IRISA/D5 Thematic Seminar

Source separation

Emmanuel Vincent

Inria Nancy – Grand Est

Source separation

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

Overview

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

...but few are yet applicable to real-world audio, as illustrated by the recent Signal Separation Evaluation Campaigns (SiSEC).

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

- computational auditory scene analysis (CASA),
- beamforming and post-filtering,
- 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. Beamforming and post-filtering
2. Probabilistic linear modeling
3. Probabilistic variance modeling
4. Summary

Paradigm 1: separation of a target source in ambient noise

Early studies on array processing focused on the extraction of a target point source in ambient noise.

In each time-frequency bin (n, f), the model used for source localization is assumed here again

$$\mathbf{X}_{nf} = S_{nf}\mathbf{D}_f + \mathbf{B}_{nf}$$

$\mathbf{X}_{nf}$: mixture STFT coeff.
S_{nf}: target STFT coeff.
$\mathbf{D}_f$: steering vector
$\mathbf{B}_{nf}$: ambient noise

where the steering vector $\mathbf{D}_f$ encode the ITDs τ_i and the IIDs g_i between the I microphones

$$\mathbf{D}_f \propto \begin{pmatrix} 1 \\ g_2 e^{-2i\pi f \tau_2} \\ \vdots \\ g_I e^{-2i\pi f \tau_I} \end{pmatrix}$$

Beamforming and post-filtering

The optimal linear estimator in the minimum mean square error (MMSE) sense is the multichannel Wiener filter

$$\widehat{S}_{nf} = V_{nf}^{S} \mathbf{D}_f^H (\Sigma_{nf}^{\mathbf{X}})^{-1} \mathbf{X}_{nf}$$

where V_{nf}^{S} is the variance of S_{nf} and $\Sigma_{nf}^{\mathbf{X}}$ the covariance of $\mathbf{X}_{nf}$.

This estimator is in fact the combination of

- a multichannel spatial filter known as the minimum variance distortionless response (MVDR) beamformer

$$Y_{nf} = \frac{\mathbf{D}_f^H (\Sigma_{nf}^{\mathbf{X}})^{-1} \mathbf{X}_{nf}}{\mathbf{D}_f^H (\Sigma_{nf}^{\mathbf{X}})^{-1} \mathbf{D}_f}$$

- a single-channe spectral filter known as the Wiener post-filter

$$\widehat{S}_{nf} = \frac{V_{nf}^{S}}{V_{nf}^{Y}} Y_{nf}$$

where V_{nf}^{Y} is the variance of Y_{nf} and V_{nf}^{S}/V_{nf}^{Y} is the SNR.

Estimation algorithms

The steering vector $\mathbf{D}_f$ is derived from the spatial position of the target obtained via a source localization algorithm.

The covariance of the mixture $\Sigma_{nf}^{\mathbf{X}}$ is computed empirically by local averaging of squared STFT coefficients in the time-frequency plane.

The variance of the target is often estimated by spectral subtraction

$$V_{nf}^S = \max\{0, V_{nf}^Y - V_{nf}^B\}$$

where V_{nf}^B is the assumed noise variance in V_{nf}^Y.

V_{nf}^B is estimated for example by the MCRA method for silence detection.

Summary of beamforming and post-filtering

The algorithms stemming from paradigm 1 exhibit two main limitations:

- performance is very sensitive to localization accuracy,
- the MCRA algorithm assumes quasi-stationary noise and fails in a multi-source context where V_{nf}^{B} varies a lot from one time frame to the next.

In order to overcome the latter limitation, multiple sources must be explicitly modeled.

Paradigm 2: 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)

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

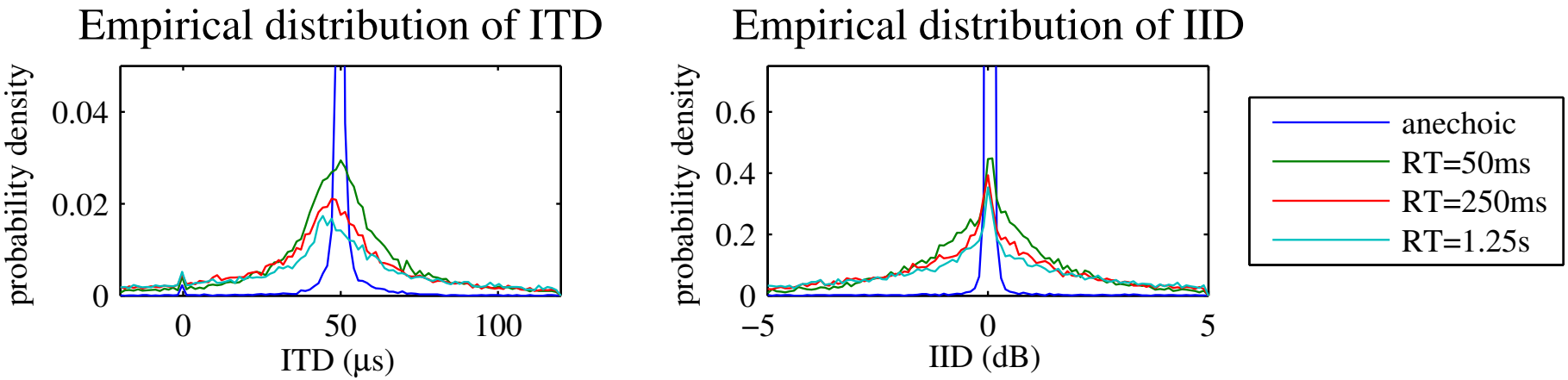
$\mathbf{X}_{nf}$: vector of mixture STFT coeff.
J: number of sources
S_{jnf}: jth source STFT coeff.
$\mathbf{A}_{jf}$: jth mixing vector

Modeling of the mixing vectors

The mixing vectors $\mathbf{A}_{jf}$ encode the ITD and IID of each source at each frequency.

For anechoic mixtures, $\mathbf{A}_{jf}$ is equal to the steering vector $\mathbf{D}_{jf}$.

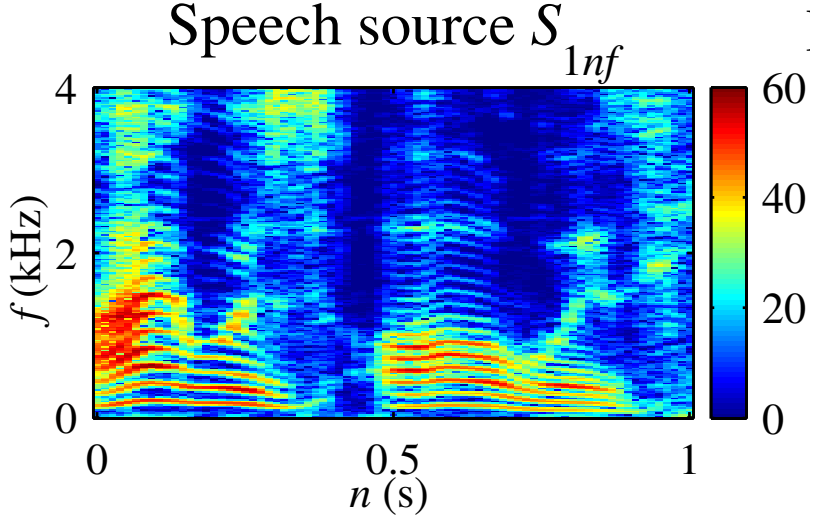
For echoic mixtures, ITDs and IIDs follow a smeared distribution $P(\mathbf{A}_{jf}|\theta_j)$



Sparsity of the source STFT coefficients

Let us suppose for the moment that the source STFT coefficients S_{jnf} are independent and identically distributed (i.i.d.).

