Audio rendering, coding and separation

Source localization and separation

N. Bertin

ISTIC / M2RI / P5 / CTR
Motivation

Given a multichannel mixture signal obtained by recording or rendering of several sources, an alternative rendering may be desired with e.g.:

- more or less channels (upmix/downmix)
- different source locations (respatialization)
- different source intensities (balance control)

Rendering methods take single-channel source signals as inputs. Source separation, *i.e.* estimation of the single-channel source signals underlying the mixture signal, must thus be performed prior to rendering.
First part: localization

1. Sound scenes
2. Auditory scene analysis
3. Two-channel localization using angular spectra
4. Two-channel localization using clustering
5. Multichannel localization
6. Summary
Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment...

Many applications are already available for, e.g., speech from a single speaker in a quiet environment.

But audio scenes are often more complex than we would like!

Ex: TV series
Sound scene analysis consists of analyzing a mixture of several sound sources in order to

1. describe the environment,
2. localize the sources,
3. describe them,
4. separate them.

Humans are able to perform the three first tasks above in many situations.

This has applications such as:

- hearing aids, denoising for handheld devices,
- post-production, remixing and 3D upmixing of music or movies,
- spoken/multimedia document retrieval, music information retrieval.
Example of sound scene

These tasks require the exploitation of the characteristics of the sound sources and the mixing process, which may be quite diverse.
Source characteristics

Audio sources include speech, music, and environmental sounds.

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:

- **temporal shape** (transitory, constant or variable)
- **spectral fine structure** (random or pitched)
- **spectral envelope**
Mixing characteristics

For point sources, room acoustics result in filtering of the recorded signal.

![Graph showing Amplitude vs. Time Delay with Early Reflections and Late Reverberations]

Software mixing has a similar effect.

In either case, the intensity and delay of direct sound are governed by the source position relative to the microphone.

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

2. Auditory scene analysis

3. Two-channel localization using angular spectra

4. Two-channel localization using clustering

5. Multichannel localization

6. Summary
Computational auditory scene analysis (CASA)

In order to design algorithms, it is useful to understand how the human auditory system works (but not to fully emulate it).

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
Spatial cues

Spatial proximity is assessed by comparing the observed

- interchannel time difference (ITD), also known as the time difference of arrival (TDOA),
- interchannel intensity difference (IID).

![ITD (anechoic)](image1)

![IID (anechoic)](image2)

![ITD (reverberant)](image3)

![IID (reverberant)](image4)
Spectral cues

The *Gestalt* rules also translate into spectral cues:

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

The estimation of *pitch* relies for instance on the cross-correlation of the auditory nerve representation in each band.
Learned cues and cues stemming from other modalities

In addition to the *primitive* cues above, the human auditory system exploits *learned* cues:

- episodic memory: "I know this poem"
- schematic memory: "The inaudible word after *presidential* must be *election*"
- short term memory: "I already heard this one minute ago"

Cues stemming from other modalities, in particular from *vision*, also play an essential role.
Source formation

It is thought that the brain associates each time-frequency bin with a single source according to these cues: this is known as binary masking.

After suppression

Piano mask

Estimated piano
Integration of multiple cues

Each cue alone is ambiguous:
- a given IID/ITD may be due to a single source in the corresponding direction or to multiple sources around that direction,
- several sources can have similar spectral characteristics.

In order to address these ambiguities, the available information must be integrated over
- several time-frequency bins
- (sometimes) several cues.

Two alternative integration approaches exist in the context of source localization:
- angular spectrum-based algorithms
- clustering-based algorithms.
1. Sound scenes
2. Auditory scene analysis
3. Two-channel localization using angular spectra
4. Two-channel localization using clustering
5. Multichannel localization
6. Summary
Time-frequency representation and steering vector

Most localization algorithms operate in the short time Fourier transform (STFT) domain.

In each time-frequency bin \((t, f)\),

\[
x(t, f) = \sum_{n=1}^{N} d(f, \tau_n) s_n(t, f) + b(t, f)
\]

where

\[
dx(f, \tau_n) = \left( \begin{array}{c} 1 \\ g_n e^{-2i\pi f \tau_n} \end{array} \right)
\]

is the steering vector associated with the ITD \(\tau_n\) and the IID \(g_n\) of the direct sound of source \(n\).
Link between DOA and TDOA

ITD and IID are governed by

- the azimuth (DOA) $\theta$, elevation $\delta$ and distance of the source,
- the spacing $d$ and directivity of the microphones,
- the presence of obstacles between the source and the microphones.

