{-\left| \frac{S_{jnf}}{\beta_f} \right|^p} \quad \begin{array}{l} p: \text{shape parameter} \\ \beta_f: \text{scale parameter} \end{array}$$



Note: coarser binary activity priors have also been employed.

Inference algorithms

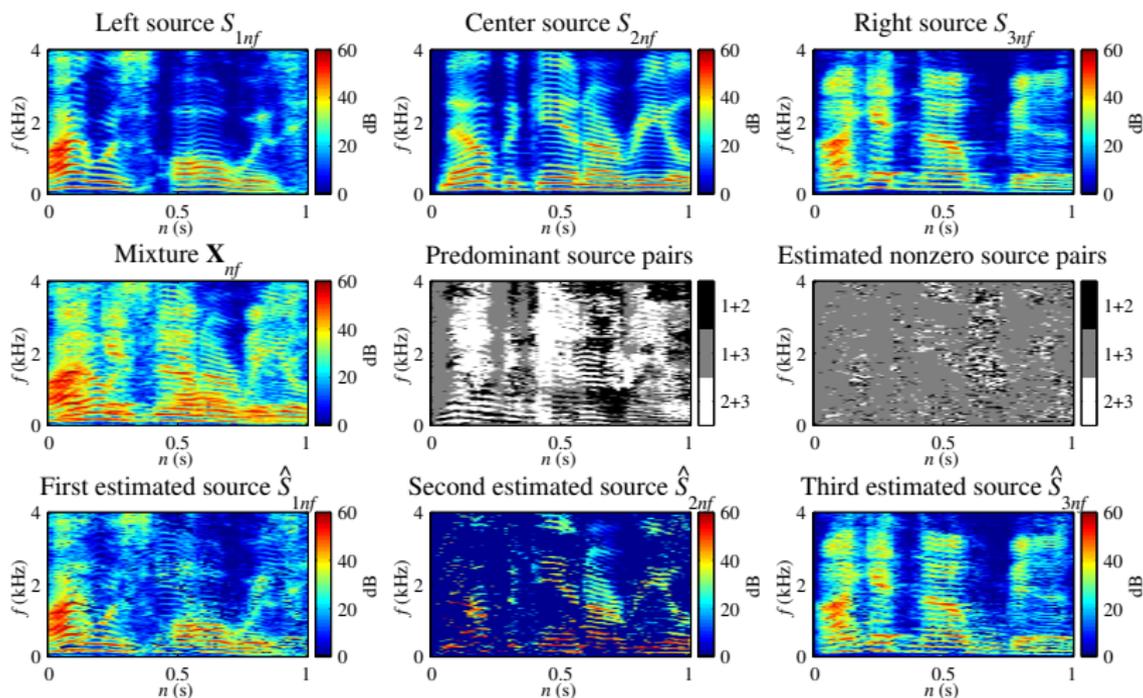
Given the above priors, source separation is typically achieved by joint MAP estimation of the source STFT coefficients S_{jnf} and other latent variables $(\mathbf{A}_{jf}, g_j, \tau_j, \rho, \beta_j)$ via **alternating nonlinear optimization**.

This objective is called sparse component analysis (SCA).

For typical values of ρ , the MAP source STFT coefficients are **nonzero for at most two sources** in a stereo setting.

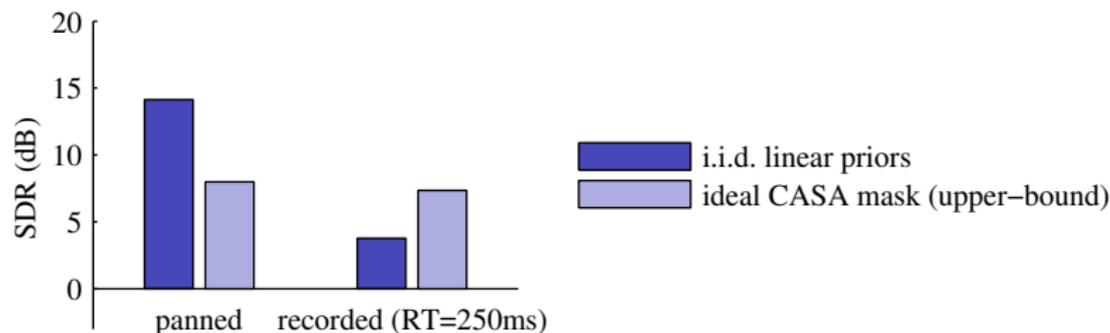
When the number of sources is $J = 2$, SCA is renamed nongaussianity-based frequency-domain independent component analysis (FDICA).