These coefficients are sparse: at each frequency, a few coefficients are large and most are close to zero.



Sparse i.i.d. modeling of the sources

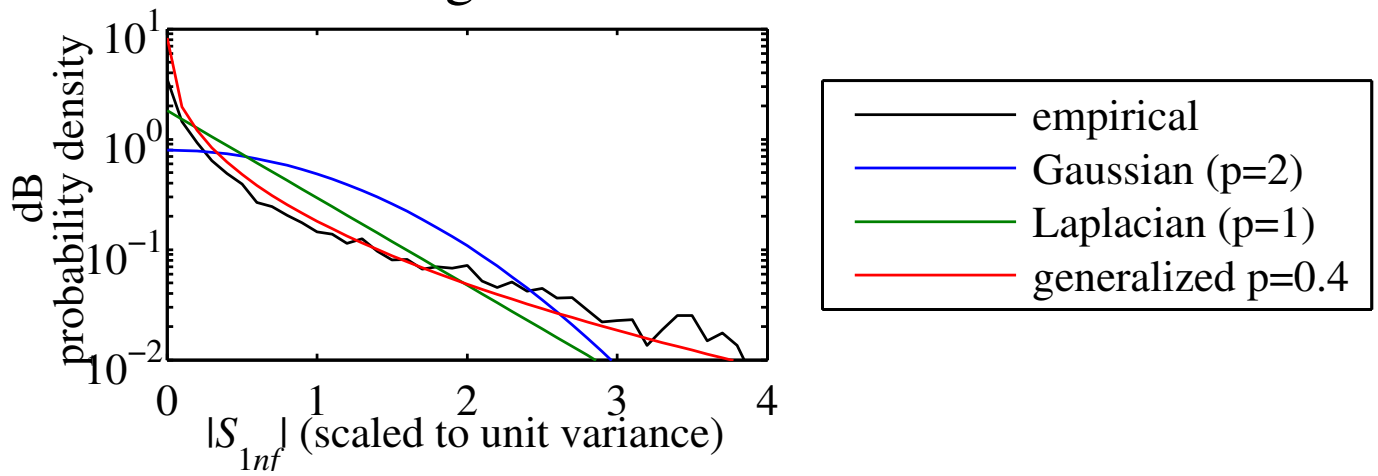
This property can be modeled in several ways:

- binary masking: single active source j_{nf}^{act} in each time-frequency bin with, e.g., uniform $P(j_{nf}^{\mathrm{act}})$,
- 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

Distribution of magnitude STFT coeff.



Inference algorithms

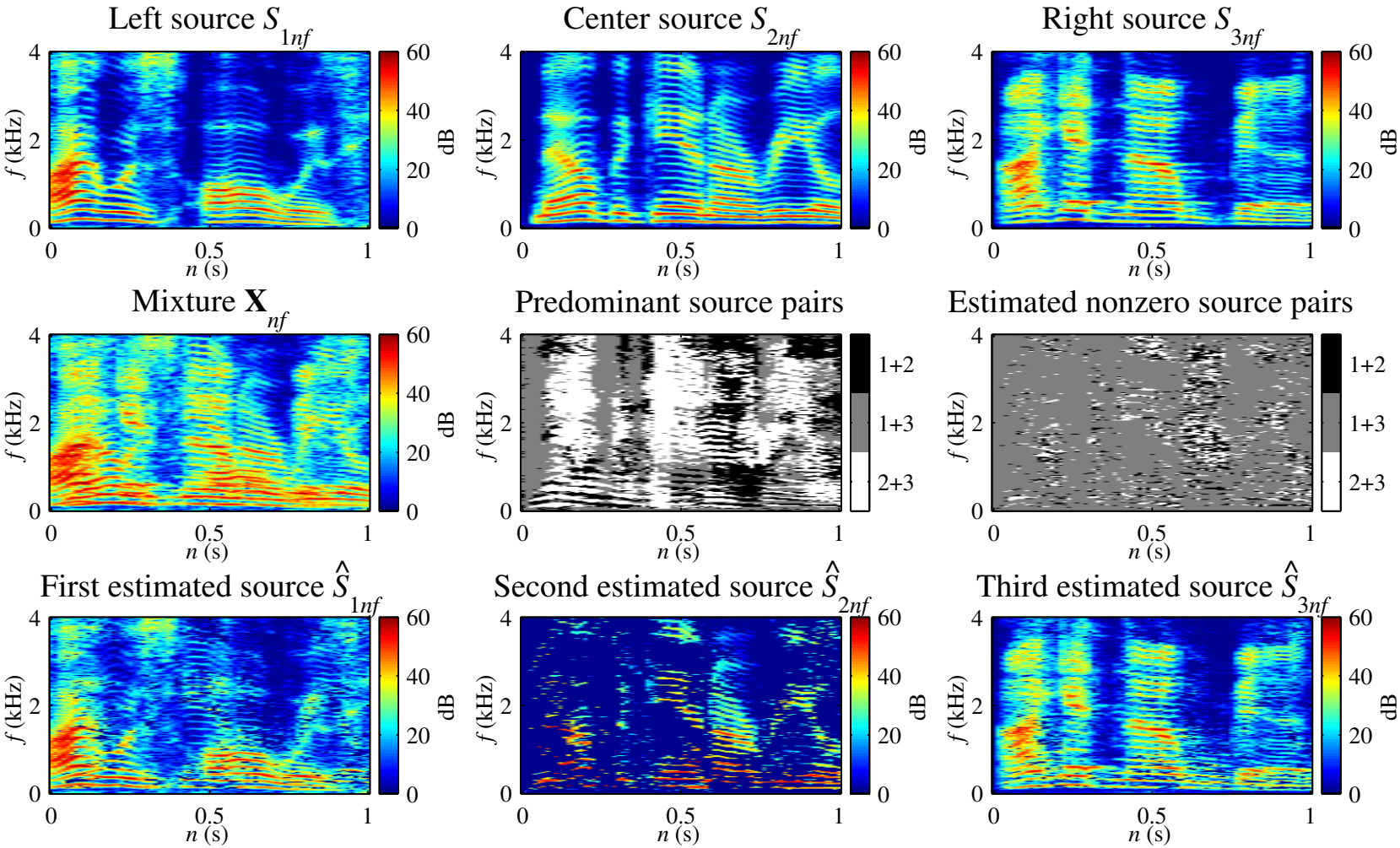
Given the above priors, source separation is typically achieved by joint MAP estimation of the source STFT coefficients S_{jnf} and other latent variables ($\mathbf{A}_{jf}$, g_j, τ_j, p, β_j) via alternating nonlinear optimization.

This objective is called sparse component analysis (SCA).

For typical values of p, the MAP source STFT coefficients are nonzero for at most I sources.