Most studies rely on a far field assumption, so that the ITD and IID do not depend on the distance to the source.

Most studies also assume omnidirectional microphones without obstacles. In this case,

$$\tau \approx \frac{d}{c} \cos \theta$$

$$g \approx 0$$
General principle of angular spectrum-based algorithms

The general principle of angular spectrum-based algorithms is to:

- build in each time-frequency bin a local angular spectrum function $\phi(t, f, \theta, \delta)$ that exhibits large values for the values of $\theta$ and $\delta$ which are compatible with the observed signal and small values otherwise,
- integrate this function over the time-frequency plane, leading to a global angular spectrum,
- find the peaks of this spectrum above a certain threshold and distant from a certain minimum angle.

These algorithms originate from the microphone array processing community.
Spatial aliasing and frequency integration

At high frequency, it is impossible to estimate the direction of sound in a given time-frequency bin alone.

Indeed, the observed interchannel phase difference $2\pi f \tau_n$ is compatible with several possible TDOAs $\tau_n + \frac{k}{f}$ with integer $k$.

When $f > c/d$, several values of $k$ are possible.

This spatial aliasing phenomenon requires the integration of the angular spectrum over all frequencies in a given time frame, typically by simple summation.
Temporal integration

All sources are generally not active on all time frames: some temporal integration is also necessary.

It is usually carried by simple summation

$$\phi^{\text{sum}}(\theta, \delta) = \sum_{t=1}^{T} \sum_{f=1}^{F} \phi(t, f, \theta, \delta)$$

or by taking the maximum

$$\phi^{\text{max}}(\theta, \delta) = \max_t \sum_{f=1}^{F} \phi(t, f, \theta, \delta)$$

Existing methods differ by the choice of the angular spectrum function $\phi(t, f, \theta, \delta)$ and the integration method.
GCC-PHAT and its variants (1)

The generalized cross-correlation (GCC) algorithm employs a sinusoidal function as the local angular spectrum, whose amplitude may depend on the signal.

Note that this function exploits ITD only.

The phase transform (PHAT) weighting consists of fixing the same amplitude for all time-frequency bins

$$\phi_{\text{GCC-PHAT}}(t, f, \theta, \delta) = \Re \left( \frac{x_1(t, f)x_2^*(t, f)}{|x_1(t, f)x_2^*(t, f)|} e^{-2\pi f \tau(\theta, \delta)} \right)$$

There exists nonlinear variants of GCC-PHAT based on applying a nonlinear function $\rho(u) = 1 - \tanh(\alpha \sqrt{1 - u}) \Rightarrow \phi^{GCC}(t, f, \theta, \delta)$ to enhance the peaks.
GCC-PHAT and its variants (2)

\[ \phi^{GCC} (t=0, f, \tau) \]

\[ \phi^{GCC_{sum}} (\tau) \times 10^4 \]

\[ N = 3, \ r = 50 \text{ cm}, \ d = 15 \text{ cm}, \ RT_{60} = 500 \text{ ms} \]
MUSIC (1)

The multiple signal classification (MUSIC) algorithm relies instead on analyzing the empirical covariance matrix of the signal

\[
\hat{R}_{xx}(t, f) = \frac{\sum_{t', f'} w(t' - t, f' - f)x(t', f')x(t', f')^H}{\sum_{t', f'} w(t' - t, f' - f)}
\]

The local angular spectrum is defined by

\[
\phi^{\text{MUSIC}}(t, f, \theta, \delta) = \left(1 - \frac{1}{2} \left| d(f, \theta, \delta)^H v(t, f) \right|^2 \right)^{-1}
\]

where \(v(t, f)\) is the principal eigenvector of \(\hat{R}_{xx}(t, f)\).
Two-channel localization using angular spectra

**MUSIC (2)**

\[
\phi_{\text{MUSIC}}(t=0, f, \tau)
\]

\[
\phi_{\text{MUSICsum}}(\tau)
\]

\[N = 3, \ r = 50 \text{ cm}, \ d = 15 \text{ cm}, \ RT_{60} = 500 \text{ ms}\]
Two-channel localization using angular spectra

SNR-based angular spectra (1)

GCC-PHAT and MUSIC give the same weight to all time-frequency bins, independently of the amount of interference, reverberation and noise.

