Audio Source Separation Using Hierarchical Phase-Invariant Models

Emmanuel Vincent

METISS Group, IRISA-INRIA, Rennes, France

1. Model-based audio source separation
2. Linear modeling
3. Hierarchical phase-invariant modeling
4. Summary and future challenges

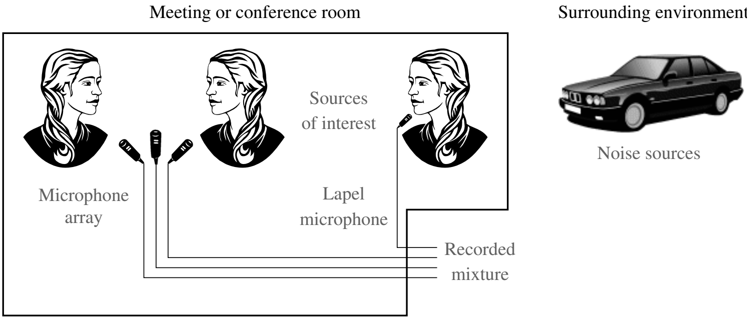
The source separation problem

Many sound scenes are mixtures of several concurrent sound sources.

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

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

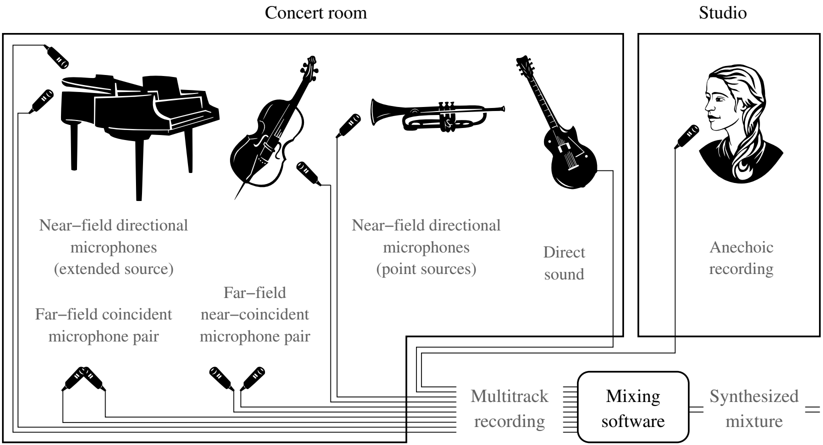
Scenario 1: recorded speech mixture



Example applications:

- speech enhancement,
- automatic speech and speaker recognition.

Scenario 2: synthesized music mixture

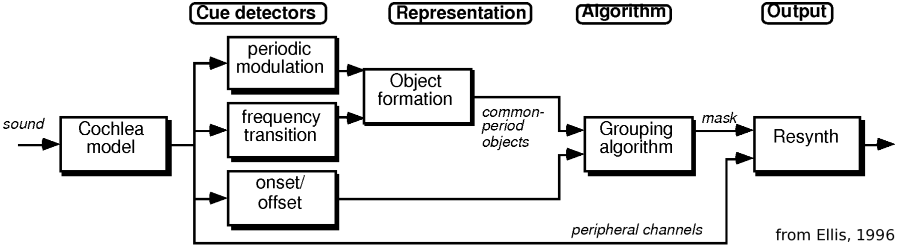

Concert room
Studio
Near–field directional microphones (extended source)
Far–field coincident microphone pair
Far–field near–coincident microphone pair
Near–field directional microphones (point sources)
Direct sound
Anechoic recording
Multitrack recording
Mixing software
Synthesized mixture

Example applications:

- post-production and multichannel upmixing,
- browsing by content.

Computational auditory scene analysis (CASA)

CASA achieves source separation by emulating the human auditory system.



from Ellis, 1996

Source segregation cues are exploited in some precedence order, assuming that small time-frequency regions are dominated by a single source.

This bottom-up approach is fast but not robust.

Model-based audio source separation

An alternative approach consists of finding the source signals that best fit 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.

This top-down approach is more robust than CASA since

- all source segregation cues are exploited at the same time,
- the number of active sources per time-frequency region is not restricted.