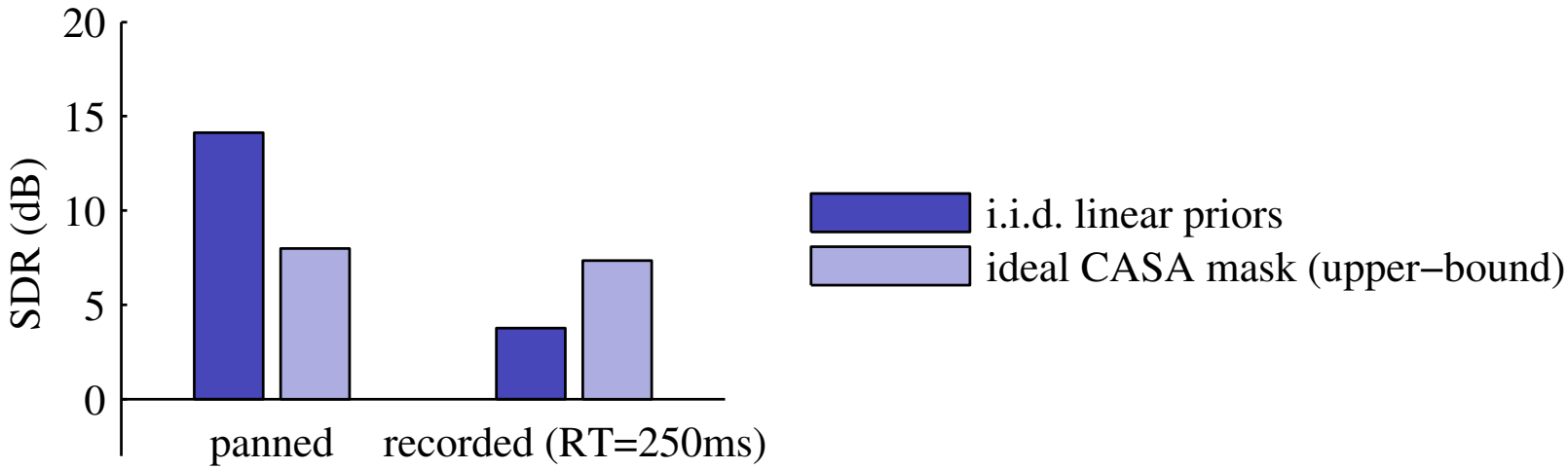
When the number of sources is $J = I$, SCA is renamed nongaussianity-based frequency-domain independent component analysis (FDICA).

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 music mixtures



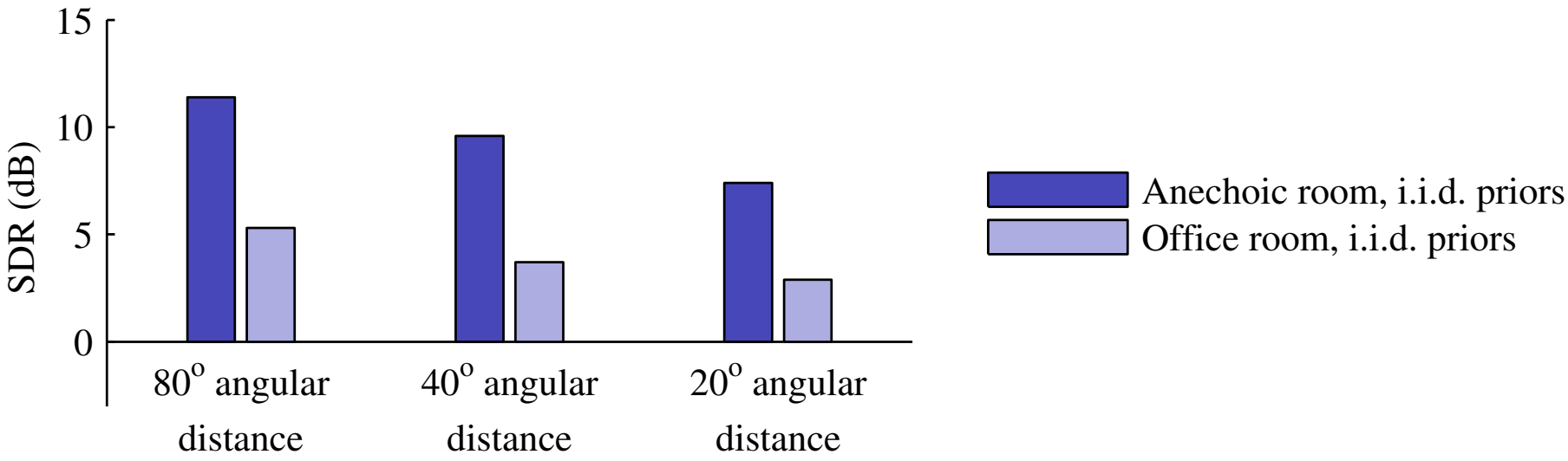
Panned mixture

Estimated sources using i.i.d. linear priors

Recorded reverberant mixture

Estimated sources using i.i.d. linear priors

SiSEC results on speech mixtures



Anechoic recording, 80° spacing

Estimated sources

Office recording, 80° spacing

Estimated sources

Summary of probabilistic linear modeling

Advantages:

- explicitly models multiple sources

Limitations:

- restricted to mixtures of non-reverberated point sources
- the sources must have different spatial cues (ITD, IID)
- at most two sources can be separated in each time-frequency bin, and their are often badly identified due to the ambiguities of spatial cues

Idea 1: from sources to source spatial images

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 the mixture, a.k.a. its spatial image.

Background noise becomes a source as any other.

Source separation becomes the problem of estimating the spatial images of all sources.

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 spatial image

Idea 2: translation and phase invariance

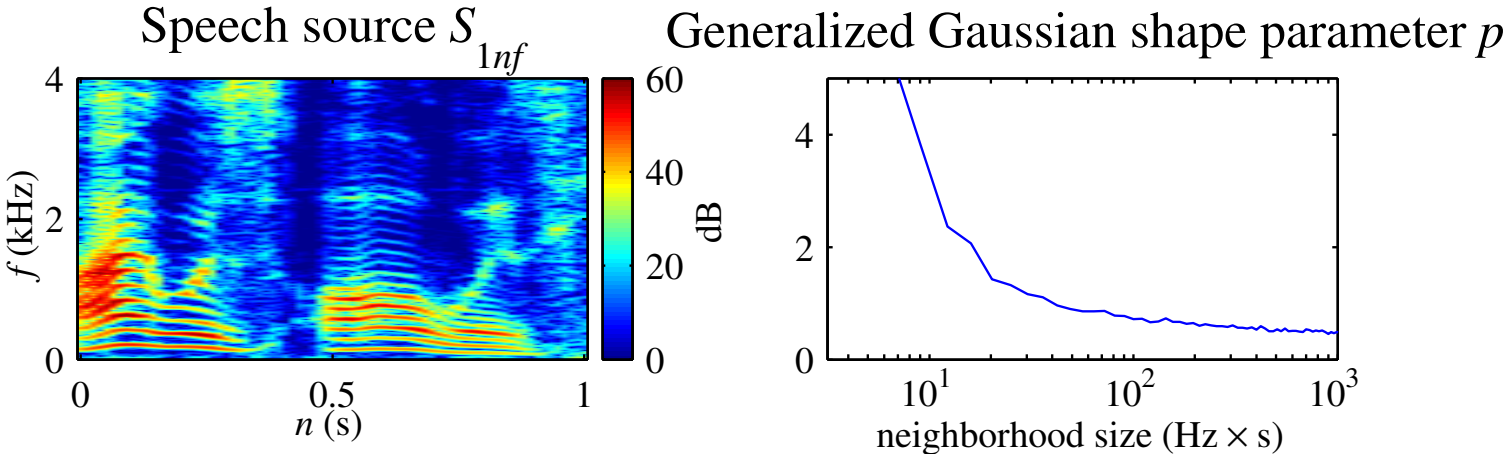
In order to overcome the ambiguities of spatial cues, additional spectral cues are needed as shown by CASA.

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

Paradigm 3: variance modeling

Variance modeling combines these two ideas by modeling the STFT coefficients of individual source spatial images 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.

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

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

The covariance matrix $\mathbf{\Sigma}_{jnf}$ of each source can be factored as the product of a scalar nonnegative variance V_{jnf} and a spatial 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 source 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 spatial covariance

The spatial 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 1, *i.e.*, $\mathbf{R}_{jf}$ has rank 1

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

In this case, the prior distributions $P(\mathbf{A}_{jf}|\theta_j)$ used with linear modeling can be reused.