A first idea is to define angular spectra whose amplitude is related to the signal-to-noise ratio (SNR).

For instance, in the framework of GCC, the optimal weighting for an additive noise uncorrelated with the sources is equal to

$$
\phi_{\text{GCC-ML}}(t, f, \theta, \delta) = \Re \left( \frac{x_1(t, f)x_2^*(t, f)}{|x_1(t, f)x_2^*(t, f)|} \frac{\gamma^\text{coh}(t, f)^2}{1 - \gamma^\text{coh}(t, f)^2} e^{-2i\pi f \tau(\theta, \delta)} \right)
$$

where

$$
\gamma^\text{coh}(t, f) = \frac{\text{SNR}}{1 + \text{SNR}} = \frac{|R_{x_1x_2}(t, f)|}{\sqrt{R_{x_1x_1}(t, f)R_{x_2x_2}(t, f)}}
$$

is the interchannel coherence.
SNR-based angular spectra (2)

Valin et al. use the following weighting instead:

\[
\phi^{\text{GCC-MCRA}}(t, f, \theta, \delta) = \Re \left( \frac{x_1(t, f)x_2^*(t, f)}{|x_1(t, f)x_2^*(t, f)|} \gamma_1^{\text{MCRA}}(t, f)\gamma_2^{\text{MCRA}}(t, f)e^{-2i\pi f \tau(\theta, \delta)} \right)
\]

where

\[
\gamma_i^{\text{MCRA}}(t, f) = \frac{\text{SNR}_i}{1 + \text{SNR}_i}
\]

is related to the SNR estimated at each microphone \(i\) by the minima controlled recursive averaging (MCRA) method for silence detection.
SNR-based angular spectra (3)

One may also define the angular spectrum to be the SNR itself.

By calculating the power of direct sound using the minimum variance distortionless response (MVDR) beamformer (see next part), we obtain

$$\phi_{\text{MVDR}}(t, f, \theta, \delta) = \frac{\left( d(f, \theta, \delta)^H \hat{R}_{xx}(t, f)^{-1} d(f, \theta, \delta) \right)^{-1}}{\frac{1}{2} \text{tr} \left( \hat{R}_{xx}(t, f) \right) - \left( d(f, \theta, \delta)^H \hat{R}_{xx}(t, f)^{-1} d(f, \theta, \delta) \right)^{-1}}$$

We will consider in the following a frequency-weighted version of this criterion called MVDRW.
SNR-based angular spectra (4)

$\phi_{\text{MVDRW}}(t=0,f, \tau)$

$N = 3, \ r = 50 \text{ cm}, \ d = 15 \text{ cm}, \ RT_{60} = 500 \text{ ms}$
Another idea is to separate the predominant source from the other sources in the neighborhood of each time-frequency bin by means of independent component analysis (ICA) (see next part).

ICA then returns two mixing coefficients \( a_1(t, f) \) and \( a_2(t, f) \) which are close in theory to \( e^{-2i\pi f \tau_1} \) and \( e^{-2i\pi f \tau_2} \), where \( \tau_1 \) and \( \tau_2 \) are the TDOAs of the two predominant sources in this neighborhood.

The state coherence transform (SCT) spectrum is defined by

\[
\phi^{c\text{SCT}}(t, f, \tau) = \sum_{n=1}^{2} \rho \left( \frac{1}{2} \left| e^{-2i\pi f \tau(\theta, \delta)} - a_n(t, f) \right| \right)
\]

where \( \rho(u) = 1 - \tanh(\alpha \sqrt{u}) \).

Note again that this exploits ITD only.
Evaluation (1)

We evaluated the above algorithms on 4446 signals sampled at 16 kHz:
- number of sources from $N = 2$ to 6,
- room reverberation time $RT_{60}$ equal to 50, 100, 150, 250, 500 or 750 ms,
- microphone spacing $d$ equal to 5, 15, 30 or 100 cm,
- distance to the sources $r$ equal to 20, 50, 100 or 200 cm,
- 3 to 5 random DOAs depending on the number of sources, between 30 and 150° and spaced by 10° minimum,
- 3 types of source signals (female, male, music) of 12 s duration.