Practical illustration of separation using i.i.d. linear priors



Time-frequency bins dominated by the center source are often erroneously associated with the two other sources.

SiSEC results on toy mixtures of 3 sources



Panned mixture



Estimated sources using i.i.d. linear priors



Recorded reverberant mixture



Estimated sources using i.i.d. linear priors



Summary of probabilistic linear modeling

Advantages:

- top-down approach
- separation of more than one source per time-frequency bin

Limitations:

- restricted to mixtures of non-reverberated point sources
- separation of at most two sources per time-frequency bin
- musical noise artifacts due to the ambiguities of spatial cues
- no straightforward framework for the integration of spectral cues

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

Idea 1: from sources to mixture components

Diffuse or semi-diffuse sources cannot be modeled as single-channel signals and not even as finite dimensional signals.

Instead of considering the signal produced by each source, one may consider its contribution to each channel of the mixture signal.

Source separation becomes the problem of estimating the **multichannel mixture components** underlying the mixture.

In each time-frequency bin (n, f)

$$\mathbf{x}_{nf} = \sum_{j=1}^J \mathbf{c}_{jnf}$$

\mathbf{x}_{nf} : vector of mixture STFT coeff.

J : number of sources

\mathbf{c}_{jnf} : j th mixture component

Idea 2: translation and phase invariance

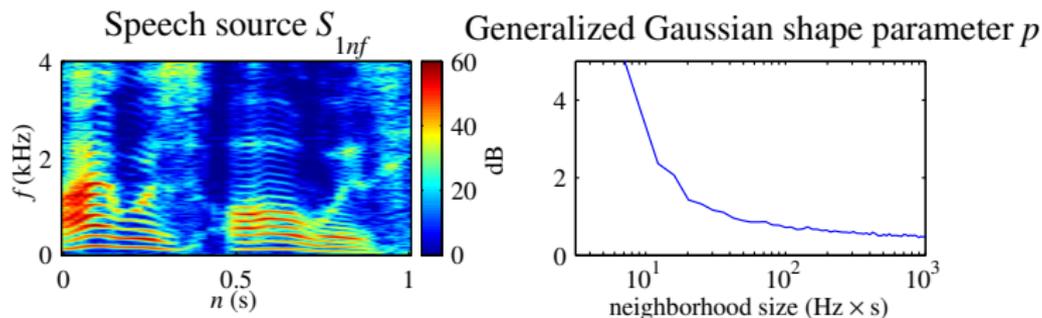
In order to overcome the ambiguities of spatial cues, additional spectral cues are needed as shown by CASA.

Most audio sources are **translation- and phase-invariant**: a given sound may be produced at any time with any relative phase across frequency.

Variance modeling

Variance modeling combines these two ideas by modeling the STFT coefficients of individual mixture components by a **circular multivariate distribution** whose parameters vary over time and frequency.

The non-sparsity of source STFT coefficients over small time-frequency regions suggests the use of a **non-sparse distribution**.



Choice of the distribution

For historical reasons, several distributions have been preferred in a mono context, which can equivalently be expressed as **divergence** functions over the source magnitude/power STFT coefficients:

- Poisson \leftrightarrow Kullback-Leibler divergence aka I-divergence
- tied-variance Gaussian \leftrightarrow Euclidean distance
- log-Gaussian \leftrightarrow weighted log-Euclidean distance

These distributions do not easily generalize to multichannel data.