Full-rank spatial covariance

For reverberated or diffuse sources, the interchannel coherence is smaller than 1, *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

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

d: microphone spacing
c: sound speed

Modeling of $\mathbf{R}_{jf}$ as an unconstrained full-rank matrix is also possible.

I.i.d. modeling of the source variances

Baseline systems rely model the source variances V_{jnf} as i.i.d. and locally constant within small time-frequency regions again.

It can then be shown that the MAP variances are nonzero for up to I^2 sources.

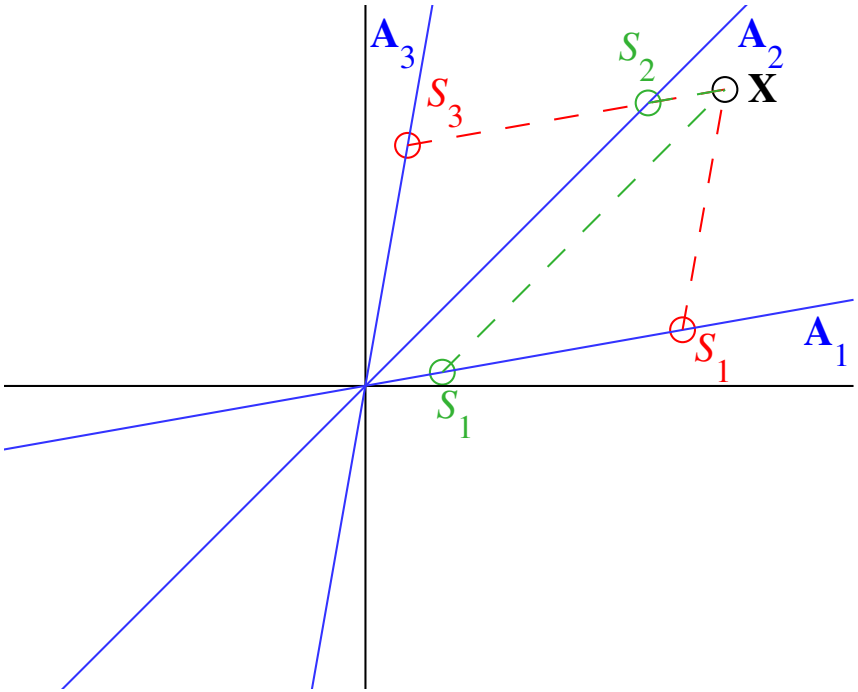
Discrete priors constraining the number of nonzero variances to a smaller number have also been employed.

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

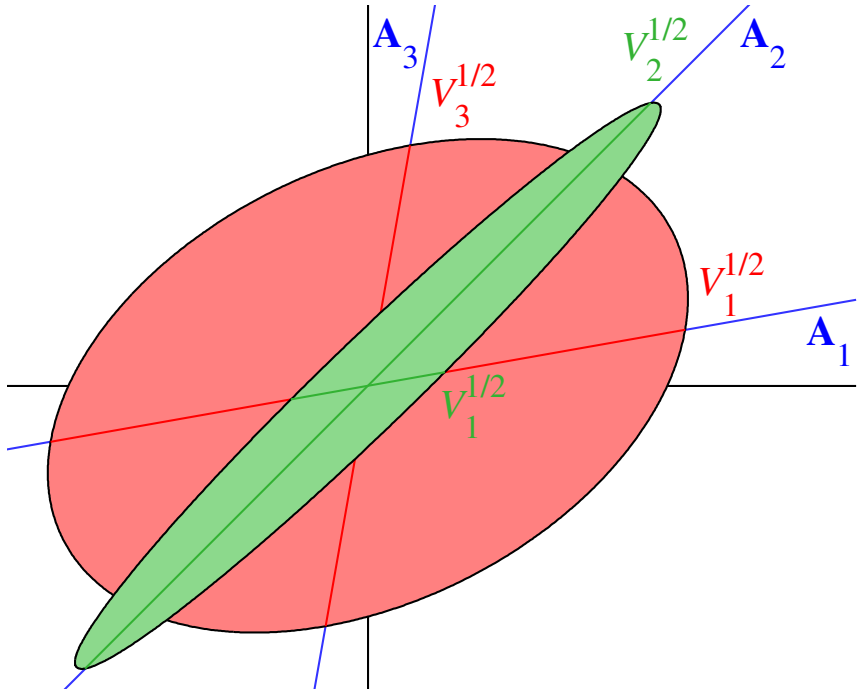
Benefit of exploiting interchannel coherence