The mixing filters were simulated by the source image method, for a room of size $4.45 \times 3.55 \times 2.5$ m.
Evaluation (2)

For all algorithms, the following parameter values were used for the computation of the STFT and the empirical covariance of the mixture:

- half-overlapping Hanning windows of 1024 samples (64 ms)
- neighborhood of $\pm 7$ frequency bands and $\pm 1$ time frame
- linear TDOA search grid, corresponding to a resolution of $0.6^\circ$ in the center and $1.3^\circ$ on the sides

Only peaks differing by at least $5^\circ$ are selected.

Localization performance is evaluated in terms of

- detection of the correct source locations, as measured by the F-measure for a tolerance of $5^\circ$ (optimal number of peaks known),
- precision of the correctly detected locations, as measured by the standard deviation for all methods with an F-measure larger than 0.7.
### Evaluation (3)

<table>
<thead>
<tr>
<th>Local angular spectrum function</th>
<th>F-measure (0 to 1)</th>
<th>Accuracy (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \phi_{\text{sum}} )</td>
<td>( \phi_{\text{max}} )</td>
</tr>
<tr>
<td>GCC-PHAT</td>
<td>0.65</td>
<td>0.84</td>
</tr>
<tr>
<td>GCC-NONLIN</td>
<td>0.69</td>
<td>0.85</td>
</tr>
<tr>
<td>MUSIC</td>
<td>0.43</td>
<td>0.41</td>
</tr>
<tr>
<td>MVDRW</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>cSCT</td>
<td>0.66</td>
<td>-</td>
</tr>
</tbody>
</table>
Evaluation (4)

![Graphs showing the average F-measure vs microphone spacing, reverberation time, and number of sources. The graphs compare different methods: GCC-NONLINsum, GCC-NONLINmax, cSCTsum, MVDRWsum, and MVDRWmax. Each method is represented by a different symbol or line pattern.](image)

**Average F-measure vs d**
- Microphone spacing $d$ (m)
- GCC-NONLINsum, GCC-NONLINmax, cSCTsum, MVDRWsum, MVDRWmax

**Average F-measure vs RT$_{60}$**
- Reverberation time RT$_{60}$ (ms)
- GCC-NONLINsum, GCC-NONLINmax, cSCTsum, MVDRWsum, MVDRWmax

**Average F-measure vs N**
- Number of sources $N$
- GCC-NONLINsum, GCC-NONLINmax, cSCTsum, MVDRWsum, MVDRWmax
1. Sound scenes
2. Auditory scene analysis
3. Two-channel localization using angular spectra
4. Two-channel localization using clustering
5. Multichannel localization
6. Summary
General principle of clustering-based algorithms

Angular spectrum-based algorithms integrate spatial cues over the whole time-frequency plane.

A peak corresponding to a source may be masked by gaps in the time-frequency bins where this source is inactive.

The general principle of clustering-based algorithms is to jointly estimate the source TDOAs and the active source(s) in each time-frequency bin.

These algorithms originate from the source separation community.

They generally rely on an iterative algorithm:
- estimation of the time-frequency activity map of the sources given their TDOAs,
- estimation of the TDOA of each source given its activity map.
Clustering according to the Euclidean distance

Sawada uses a standard **hard clustering** algorithm based on the Euclidean distance.

This algorithm is applied to the **phase and amplitude normalized** STFT coefficients of the mixture

\[
\tilde{x}(t, f) = \frac{\sqrt{2}}{\|x(t, f)\|} \frac{x_1^*(t, f)}{|x_1(t, f)|} x(t, f)
\]

which are in theory close to \(d(f, \tau_n)\) for the predominant source \(n\).

The algorithm consists of:

- associating bin \((t, f)\) with the source \(n\) minimizing \(\|\tilde{x}(t, f) - d(f, \tau_n)\|\)
- reestimating \(\tau_n\) from the bins \((t, f)\) associated with source \(n\) so as to minimize the sum of the these squared distances (criterion equivalent to GCC-PHAT summed on these bins)
Clustering using a binary activation model (1)

The Euclidean distance does not fit well the deviations of the apparent TDOA due to interference, reverberation and noise.