The multichannel Gaussian model

The **zero-mean Gaussian distribution** is a simple multichannel model.

$$P(\mathbf{C}_{jnf} | \boldsymbol{\Sigma}_{jnf}) = \frac{1}{\det(\pi \boldsymbol{\Sigma}_{jnf})} e^{-\mathbf{C}_{jnf}^H \boldsymbol{\Sigma}_{jnf}^{-1} \mathbf{C}_{jnf}} \quad \boldsymbol{\Sigma}_{jnf}: jth \text{ component covariance matrix}$$

The covariance matrix $\boldsymbol{\Sigma}_{jnf}$ of each mixture component can be factored as the product of a **scalar nonnegative variance** V_{jnf} and a **mixing covariance matrix** \mathbf{R}_{jf} respectively modeling spectral and spatial properties

$$\boldsymbol{\Sigma}_{jnf} = V_{jnf} \mathbf{R}_{jf}$$

Under this model, the mixture STFT coefficients also follow a Gaussian distribution whose covariance is the sum of the component covariances

$$P(\mathbf{X}_{nf} | V_{jnf}, \mathbf{R}_{jf}) = \frac{1}{\det\left(\pi \sum_{j=1}^J V_{jnf} \mathbf{R}_{jf}\right)} e^{-\mathbf{X}_{nf}^H \left(\sum_{j=1}^J V_{jnf} \mathbf{R}_{jf}\right)^{-1} \mathbf{X}_{nf}}$$

General inference algorithm

Independently of the priors over V_{jnf} and \mathbf{R}_{jf} , source separation is typically achieved in two steps:

- joint MAP estimation of all model parameters using the **expectation maximization** (EM) algorithm,
- MAP estimation of the source STFT coefficients conditional to the model parameters by **multichannel Wiener filtering**

$$\hat{\mathbf{C}}_{jnf} = V_{jnf} \mathbf{R}_{jf} \left(\sum_{j'=1}^J V_{j'nf} \mathbf{R}_{j'f} \right)^{-1} \mathbf{X}_{nf}.$$

Rank-1 priors over the mixing covariances

The mixing covariances \mathbf{R}_{jf} encode the apparent spatial direction and spatial spread of sound in terms of

- ITD,
- IID,
- normalized interchannel correlation a.k.a. [interchannel coherence](#).

For non-reverberated point sources, the interchannel coherence is equal to one, *i.e.* \mathbf{R}_{jf} has [rank 1](#)

$$\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H$$

The priors $P(\mathbf{A}_{jf} | \theta_j)$ used with linear modeling can then be simply reused.

Full-rank priors over the mixing covariances

For reverberated or diffuse sources, the interchannel coherence is smaller than one, *i.e.* \mathbf{R}_{jf} has **full rank**.

The theory of statistical room acoustics suggests the **direct+diffuse model**

$$\mathbf{R}_{jf} \propto \lambda_j \mathbf{A}_{jf} \mathbf{A}_{jf}^H + \mathbf{B}_f$$

λ_j : direct-to-reverberant ratio

\mathbf{A}_{jf} : direct mixing vector

\mathbf{B}_f : diffuse noise covariance

with

$$\mathbf{A}_{jf} = \sqrt{\frac{2}{1 + g_j^2}} \begin{pmatrix} 1 \\ g_j e^{-2i\pi f \tau_j} \end{pmatrix}$$

τ_j : ITD of direct sound

g_j : IID of direct sound

$$\mathbf{B}_f = \begin{pmatrix} 1 & \text{sinc}(2\pi fd/c) \\ \text{sinc}(2\pi fd/c) & 1 \end{pmatrix}$$

d : microphone spacing

c : sound speed

I.i.d. priors over the source variances

Baseline systems rely again on the assumption that the sources have random spectra and model the source variances V_{jnf} as **i.i.d. and locally constant** within small time-frequency regions.