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

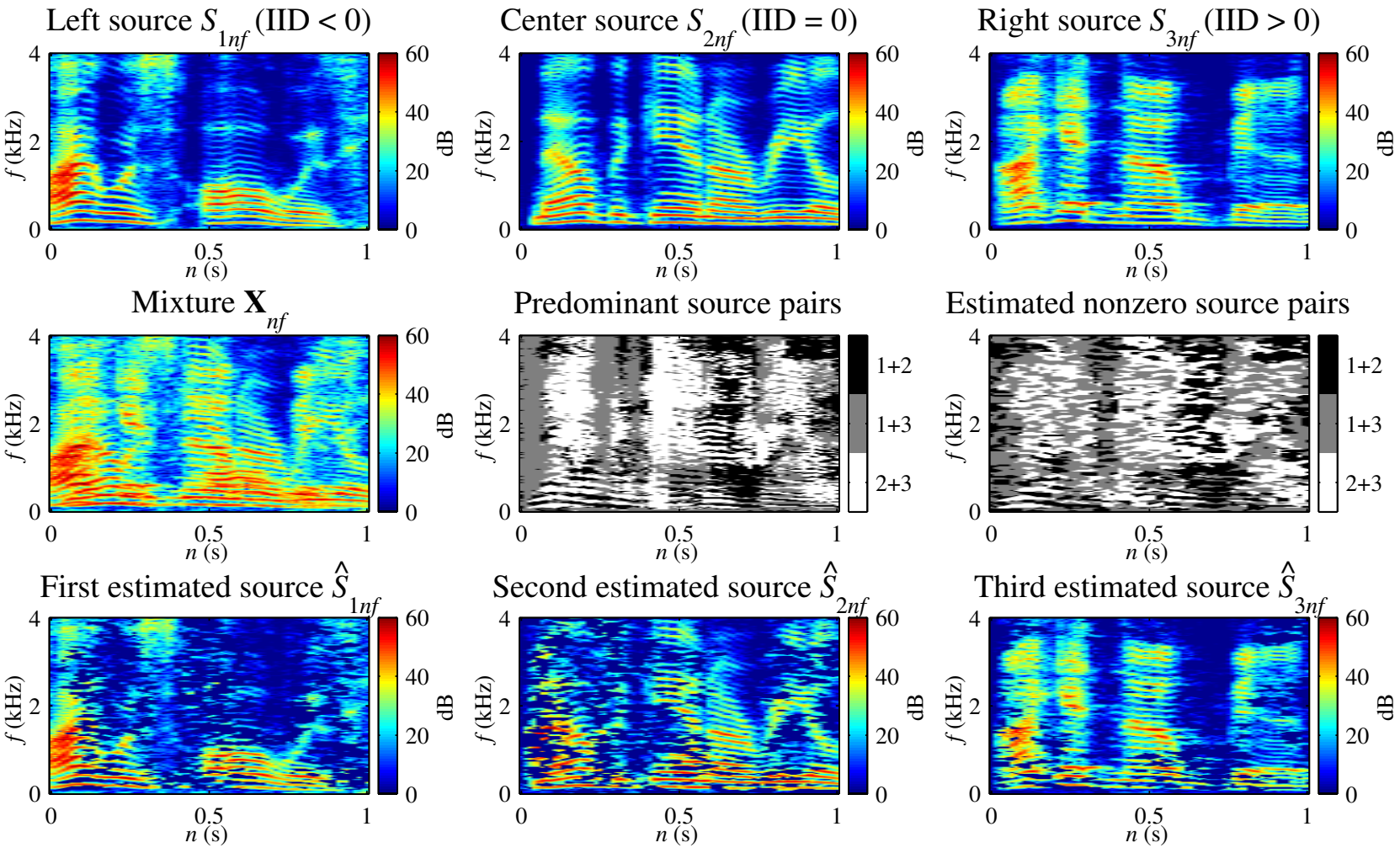
Linear model



Covariance model



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



Spectral modeling using template spectra

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

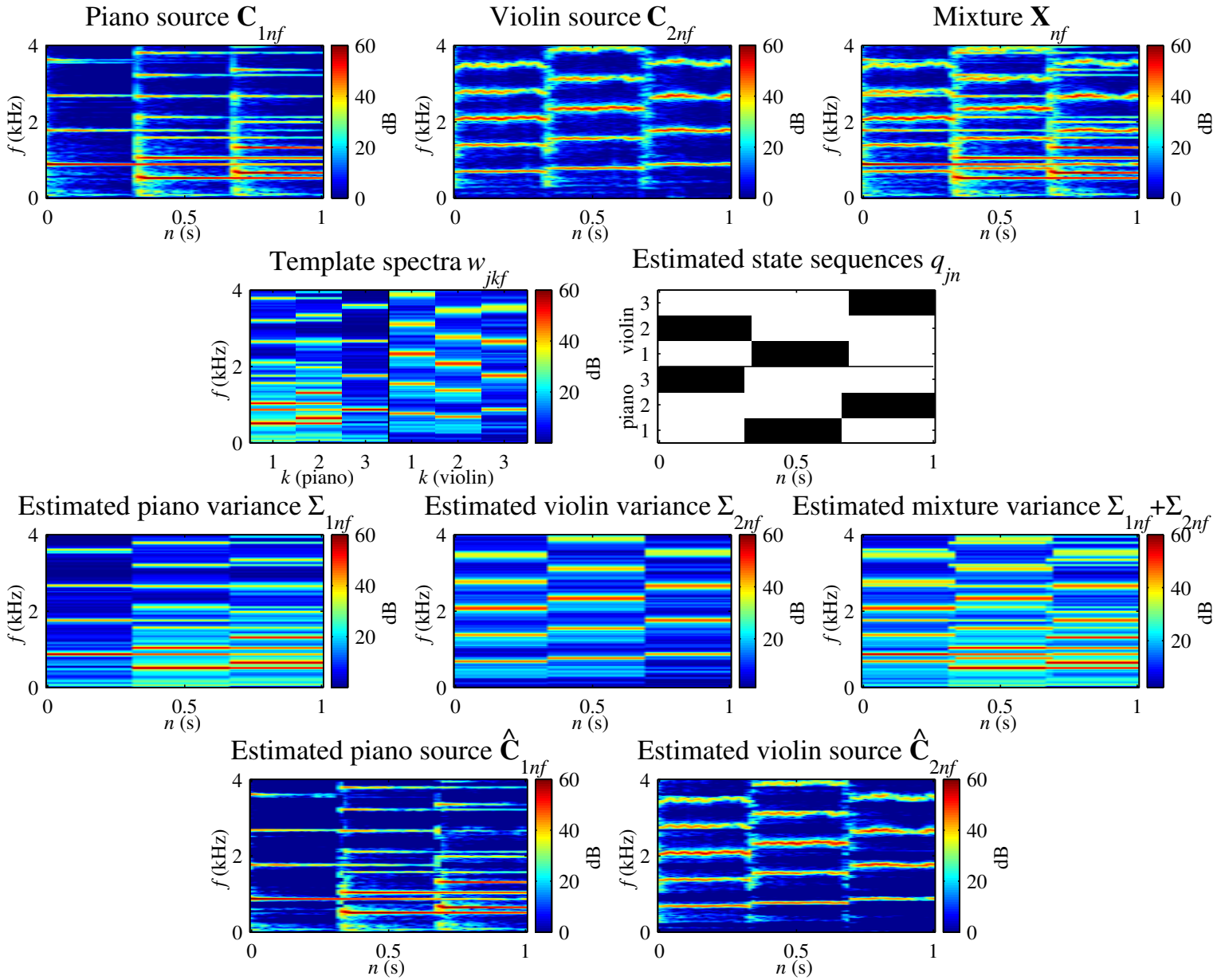
The Gaussian mixture model (GMM) represents the variance V_{jnf} of each source at a given time by one of K template spectra w_{jkf} indexed by a discrete state q_{jn}

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

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 modeling using basis spectra

The GMM does not efficiently model polyphonic sound sources.

The variance V_{jnf} of each source can be modeled instead as the linear combination of K basis spectra w_{jkf} multiplied by time activation coefficients h_{jkn}

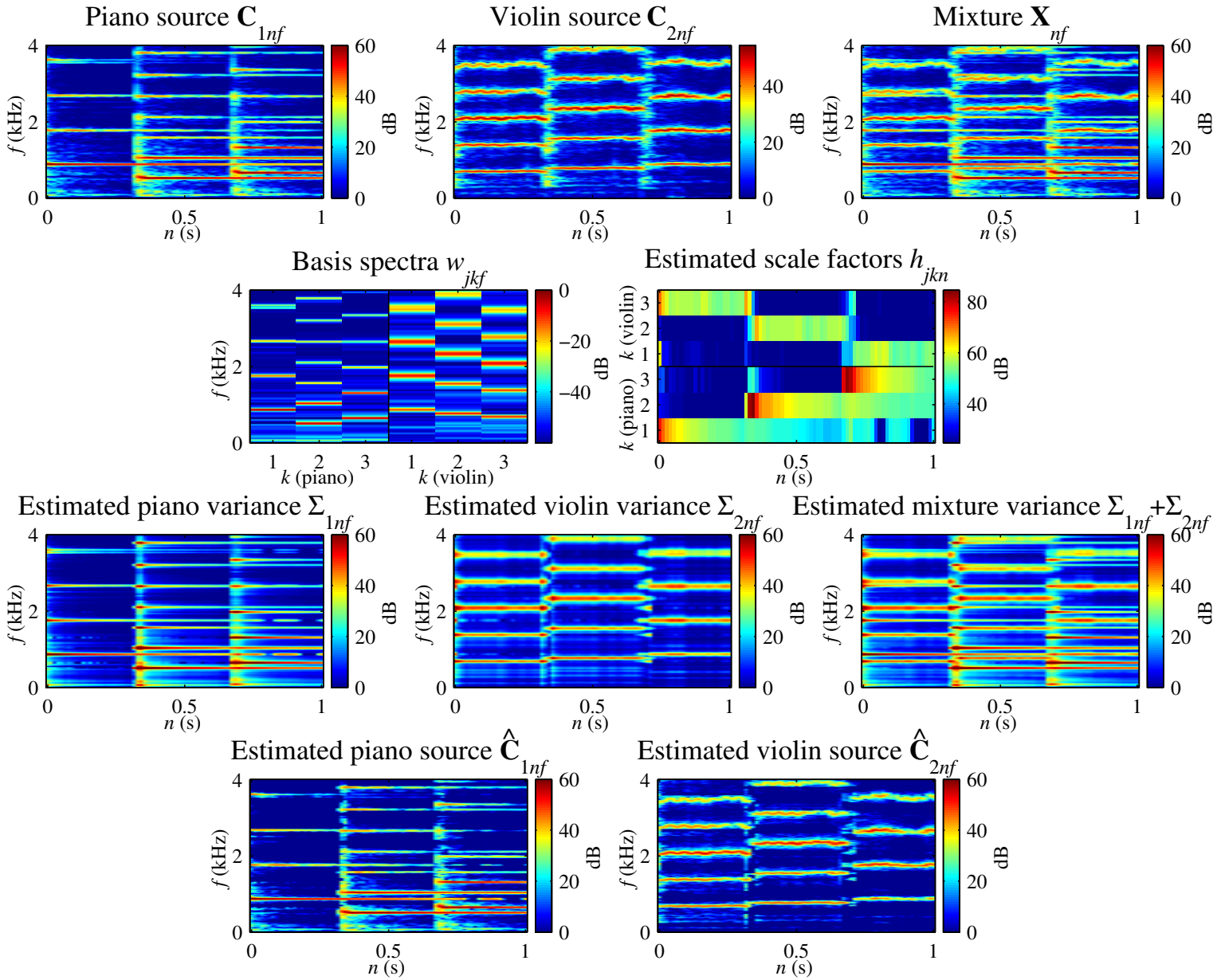
$$V_{jnf} = \sum_{k=1}^{K} h_{jkn} w_{jkf}$$