Izumi considers the following probabilistic model:

\[ x(t, f) = s_{ntf}(t, f)d(f, \tau_{ntf}) + b(t, f) \]

where \( s_{ntf}(t, f) \) is a deterministic and \( b(t, f) \) follows a Gaussian distribution with covariance \( \nu^b(t, f)\Psi(f) \) where

\[ \Psi(f) = \begin{pmatrix} 1 & \text{sinc}(2\pi f \frac{d}{c}) \\ \text{sinc}(2\pi f \frac{d}{c}) & 1 \end{pmatrix} \]

is the theoretical covariance of a diffuse noise.

Clustering is then performed in the maximum likelihood (ML) sense via an expectation-maximization (EM) algorithm

- E-step: estimate the posterior probability of \( n_{ntf} \).
- M-step: update \( s_{ntf}(t, f), \nu^b(t, f) \) and \( \tau_n \).
Clustering using a binary activation model (2)

A variant of this model (EM-predom) is to assume that $s_{n_tf}(t, f)$ is also Gaussian with variance $\nu^s(t, f)$.

Clustering is then performed in the ML sense via an EM algorithm:

- E-step: estimate the posterior probability of $n_{tf}$,
- M-step: update $\nu^s(t, f)$, $\nu^b(t, f)$ and $\tau_n$. 
Clustering using a multi-source model

Finally, one may seek to better take into account the presence of multiple sources in each time-frequency bin via the following model (EM-multi):

\[
x(t, f) = \sum_{n=1}^{N} s_n(t, f) d(f, \tau_n) + b(t, f)
\]

where \(s_n(t, f)\) and \(b(t, f)\) follow Gaussian distributions with variance \(\nu^s_n(t, f)\) and covariance \(\nu^b(t, f)\psi(f)\).

Clustering is then performed in the ML sense via an EM algorithm:
- E-step: estimate the posterior mean and covariance of \(s_n(t, f)\),
- M-step: update \(\nu^s_n(t, f)\), \(\nu^b(t, f)\) and \(\tau_n\).
Evaluation (1)

We evaluated these algorithms on mixtures of female speech among the mixtures considered above.

Due to the nonconvex nature of clustering problems, initialization is crucial.

The initial TDOAs are either

- randomly generated,
- estimated by GCC-NONLINmax (best angular spectrum-based algorithm).

For all algorithms, the following parameter values were used for the computation of the STFT and the empirical covariance of the mixture:

- half-overlapping Hanning windows of 1024 samples (64 ms)
- neighborhood of $\pm 1$ frequency band and $\pm 1$ time frame
- 100 iterations at most
## Evaluation (2)

<table>
<thead>
<tr>
<th>Clustering algorithm</th>
<th>F-measure (0 to 1)</th>
<th>Accuracy (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rand</td>
<td>init</td>
</tr>
<tr>
<td>None (GCC-NONLINmax)</td>
<td>-</td>
<td>0.90</td>
</tr>
<tr>
<td>Sawada et al.</td>
<td>0.49</td>
<td>0.87</td>
</tr>
<tr>
<td>Izumi et al.</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>EM-predom</td>
<td>0.52</td>
<td>0.85</td>
</tr>
<tr>
<td>EM-multi</td>
<td>0.32</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Evaluation (3)

Average F−measure vs $d$

- GCC–NONLINmax
- Sawada
- Izumi
- EM–predom

Average F−measure vs $RT_{60}$

Average F−measure vs $N$
1. Sound scenes
2. Auditory scene analysis
3. Two-channel localization using angular spectra
4. Two-channel localization using clustering
5. Multichannel localization
6. Summary
Multichannel localization

A pair of in-air omnidirectional microphones for 2D localization of a source in terms of azimuth $\theta$ and elevation $\delta$.

This requires

- either 2 microphones around an obstacle which is asymmetric with respect to all the plans intersecting the microphones and of sufficient size with respect to the wavelength (head),
- $l > 3$ unaligned microphones (array).