Overview of the talk

Hundreds of model-based source separation systems were designed in the last 20 years, based on *e.g.*

- frequency-domain independent component analysis (FDICA),
- sparse component analysis (SCA),
- hidden Markov models (HMM),
- nonnegative matrix factorization (NMF).

30 systems submitted by 15 research groups were evaluated during the 2008 Signal Separation Evaluation Campaign (SiSEC'08).

In this talk, we will

- show that all systems boil down to one of two modeling paradigms: linear modeling *vs.* hierarchical phase-invariant modeling,
- illustrate state-of-the-art performance as measured by SiSEC'08,
- consequently identify the most promising paradigm for future research.

For simplicity, we will focus on stereo (two-channel) mixtures.

1. Model-based audio source separation
2. Linear modeling
3. Hierarchical phase-invariant modeling
4. Summary and future challenges

Paradigm 1: 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}: jth source STFT coeff.
$\mathbf{A}_{jf}$: jth mixing vector

Priors over the mixing vectors

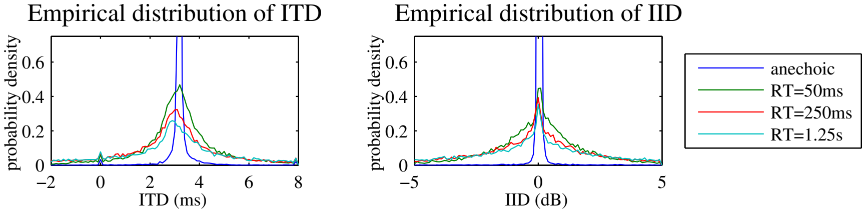
The mixing vectors $\mathbf{A}_{jf}$ encode the apparent sound direction in terms of

- interchannel time difference (ITD) g_{jf},
- interchannel intensity difference (IID) τ_{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 smeared distribution $P(\mathbf{A}_{jf}|\theta_j)$



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

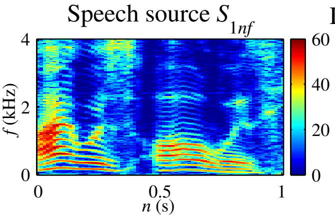
Most systems assume that the sources have random spectra, *i.e.* their STFT coefficients S_{jnf} are independent and identically distributed (i.i.d.).

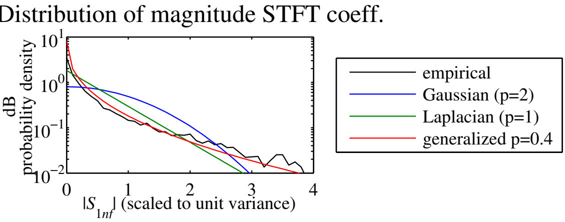
The magnitude STFT coefficients of audio sources are sparse: at each frequency, few coefficients have large values while most are close to zero.

This property is well modeled by the generalized exponential distribution

$$P(|S_{jnf}||p,\beta_f) = \frac{p}{\beta_f \Gamma(1/p)} e^{-\left|\frac{S_{jnf}}{\beta_f}\right|^p}$$

p: shape parameter
β_j: scale parameter


Speech source S_{1nf}


Distribution of magnitude STFT coeff.

Coarser binary priors have also been employed.