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

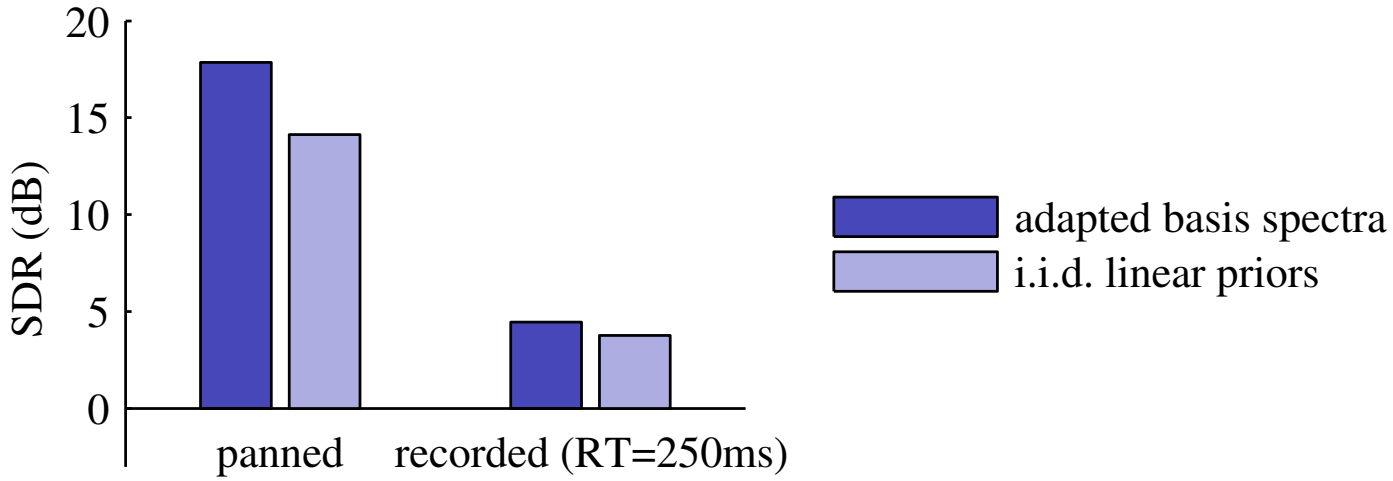
A range of strategies have been used to learn these spectra:

- 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



SiSEC results on music mixtures



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

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

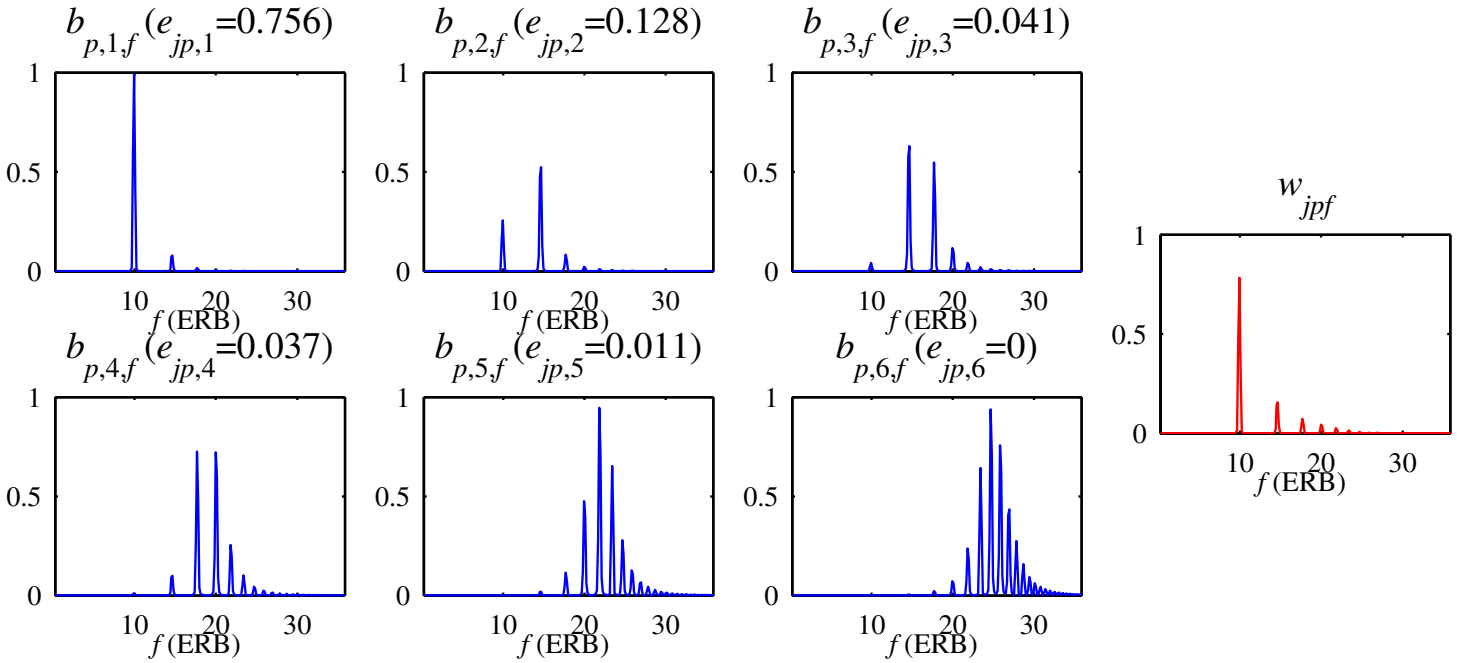
For instance, harmonicity and spectral smoothness can be enforced by