In the first case, $\theta$ is found via the ITD and $\delta$ via the IID resulting from the shape of the head.
Approaches for multichannel localization

There exist three approaches for an array with any geometry:

1. **triangulation** of the TDOAs estimated from all pairs of microphones,
2. **summation** of the angular spectra or the clustering criteria over all pairs of microphones,
3. **generalization** of the angular spectra and the clustering criteria to multichannel input, as parameterized by $\hat{R}_{xx}(t, f)$ and $d(f, \tau)$.

DiBiase has shown that approach 1 is less accurate than approach 2.
SRP-PHAT

Following approach 2, GCC-PHAT becomes

$$\phi^{\text{SRP-PHAT}}(t, f, \theta, \delta) = \sum_i \sum_{i'} \phi^{\text{GCC-PHAT}}_{ii'}(t, f, \theta, \delta)$$

where $$\phi^{\text{GCC-PHAT}}_{ii'}(t, f, \theta, \delta)$$ is the GCC-PHAT spectrum between microphones $$i$$ and $$i'$$.

This is called the steered response power PHAT (SRP-PHAT) spectrum.

Approach 3 boils down to considering only certain pairs of microphones with respect to a reference microphone $$i$$ (Loesch)

$$\phi^{\text{GCC-PHAT ref}}(t, f, \theta, \delta) = \sum_{i'} \phi^{\text{GCC-PHAT}}_{ii'}(t, f, \theta, \delta)$$
Evaluation

Few evaluation results are available for multi-microphone scenarios to date.

Valin evaluated SRP-PHAT using a cubic array of 8 microphones with 16 cm side.

He obtained a recall of 98 to 100% and a standard deviation of 1° for a single source, after integrating on a certain temporal duration.
Exploitation of visual cues

Maganti proposed a particle filtering approach for audiovisual speaker tracking using calibrated cameras and microphones:

- audio cues extracted by SRP-PHAT,
- visual shape and color cues,
- probabilistic combination of the 2 cues, using visual cues alone when a speaker is inactive
- probabilistic modeling of the movements and the activity/inactivity of the speakers

According to him, the combination of both cues reduces false positives/negatives and visual cues alone provide a smaller standard deviation among the correctly detected sources.
1. Sound scenes
2. Auditory scene analysis
3. Two-channel localization using angular spectra
4. Two-channel localization using clustering
5. Multichannel localization
6. Summary
Summary

This state of the art showed that
- angular spectrum-based algorithms provide state-of-the-art detection performance and accuracy and they can be used in real time,
- temporal integration over several 100 ms is necessary, because of source inactivity and echoes,
- visual cues help improving detection performance and accuracy.

Clustering-based algorithms currently suffer from local optima issues, but this may change soon thanks to convex formulations of the problem.
References


B. Loesch, and B. Yang, ”Blind source separation based on time-frequency sparseness in the presence of spatial aliasing,” in Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA), pp. 1-8, 2010.


Software

**HARK**: http://winnie.kuis.kyoto-u.ac.jp/HARK/
Software platform for robot audition, including a localization module based on MUSIC

**ManyEars**: http://sourceforge.net/projects/manyyears/
Software platform for robot audition, including a localization module based on SRP-MCRA

**BSS Locate**: http://bass-db.gforge.inria.fr/bss_locate/
Two-channel localization toolbox (Matlab)

**Roomsimove**: http://www.loria.fr/~evincent/Roomsimove.zip
Simulation of room impulse responses (Matlab)
Second part: separation

1. Beamforming and post-filtering
2. Probabilistic linear modeling
3. Probabilistic variance modeling
4. Summary
Source separation

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

Perfect source recovery is rarely feasible. Instead, the estimated sources are often distorted in various ways.

- Mixture
- Perfect source recovery
- Interference
- Artifacts
- Timbre distortion

The global distortion level can be measured via the Signal-to-Distortion Ratio (SDR) in decibels (dB).
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).
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

\[
X_{nf} = S_{nf}D_f + B_{nf}
\]

where the steering vector \(D_f\) encode the ITDs \(\tau_i\) and the IIDs \(g_i\) between the \(I\) microphones

\[
D_f \propto \begin{pmatrix}
1 \\
g_2 e^{-2i\pi f\tau_2} \\
\vdots \\
g_I e^{-2i\pi f\tau_I}
\end{pmatrix}
\]
Beamforming and post-filtering