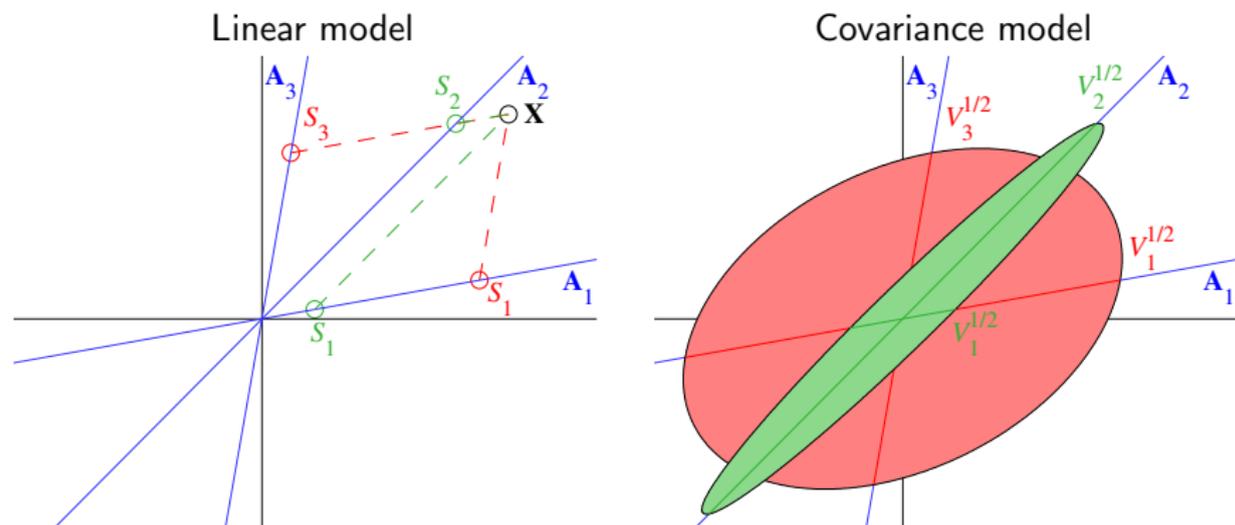
When these follow a **mildly sparse prior**, it can be shown that the MAP variances are **nonzero for up to four sources**.

Discrete priors constraining the number of nonzero variances to one or two have also been employed.

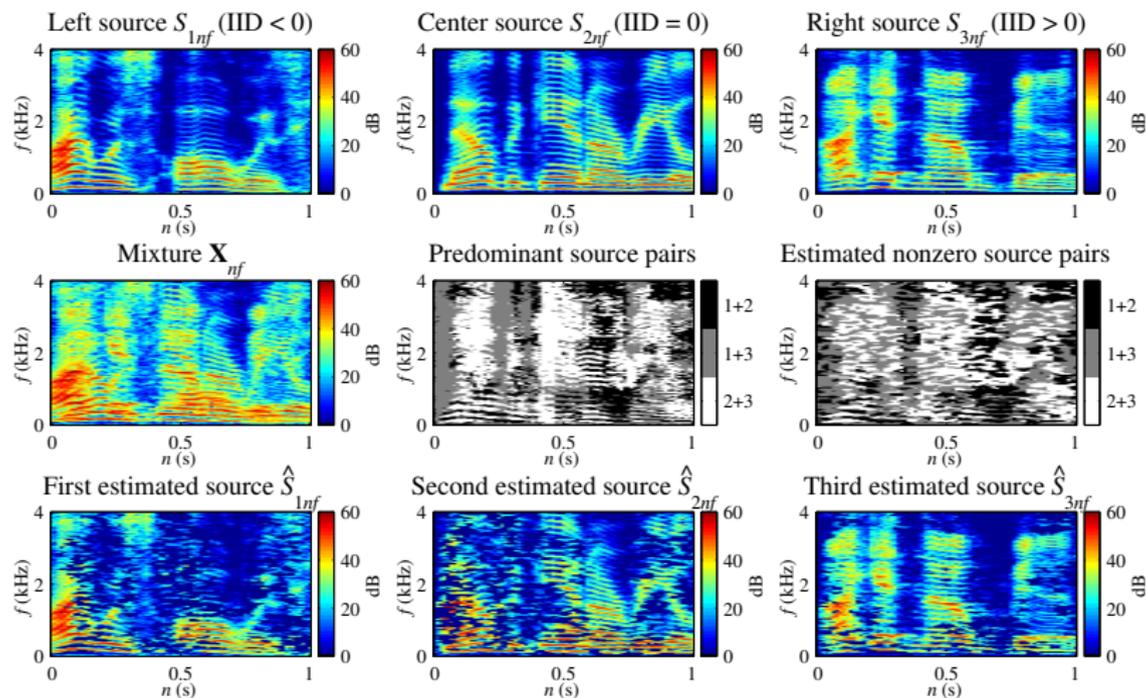
When the number of sources is $J = 2$, this model is also called nonstationarity-based FDICA.