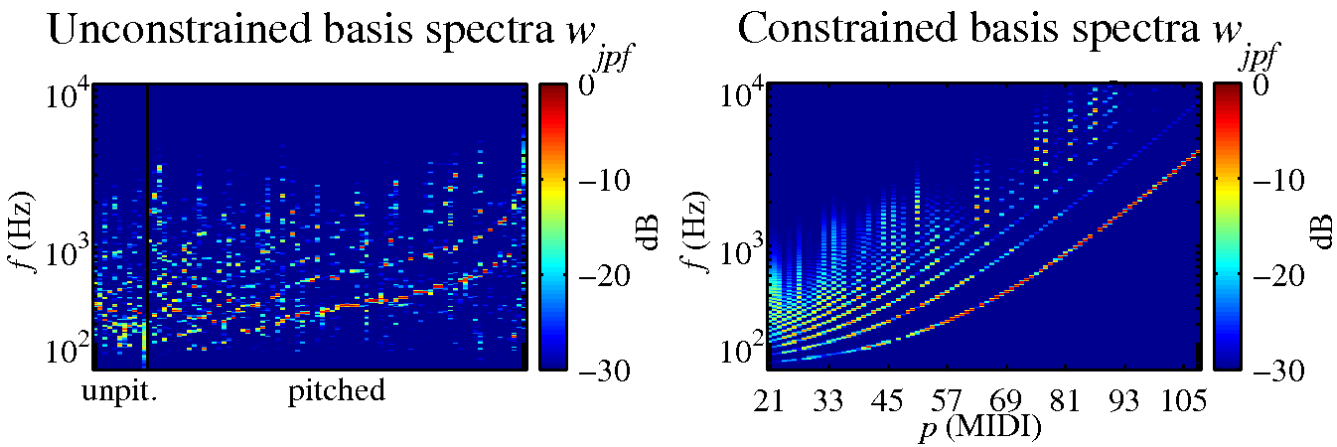
- associating each basis spectrum with some a priori pitch p
- modeling w_{jpf} as the sum of fixed narrowband spectra b_{plf} representing adjacent partials at harmonic frequencies scaled by spectral envelope coefficients e_{jpl}

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

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

Practical illustration of harmonicity constraints



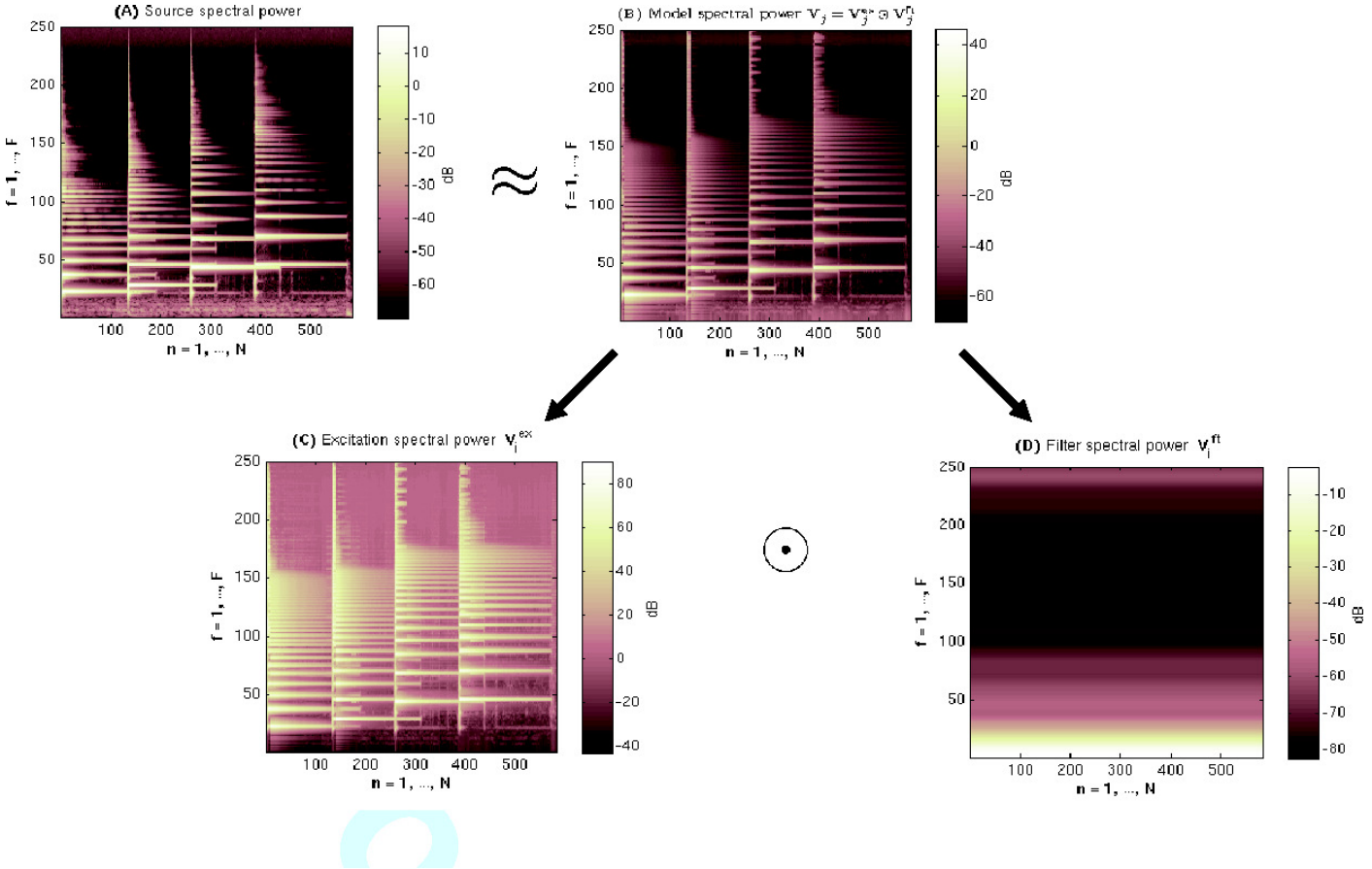


A flexible spectral model

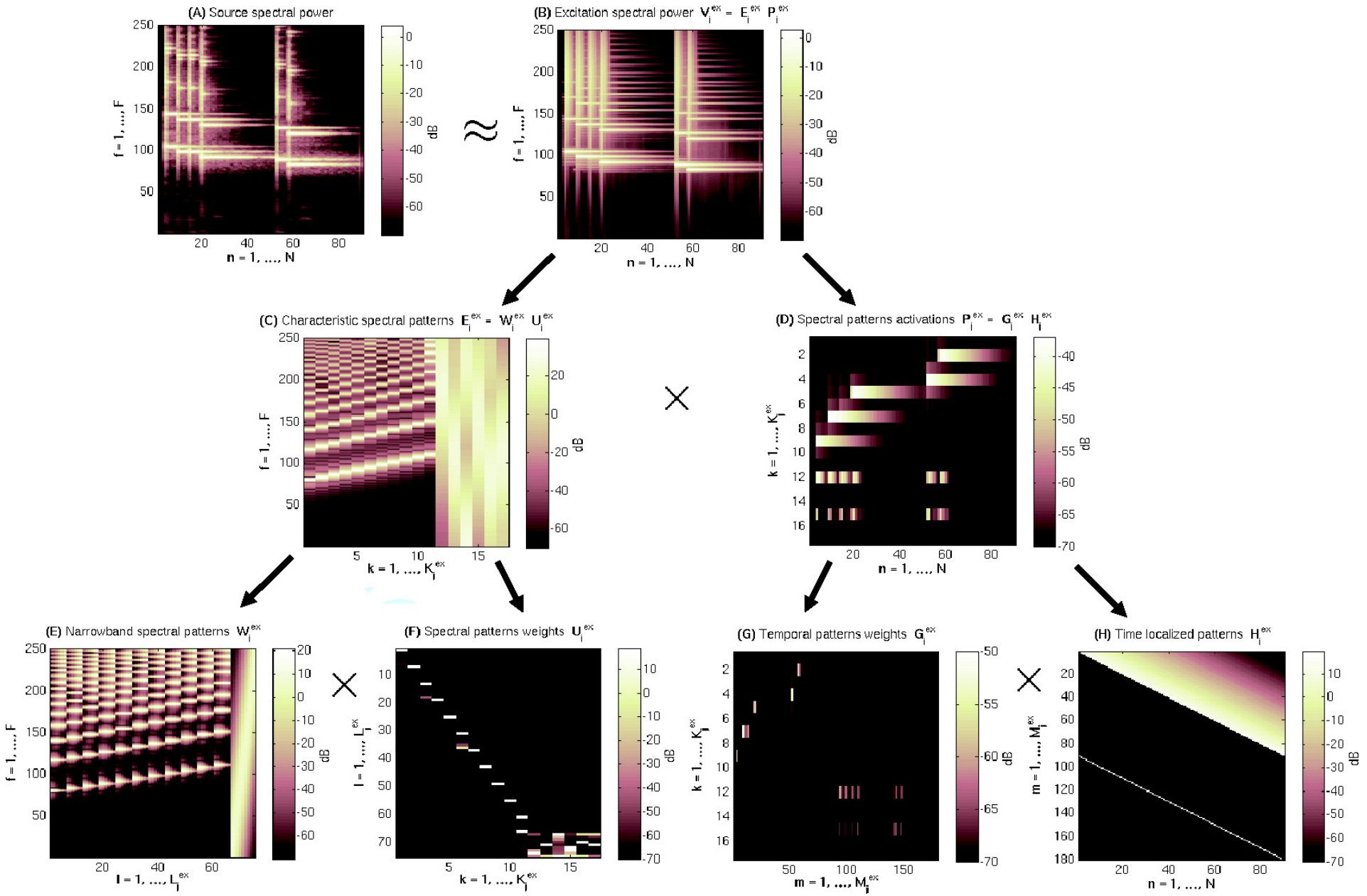
We have built upon this idea and proposed a flexible framework enabling the joint exploitation of a wide range of cues by:

- factorization of the variance assuming the excitation-filter model $V_{jnf} = V^{\mathrm{ex}}_{jnf} V^{\mathrm{ft}}_{jnf}$
- further factorization of each part into basis spectra and time activation coefficients *e.g.* $V^{\mathrm{ex}}_{jnf} = \sum_k h^{\mathrm{ex}}_{jkn} w^{\mathrm{ex}}_{jkf}$
- further factorization of the basis spectra and time activation series into fine structure and envelope coefficients *e.g.* $w^{\mathrm{ex}}_{jkf} = \sum_l e^{\mathrm{ex}}_{jlk} f^{\mathrm{ex}}_{jlf}$

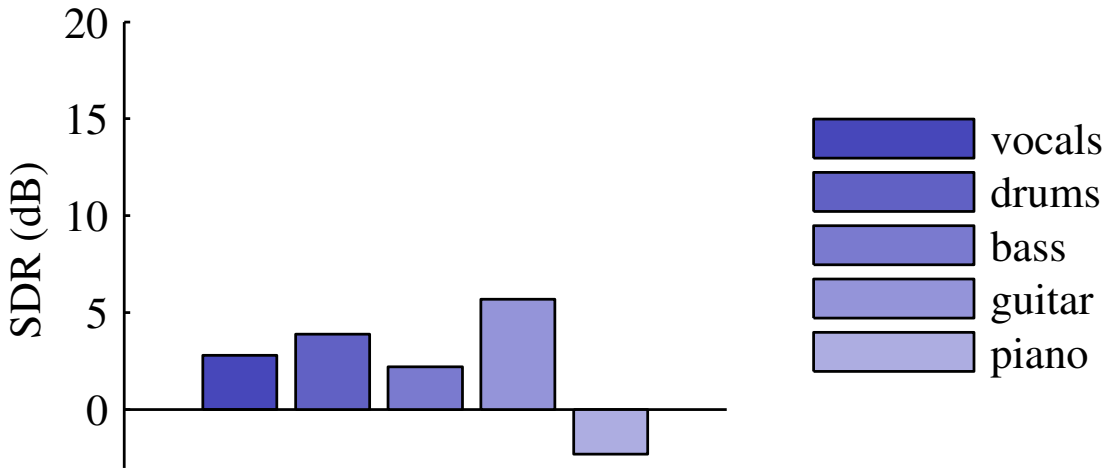
Source-filter factorization



Fine structure and envelope factorization



SiSEC results on professional music mixtures



Tamy (2 sources)

Estimated sources using the flexible framework

Bearlin (10 sources)

Estimated sources using the flexible framework

Results on a speech mixture

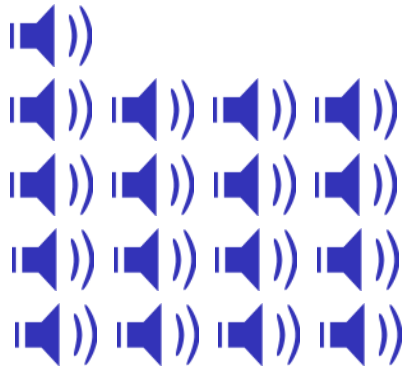
Recorded mixture of 4 sources

Estimated sources using rank-1 mixing covariance

full-rank mixing covariance

rank-1 and harmonicity

full-rank and harmonicity



Separation of single-channel recordings

The separation of single-channel recordings is more difficult than that of multichannel recordings since it relies on spectral cues only.

A specific model must be learned a priori for each source.

This makes it possible to separate the sources in each time frame (using pitch for instance).

For mixtures of 2 speakers

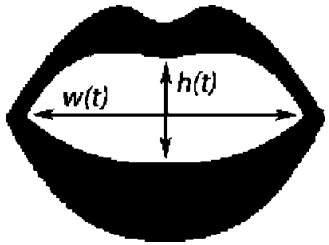
- Schmidt & Olsson obtained a SDR of 8 dB with 5 min training signals,
- Smaragdis obtained a SDR of 5 dB with 30 s training signals.

Grouping of the separated sources over time remains difficult and requires more sophisticated temporal evolution models which are currently being studied.

Exploitation of visual cues

Two approaches exist to exploit visual cues:

- activity detection of each speaker and zeroing of inactive time intervals,
- lip feature extraction and joint modeling of audio and visual features by GMMs.



The second approach performs better, but it cannot always be applied.

Most of these algorithms were tested on mixtures with $I \geq J$.

In a single-channel scenario, Llagostera obtained comparable performance to Smaragdis but with much shorter training signals.

Summary of probabilistic variance modeling

Advantages:

- virtually applicable to any mixture, including to diffuse sources
- no hard constraint on the number of sources per time-frequency bin
- the predominant sources are more accurately estimated by joint use of spatial, spectral and learned cues
- principled flexible framework for the integration of additional cues

Limitations:

- remaining musical noise artifacts
- remaining local optima of the estimation criterion

Summary

This state of the art showed that

- variance modeling algorithms have a greater potential due to the fusion of multiple cues,
- the separation quality is satisfactory for instantaneous noiseless mixtures: the handling of reverberation and noise remains a major challenge,
- single-channel separation remains difficult, especially when the sources have similar spectral cues,
- visual cues can improve performance but their use has been little studied.

Existing systems are gradually finding their way into the industry, especially for remixing applications that can accomodate a certain amount of musical noise artifacts and partial user input/feedback.

References

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, pp. 162-185, 2010.

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 Processing, 92, pp. 1928–1936, 2012.

H.K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array", IEEE Transactions on Audio, Speech and Language Processing, 15(8), pp. 2257-2268, 2007.

P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation", IEEE Transactions on Audio, Speech and Language Processing, 15(1), pp. 1-12, 2007.

A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations", IEEE Transactions on Multimedia, 12(5), pp. 358-371, 2010.

Websites and software

FASST: http://bass-db.gforge.inria.fr/fasst/
Software framework for the implementation of source separation algorithms (Matlab)

SiSEC: http://sisec.wiki.irisa.fr/
Series of evaluation campaigns