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

$$\hat{S}_{nf} = V_{nf}^S D_f^H (\Sigma_{nf}^X)^{-1} X_{nf}$$

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

This estimator is in fact the combination of

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

$$Y_{nf} = \frac{D_f^H (\Sigma_{nf}^X)^{-1} X_{nf}}{D_f^H (\Sigma_{nf}^X)^{-1} D_f}$$

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

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

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 $\Sigma_{nf}^X$ 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.
1 Beamforming and post-filtering
2 Probabilistic linear modeling
3 Probabilistic variance modeling
4 Summary
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 \((n, f)\)

\[
X_{nf} = \sum_{j=1}^{J} S_{jnf} A_{jf}
\]

- \(X_{nf}\): vector of mixture STFT coeff.
- \(J\): number of sources
- \(S_{jnf}\): \(j\)th source STFT coeff.
- \(A_{jf}\): \(j\)th mixing vector
Modeling of the mixing vectors

The mixing vectors $A_{jf}$ encode the ITD and IID of each source at each frequency.

For anechoic mixtures, $A_{jf}$ is equal to the steering vector $D_{jf}$.

For echoic mixtures, ITDs and IIIDs follow a smeared distribution $P(A_{jf} | \theta_j)$.
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 $j_{nf}^{\text{act}}$ in each time-frequency bin with, e.g., uniform $P(j_{nf}^{\text{act}})$,
- **generalized exponential distribution**

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

$p$: shape parameter

$\beta_f$: scale parameter

Distribution of magnitude STFT coeff.

![Distribution of magnitude STFT coeff.](image)
Inference algorithms

Given the above priors, source separation is typically achieved by joint MAP estimation of the source STFT coefficients $S_{jn}$ and other latent variables $(A_{jf}, g_j, \tau_j, p, \beta_j)$ 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 room, i.i.d. priors
- Office room, i.i.d. priors

### SDR (dB)

<table>
<thead>
<tr>
<th>Angular Distance</th>
<th>Anechoic</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>80°</td>
<td><img src="image1" alt="Anechoic" /></td>
<td><img src="image2" alt="Office" /></td>
</tr>
<tr>
<td>40°</td>
<td><img src="image3" alt="Anechoic" /></td>
<td><img src="image4" alt="Office" /></td>
</tr>
<tr>
<td>20°</td>
<td><img src="image5" alt="Anechoic" /></td>
<td><img src="image6" alt="Office" /></td>
</tr>
</tbody>
</table>

- 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
1. Beamforming and post-filtering
2. Probabilistic linear modeling
3. Probabilistic variance modeling
4. Summary
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 \((n, f)\)

\[
X_{nf} = \sum_{j=1}^{J} C_{jnf}
\]

- \(X_{nf}\): vector of mixture STFT coeff.
- \(J\): number of sources
- \(C_{jnf}\): \(j\)th 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.

\[ P(C_{jn|} \mid \Sigma_{jn|}) = \frac{1}{\det(\pi \Sigma_{jn|})} e^{-C_{jn}^H \Sigma_{jn|}^{-1} C_{jn}} \]

\( \Sigma_{jn|} \): \( j \)th source covariance matrix

The covariance matrix \( \Sigma_{jn} \) of each source can be factored as the product of a scalar nonnegative variance \( V_{jn} \) and a spatial covariance matrix \( R_{jf} \) respectively modeling spectral and spatial properties.

\[ \Sigma_{jn} = V_{jn} R_{jf} \]

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

\[ P(X_{nf} \mid V_{jn}, R_{jf}) = \frac{1}{\det \left( \pi \sum_{j=1}^{J} V_{jn} R_{jf} \right)} e^{-X_{nf}^H \left( \sum_{j=1}^{J} V_{jn} R_{jf} \right)^{-1} X_{nf}} \]
General inference algorithm

Independently of the priors over $V_{jnf}$ and $R_{jf}$, 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

$$
\hat{C}_{jnf} = V_{jnf} R_{jf} \left( \sum_{j'=1}^{J} V_{j'nf} R_{j'f} \right)^{-1} X_{nf}.
$$
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. $R_{jf}$ has full rank.