Inference algorithms

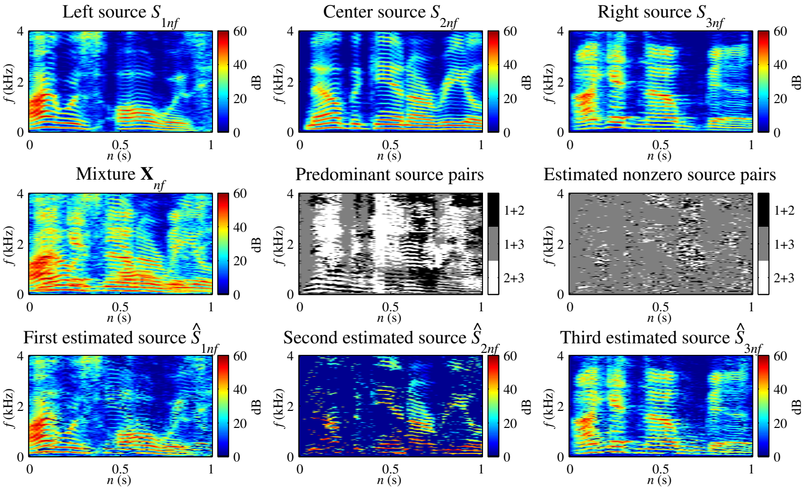
Given the above priors, source separation is typically achieved by joint MAP estimation of the source STFT coefficients S_{jnf} and other latent variables ($\mathbf{A}_{jf}$, g_j, τ_j, p, β_j) via alternating nonlinear optimization.

This objective is called sparse component analysis (SCA).

For typical values of p ($p < 0.74$), the MAP source STFT coefficients are nonzero for at most two sources. SCA then consists of finding the pair of sources with minimum ℓ_p norm.

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. priors



Time-frequency bins dominated by the center source are often erroneously associated with the two other sources, leading to musical noise artifacts.

Spectral priors based on arbitrary sound atoms

In order to avoid such errors, the spectral characteristics of each source must be exploited in addition to its spatial characteristics.

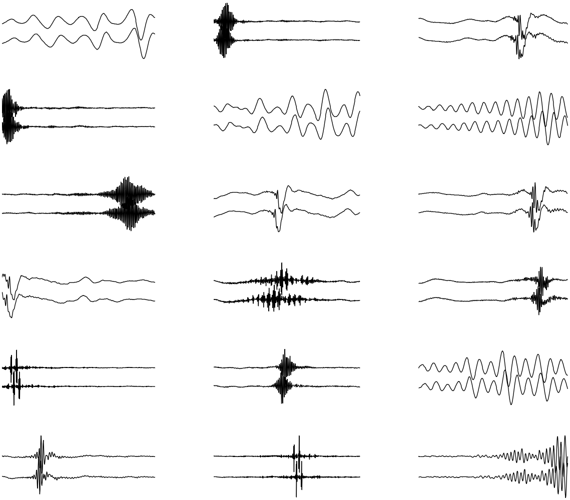
In the framework of linear modeling, this translates into representing each source as a linear combination of a set of wideband sound atoms B_{jknf} weighted by sparse activation weights α_{jk}

$$S_{jnf} = \sum_{k=1}^{K} \alpha_{jk} B_{jknf}$$

Different strategies have been proposed to learn these atoms:

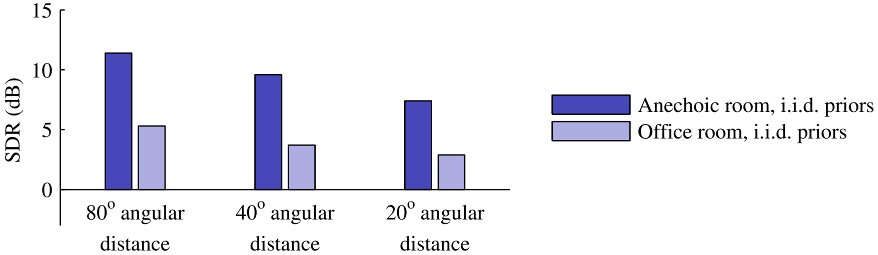
- speaker-independent training on separate single-source data,
- MAP adaptation to one mixture channel with a uniform prior.

Example sound atoms adapted to a speech mixture



from Jafari, 2008

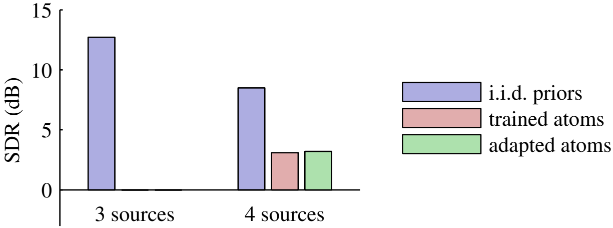
Separation results on recordings of 2 sources



Office recording, 80° angular distance 🔊

Estimated sources using i.i.d. priors 🔊 🔊

Separation results on panned mixtures of 3 or 4 sources



Panned mixture of 4 sources

Estimated sources using i.i.d. priors

trained atoms

Summary limitations of linear modeling

To sum up, state-of-the-art linear modeling-based systems exhibit two theoretical limitations:

- the mixture must consist of non-reverberated point sources,
- at most two sources can be separated in each time-frequency bin and the two predominant sources are often wrongly identified.