Benefit of exploiting interchannel coherence

Interchannel coherence helps resolving some ambiguities of ITD and IID and identify the predominant sources more accurately.



Practical illustration of separation using i.i.d. variance priors



Spectral priors based on template spectra

Variance modeling enables the design of phase-invariant spectral priors.

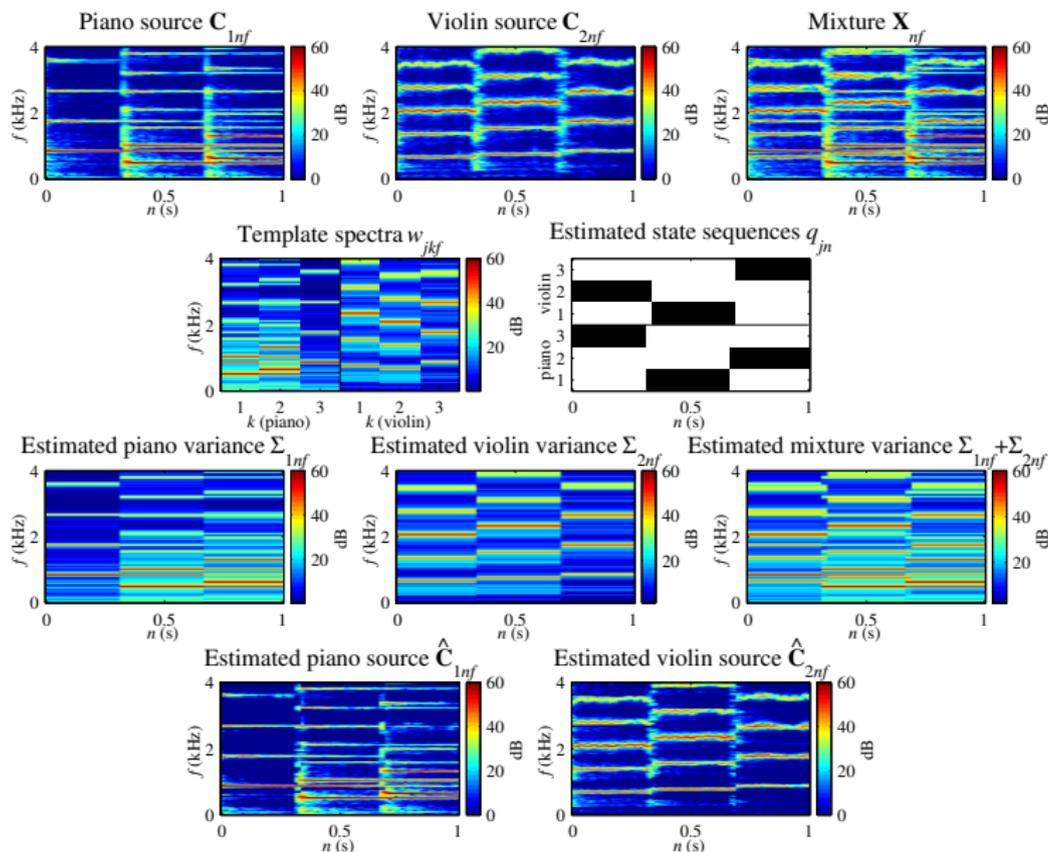
The Gaussian mixture model (GMM) represents the variance V_{jnf} of each source at a given time by one of K **template spectra** w_{jkf} indexed by a **discrete state** q_{jn}

$$V_{jnf} = w_{jq_{jn}f} \text{ with } P(q_{jn} = k) = \pi_{jk}$$

Different strategies have been proposed to learn these spectra:

- speaker-independent training on separate single-source data,
- speaker-dependent training on separate single-source data,
- MAP adaptation to the mixture using model selection or interpolation,
- MAP inference from a coarse initial separation.