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

$$R_{jf} \propto \lambda_j A_{jf} A_{jf}^H + B_f$$

with

$$A_{jf} = \sqrt{\frac{2}{1 + g_j^2}} \left( 1 \right) \left( g_j e^{-2i\pi f \tau_j} \right)$$

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

$\lambda_j$: direct-to-reverberant ratio  
$A_{jf}$: direct mixing vector  
$B_f$: diffuse noise covariance  
$\tau_j$: ITD of direct sound  
g_j: IID of direct sound  
d: microphone spacing  
c: sound speed  
Modeling of $R_{jf}$ as an unconstrained full-rank matrix is also possible.
I.i.d. modeling of the source variances

Baseline systems rely on modeling the source variances $V_{jn}$ 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.
Practical illustration of separation using i.i.d. variance priors

Left source $S_{1nf}$ (IID < 0)

Center source $S_{2nf}$ (IID = 0)

Right source $S_{3nf}$ (IID > 0)

Mixture $X_{nf}$

Predominant source pairs

Estimated nonzero source pairs

First estimated source $\hat{S}_{1nf}$

Second estimated source $\hat{S}_{2nf}$

Third estimated source $\hat{S}_{3nf}$

E. Vincent, N. Bertin
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.
Probabilistic variance modeling

Practical illustration of separation using template spectra

- **Piano source** $C_{1nf}$
- **Violin source** $C_{2nf}$
- **Mixture** $X_{nf}$

**Template spectra** $w_{jkf}$

**Estimated state sequences** $q_{jn}$

**Estimated piano variance** $\Sigma_{1nf}$

**Estimated violin variance** $\Sigma_{2nf}$

**Estimated mixture variance** $\Sigma_{1nf} + \Sigma_{2nf}$

**Estimated piano source** $\hat{C}_{1nf}$

**Estimated violin source** $\hat{C}_{2nf}$
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 $h_{jkn}$

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

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.
Probabilistic variance modeling

Practical illustration of separation using basis spectra

Piano source $C_{1nf}$

Violin source $C_{2nf}$

Mixture $X_{nf}$

Basis spectra $w_{jkf}$

Estimated scale factors $h_{jkn}$

Estimated piano variance $\Sigma_{1nf}$

Estimated violin variance $\Sigma_{2nf}$

Estimated mixture variance $\Sigma_{1nf} + \Sigma_{2nf}$

Estimated piano source $\hat{C}_{1nf}$

Estimated violin source $\hat{C}_{2nf}$

E. Vincent, N. Bertin

Source localization and separation

06/11/2013 88 / 104
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

Unconstrained basis spectra $w_{jp}$

Constrained basis spectra $w_{jp}$
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
  \[ V_{jn} = V_{jn}^{ex} V_{jn}^{ft} \]

- further factorization of each part into basis spectra and time activation coefficients e.g.
  \[ V_{jn}^{ex} = \sum_k h_{jn}^{ex} w_{jn}^{ex} \]

- further factorization of the basis spectra and time activation series into fine structure and envelope coefficients e.g.
  \[ w_{jn}^{ex} = \sum_l e_{jn}^{ex} f_{jn}^{ex} \]
Source-filter factorization

\[ (A) \text{Source spectral power} \quad \approx \quad (B) \text{Model spectral power} V_j = V_{j1} \circ V_{j2} \]

\[ (C) \text{Evolution spectral power} V_j^{(n)} \quad \quad \quad \quad (D) \text{Filter spectral power} V_j^{B} \]
Fine structure and envelope factorization

(A) Source spectral power

(B) Excitation spectral power $\Psi_{F}^{E} = \phi_{E}^{E}$

(C) Characteristic spectral patterns $E_{k}^{E} = W_{k}^{E} U_{k}^{E}$

(D) Spectral patterns activations $P_{k}^{E} = C_{k}^{E} H_{k}^{E}$

(E) Narrowband spectral patterns $W_{k}^{E}$

(F) Spectral patterns weights $U_{k}^{E}$

(G) Temporal patterns weights $G_{k}^{E}$

(H) Time localized patterns $H_{k}^{E}$
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
1. Beamforming and post-filtering
2. Probabilistic linear modeling
3. Probabilistic variance modeling
4. Summary
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 accommodate a certain amount of musical noise artifacts and partial user input/feedback.
References


Websites and software

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

Series of evaluation campaigns