These limitations are due respectively to the modeling of the sources as single-channel signals and to intrinsic ambiguities of ITD and IID cues.

1. Model-based audio source separation
2. Linear modeling
3. Hierarchical phase-invariant modeling
4. Summary and future challenges

Idea 1: from sources to source 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 source 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
$\mathbf{C}_{jnf}$: jth source component

Idea 2: translation and phase invariance

In order to overcome the ambiguities of spatial cues, additional spectral cues are needed.

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

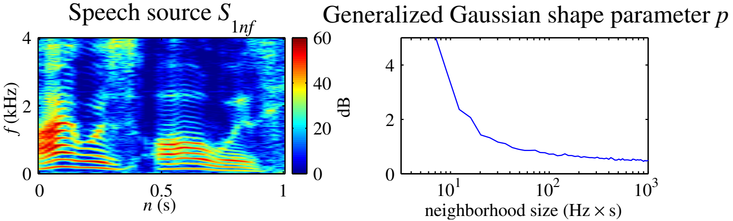
This property is not efficiently accounted for by linear modeling. The number of sound atoms needed to represent a given source is virtually infinite and a finite number of atoms leads to phase ringing artifacts.

Translation and phase invariance must be explicitly accounted for.

Paradigm 2: hierarchical phase-invariant modeling

Hierarchical phase-invariant modeling combines the two ideas by modeling the STFT coefficients of individual source 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.



Log-Gaussian and Poisson distributions over magnitude STFT coefficients have been widely used for single-channel data but do not easily generalize to multichannel data.

The Gaussian model

The zero-mean Gaussian distribution provides a simpler model.

$$P(\mathbf{C}_{jnf}|\mathbf{\Sigma}_{jnf}) = \frac{1}{\det(\pi\mathbf{\Sigma}_{jnf})} e^{-\mathbf{C}_{jnf}^H \mathbf{\Sigma}_{jnf}^{-1} \mathbf{C}_{jnf}}$$

$\mathbf{\Sigma}_{jnf}$: *jth* component covariance matrix

The covariance matrix $\mathbf{\Sigma}_{jnf}$ of each source 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

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

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

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

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

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

d: microphone spacing
c: sound speed

More accurate models remain to be found.

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.

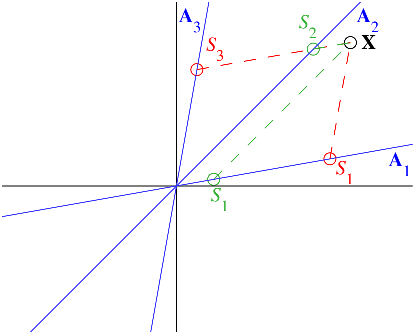
Discrete priors constraining the number of nonzero variances to one or two have also been employed.

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

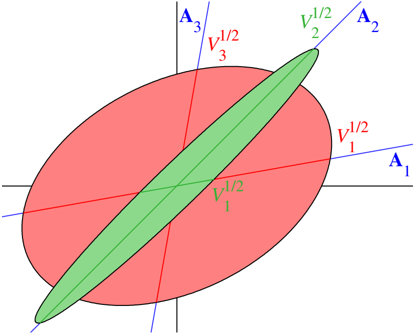
Benefit of exploiting interchannel coherence