Practical illustration of separation using template spectra



Spectral priors based on basis spectra

The GMM does not efficiently model polyphonic musical instruments.

The variance V_{jnf} of each source is then better represented as the linear combination of K **basis spectra** w_{jkn} multiplied by **time-varying scale factors** h_{jkn}

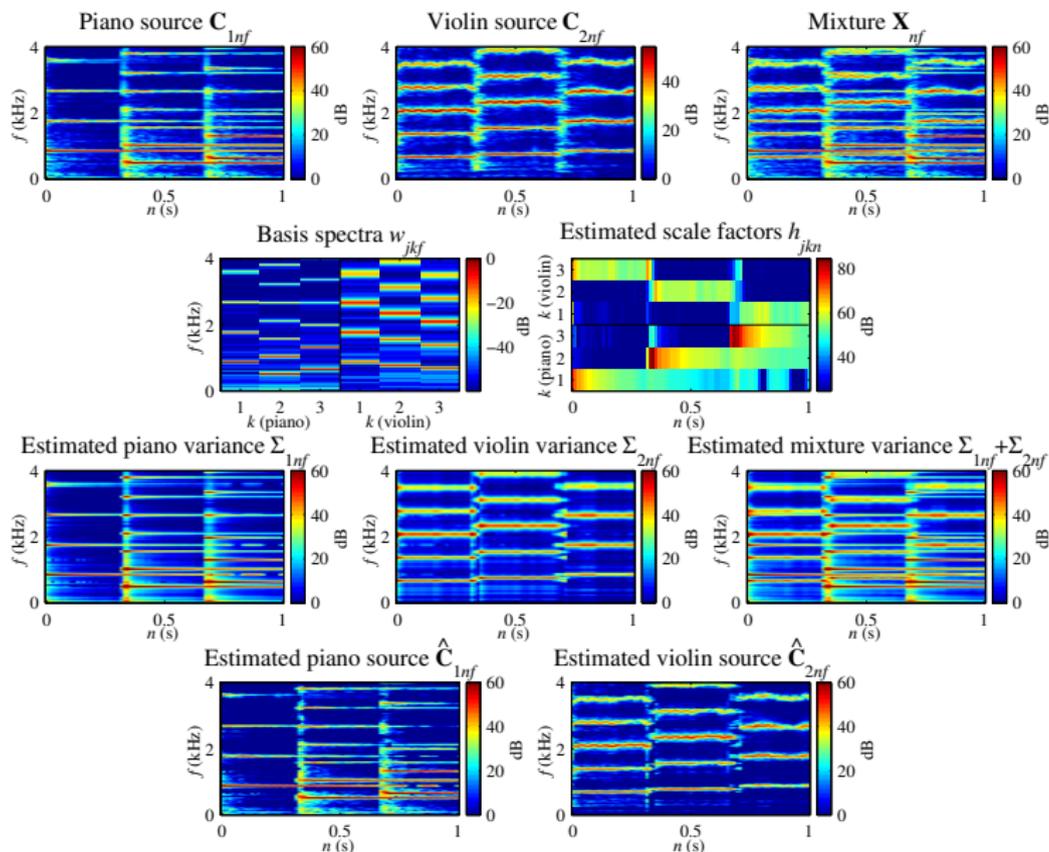
$$V_{jnf} = \sum_{k=1}^K h_{jkn} w_{jkn}$$

This model is also called nonnegative matrix factorization (NMF).

Again, a range of strategies have been used to learn these spectra:

- instrument-dependent training on separate single-source data,
- MAP adaptation to the mixture using uniform priors,
- MAP adaptation to the mixture using trained priors.

Practical illustration of separation using basis spectra



Constrained template/basis spectra

MAP adaptation or inference of the template/basis spectra is often needed due to

- the lack of training data,
- the mismatch between training and test data.

However, it is often inaccurate: additional constraints over the spectra are needed to further reduce overfitting.