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

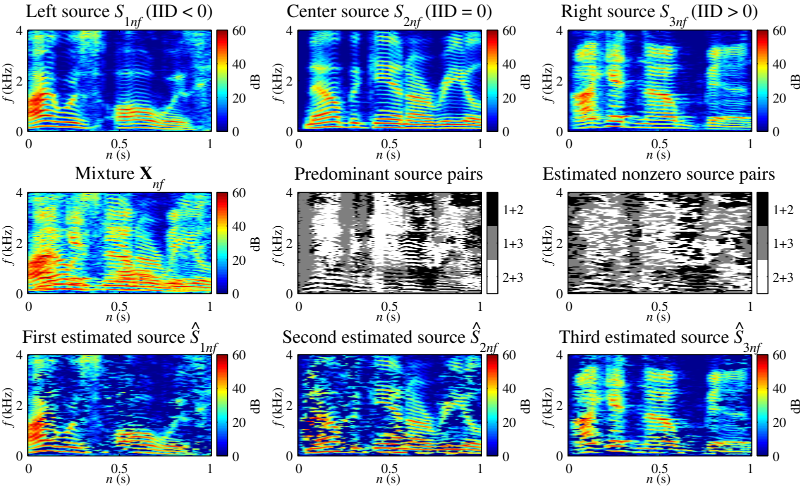
Linear model



Covariance model



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



Spectral priors based on template spectra

Hierarchical modeling eases 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}

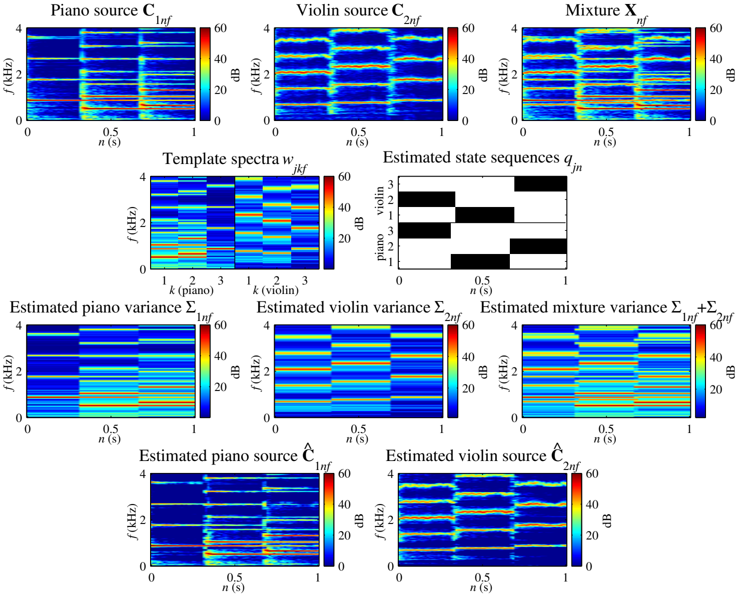
$$V_{jnf} = w_{jq_{jn}f} \text{ with } P(q_{jn} = k) = \pi_{jk}$$

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.

Smooth autoregressive (AR) parameterization of the template spectra has also been studied so as to avoid overfitting to the fundamental frequencies of sounds in the training data.

Practical illustration of separation using template spectra



Spectral priors based on basis spectra

The GMM does not efficiently model polyphonic musical instruments such as the piano, which can play several notes at a time.

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

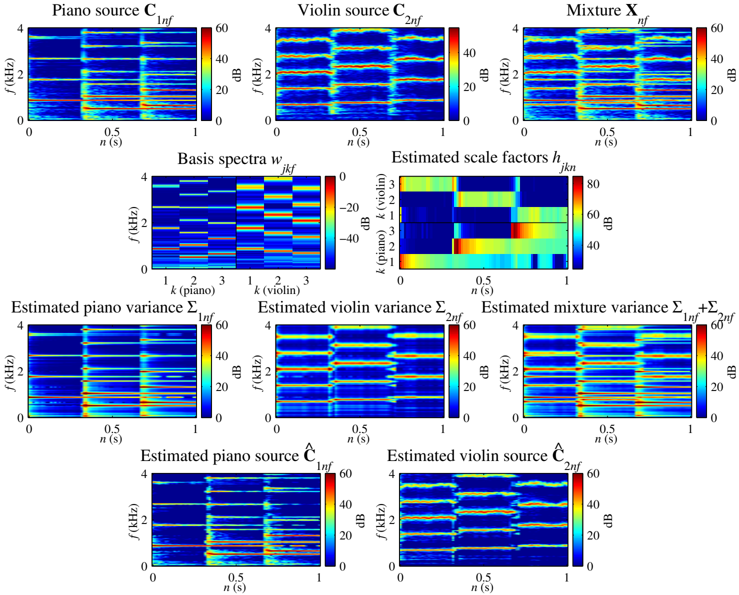
$$V_{jnf} = \sum_{k=1}^{K} h_{jkn} w_{jkf}$$