Harmonicity and spectral smoothness constraints

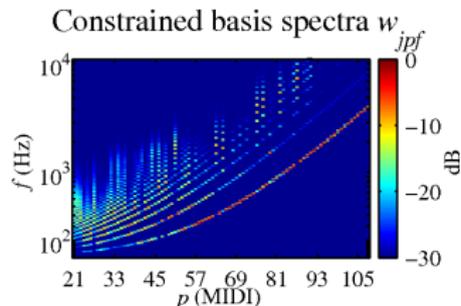
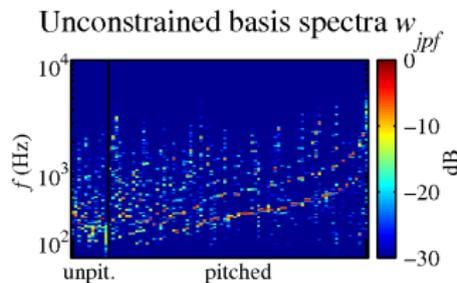
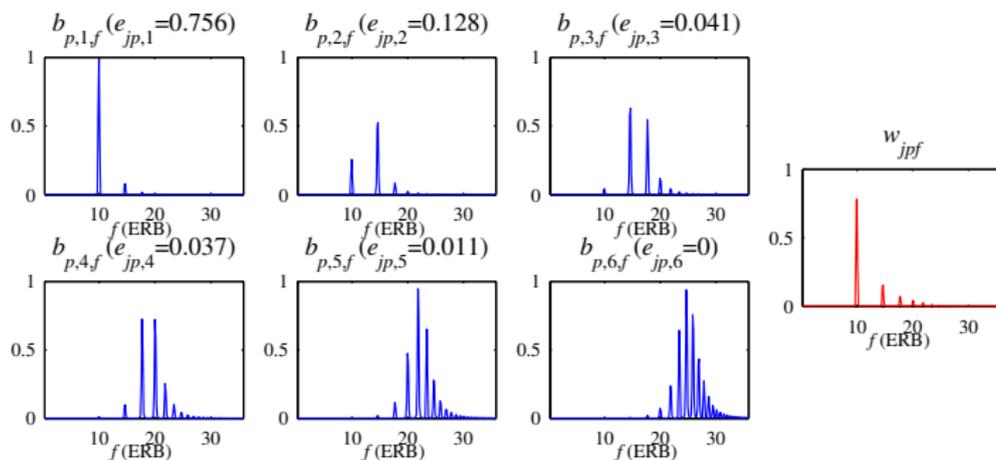
For instance, harmonicity and spectral smoothness can be enforced by

- associating each basis spectrum with some **a priori pitch** p
- modeling w_{jpf} as the sum of **fixed narrowband spectra** b_{plf} representing adjacent partials at harmonic frequencies scaled by **spectral envelope coefficients** e_{jpl}

$$w_{jpf} = \sum_{l=1}^{L_p} e_{jpl} b_{plf}.$$

Parameter estimation now amounts to estimating the active pitches and their spectral envelopes instead of their full spectra.

Practical illustration of harmonicity constraints



Further constraints

Further constraints that have been implemented in this context include

- **source-filter model** of instrumental timbre,
- inharmonicity and tuning.

Probabilistic priors are also popular:

- **state transition** priors

$$P(q_{jn} = k | q_{j,n-1} = l) = \pi_{jkl}$$

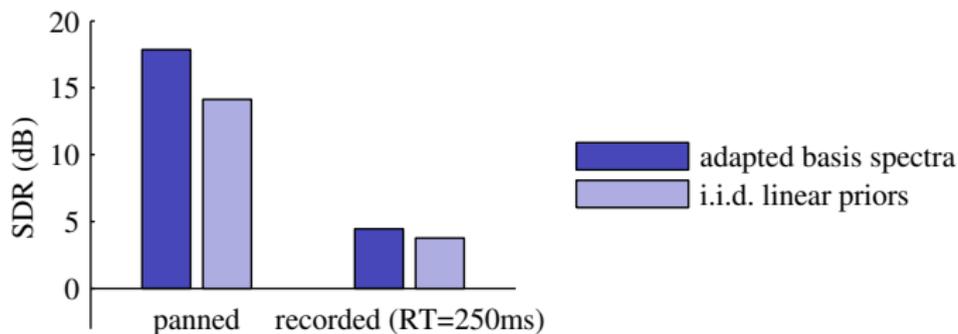
- **spectral continuity** priors (for percussive sounds)

$$P(V_{jnf} | V_{jn,f-1}) = \mathcal{N}(V_{jnf}; V_{jn,f-1}, \sigma_{\text{perc}})$$