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



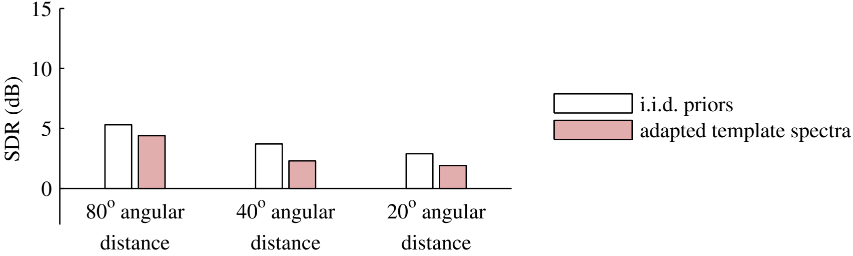
Reported performance improvements

Compared to SCA, hierarchical modeling-based systems have achieved SDR improvements as reported by their authors on the order of

- +1 dB with i.i.d. priors,
- +2 dB with adapted template spectra,
- up to +10 dB with instrument-dependent learned basis spectra.

Smaller improvements were however measured during SiSEC'08.

Separation results on recordings of 2 sources

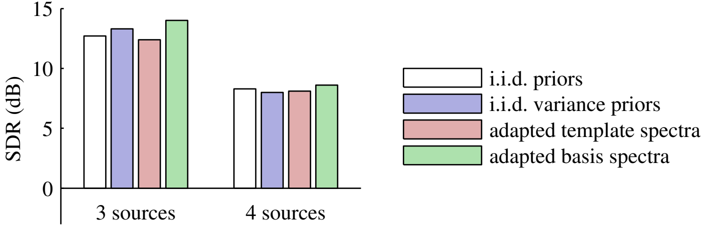


Office recording, 80° angular distance

Estimated sources using i.i.d. priors

adapted template spectra

Separation results on panned mixtures of 3 or 4 sources



Panned mixture of 4 sources 🔊

Estimated sources using i.i.d. priors 🔊 🔊 🔊 🔊

adapted basis spectra 🔊 🔊 🔊 🔊

1. Model-based audio source separation
2. Linear modeling
3. Hierarchical phase-invariant modeling
4. 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.

Summary advantages of hierarchical phase-invariant modeling

Hierarchical phase-invariant modeling exhibits three theoretical advantages compared to linear modeling:

- the mixture may contain diffuse or reverberated sources,
- spatial cues allow the separation of more sources per time-frequency bin and more accurate identification of the predominant sources,
- in addition, phase-invariant spectral cues can be efficiently exploited.

Remaining challenges

Three great challenges remain:

- separate complex mixtures involving many reverberated and/or moving sources,
- build fully blind systems able to separate any mixture without any prior knowledge,
- find a robust output representation so as to integrate source separation within a range of applications.

The Bayesian framework provides a principled way of tackling these challenges by

- designing higher-level priors based on meaningful latent variables,
- building modular systems combining alternative priors and enabling them to select the most appropriate priors for the mixture at hand,
- computing the posterior probability of the separated sources.

Conclusion

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

Existing systems have found few practical applications yet, due to their

- insufficient performance on real-world mixtures,
- need for prior knowledge,
- poor application integration.

We believe that these limitations could be addressed in the next 5 to 10 years by exploiting the full power of Bayesian inference as applied to hierarchical phase-invariant models.

For more information

`http://sisec.wiki.irisa.fr/`

`emmanuel.vincent@irisa.fr`