- **temporal continuity** priors (for sustained sounds)

$$P(V_{jnf} | V_{j,n-1,f}) = \mathcal{N}(V_{jnf}; V_{j,n-1,f}, \sigma_{\text{sust}})$$

SiSEC results on toy mixtures of 3 sources



Panned mixture

Estimated sources using adapted basis spectra

Estimated sources using i.i.d. linear priors



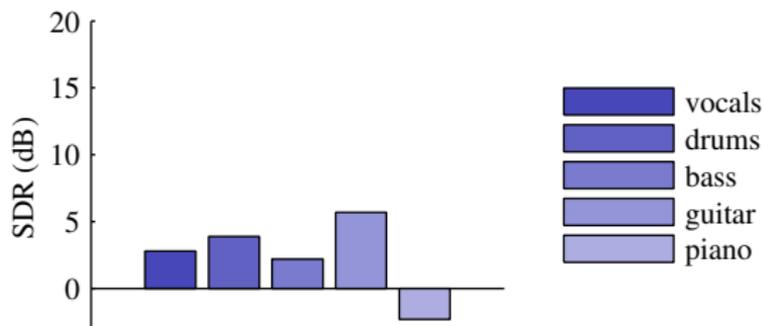
Recorded reverberant mixture

Estimated sources using adapted basis spectra

Estimated sources using i.i.d. linear priors



SiSEC results on professional mixtures



Tamy (2 sources)

Estimated sources using adapted basis spectra



Bearlin (10 sources)

Estimated sources using adapted basis spectra



Summary of probabilistic variance modeling

Advantages:

- top-down approach
- virtually applicable to any mixture, including to diffuse sources
- no hard constraint on the number of sources per time-frequency bin
- fewer musical noise artifacts by joint exploitation of spatial, spectral and learned cues
- principled modular framework for the integration of additional cues

Limitations:

- remaining musical noise artifacts
- current implementations limited to a few spectral and/or spatial cues. . . but this is gradually changing!

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

Summary principles of model-based source separation

Most model-based source separation systems rely on modeling the STFT coefficients of each source as a function of

- a **scalar variable** (S_{jnf} or V_{jnf}) encoding **spectral cues**,
- a **vector or matrix variable** (\mathbf{A}_{jf} or \mathbf{R}_{jf}) encoding **spatial cues**.

Robust source separation requires **priors over both types of cues**:

- spectral cues alone cannot discriminate sources with similar pitch range and timbre,
- spatial cues alone cannot discriminate sources with the same DOA.

A range of informative priors have been proposed, relating for example

- S_{jnf} or V_{jnf} to discrete or continuous latent states,
- \mathbf{A}_{jf} or \mathbf{R}_{jf} to the source DOAs.

Variance modeling outperforms linear modeling.

Conclusion and remaining challenges

To sum up, source separation is a **core problem of audio signal processing** with **huge potential applications**.

Existing systems are **gradually finding their way into the industry**, especially for applications that can accommodate

- a certain amount of musical noise artifacts, such as MIR,
- partial user input/feedback, such as post-production.

We believe that these two **limitations could be addressed in the next 10 years** by exploiting the full power of probabilistic modeling, especially by:

- integrating more and more spatial and spectral cues,
- making a better use of learned cues, using training data or repeated sounds

References

D.L. Wang and G.J. Brown, Eds.

Computational Auditory Scene Analysis: Principles, Algorithms and Applications

Wiley/IEEE Press, 2006.

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: Principles, Algorithms and Systems*

IGI Global, 2010.

2008 and 2010 Signal Separation Evaluation Campaigns

<http://sisec.wiki.irisa.fr/>