Blind Audio Source Separation

A review of state-of-the-art techniques

Emmanuel Vincent

Centre for Digital Music
Electronic Engineering Department
Queen Mary, University of London

The problem and its applications

Most audio signals are mixtures of several audio sources (speech, music, noises). Blind Audio Source Separation (BASS) consists in recovering one or several source signals from a given mixture signal.

Direct applications include

- real-time speaker separation for simultaneous translation,
- sampling of musical sounds for electronic music composition.

Many derived applications aim to modify the mixture signal, for example

- speech enhancement within hearing aids,
- voice cancellation for karaoke,
- rendering of stereo CDs on multichannel devices.

Scope of the review

BASS methods have been proposed independently by researchers from different communities (statistical signal processing, audio signal processing, cognitive psychology) since the early 90's.

The goal of this review is to

- make clear the modeling assumptions and performance limitations of each approach,
- emphasize the importance of audio-specific issues in the design of BASS algorithms.

Implementation details are not discussed.

Overview

1. Background and problem definition

2. Filtering techniques

3. Multi-channel identification based on sparsity

4. Single-channel identification based on advanced models

5. Hybrid identification techniques

6. Conclusion and promising approaches

Typical audio mixtures

Audio mixtures are categorized as

- music or speech,
- live recordings or synthetic mixtures.

Each category exhibits different properties that can be exploited for a specific processing.

Music *vs.* speech sources

Both music and speech sources produce a superposition of

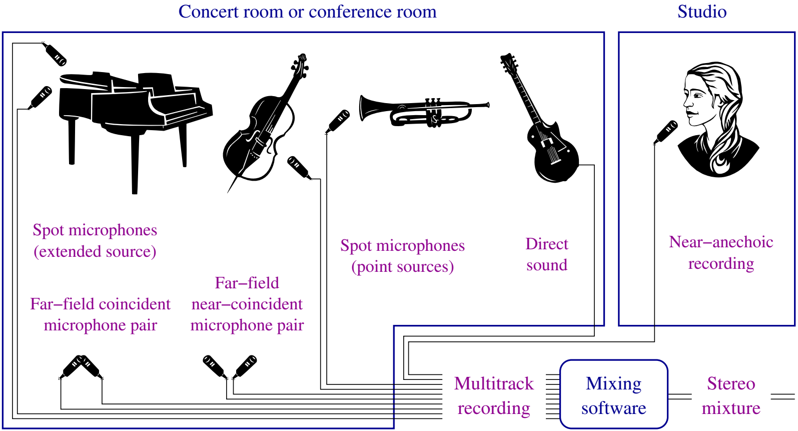
- nearly periodic signals (harmonic sinusoidal partials),
- wideband noise signals,
- transient signals.

Differences lie in

- the underlying discrete structure (chords *vs.* phonemes),
- the amount of dependence between different sources,
- the fundamental frequency and spectral envelope range.

Live recordings

Live sources can be recorded separately (pop music, movie dialogues) or together (classical music, meeting). The setup and the microphones types determine the amount of interferences and reverberation on each channel.


Concert room or conference room
Studio
Spot microphones (extended source)
Spot microphones (point sources)
Direct sound
Near–anechoic recording
Far–field coincident microphone pair
Far–field near–coincident microphone pair
Multitrack recording
Mixing software
Stereo mixture

Synthetic mixtures

If necessary, live recordings can be mixed down to a stereo (two-channel) mixture.

Typical synthetic mixing effects include

- panning,
- synthetic reverb,
- autopanning.

Classification of mixtures types

The signal processing literature classifies mixtures as

- under-determined *vs.* over-determined,
- instantaneous *vs.* convolutive,
- time-varying *vs.* time-invariant.

Live recordings are over-determined time-varying convolutive mixtures (many microphones, reverberation, moving sources).

Synthetic mixtures are most often under-determined time-invariant convolutive mixtures (two channels, synthetic reverb).

The separation problem

Under the point source assumption, the mixture channels (x_i) equal

$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=-\infty}^{+\infty} a_{ij}(t-\tau,\tau) s_j(t-\tau),$$

where (s_j) are the original source signals and (a_{ij}) the mixing filters.

More generally, the mixing process writes

$$x_i(t) = \sum_{j=1}^{J} s_{\mathrm{img}\,ij}(t),$$

where $(s_{\mathrm{img}\,ij})$ are the source images on the mixture channels.

The BASS problem consists in retrieving the source image signals $(s_{\mathrm{img}\,ij})$ given the mixture signals (x_i).

Performance evaluation

Various distortions are distinguished when evaluating the separation quality.

Mixture 🔊

Perfect separation 🔊

Pipe-noise artifacts 🔊 Interferences 🔊 Timbre distortion 🔊

These distortions should be measured by listening tests.
They can also be approximated by energy ratio criteria.

BSS_EVAL toolbox: `http://www.irisa.fr/metiss/bss_eval/`.

Identification and filtering

BASS is often addressed as a two-part problem:

- identify the number of sources and a set of sufficient parameters for each source,
- filter the mixture based on these parameters to obtain the source image signals.

A good separation may be obtained under two conditions:

- the filtering technique is actually able to separate the sources using optimal filters,
- the sufficient parameters contain enough information to derive near-optimal filters.

Filtering techniques

Common filtering techniques include

- beamforming,
- time-frequency masking.

Beamforming (1)

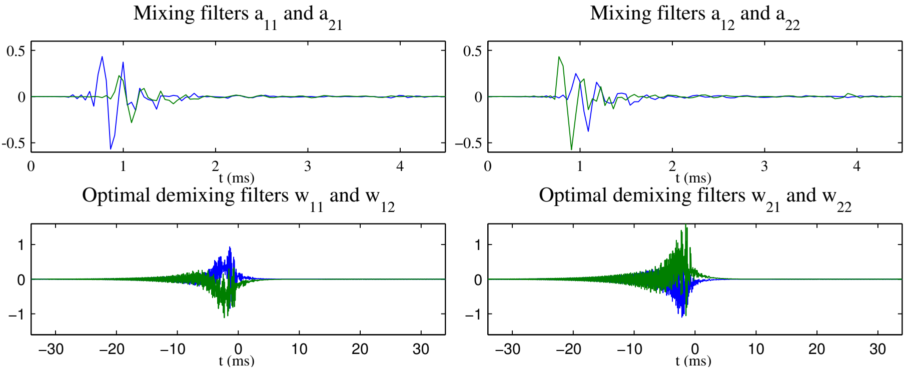
Beamforming relies on spatial diversity and exploits all the mixture channels simultaneously.

It involves filtering the mixture channels by stationary filters and summing them together. In the time-frequency domain, this translates as

$$\hat{U}_j(n,f) = \sum_{k=1}^{I} w_{jk}(f) X_k(n,f),$$
$$\hat{S}_{\mathrm{img}\,ij}(n,f) = b_{ij}(f) \hat{U}_j(n,f),$$

where $(\hat{S}_{\mathrm{img}\,ij})$ and (X_i) denote the short time Fourier transforms of the source images and the mixture channels, $(w_{jk}(f))$ are the demixing filters and $(b_{ij}(f))$ is the pseudo-inverse matrix of $(w_{jk}(f))$.

Beamforming (2)



Time-frequency masking (1)

Time-frequency masking relies on time-frequency diversity and processes each mixture channel separately.

It involves computing the STFTs of the mixture channels, multiplying them by time-frequency masks containing gains between 0 and 1 and inverting the resulting STFTs.

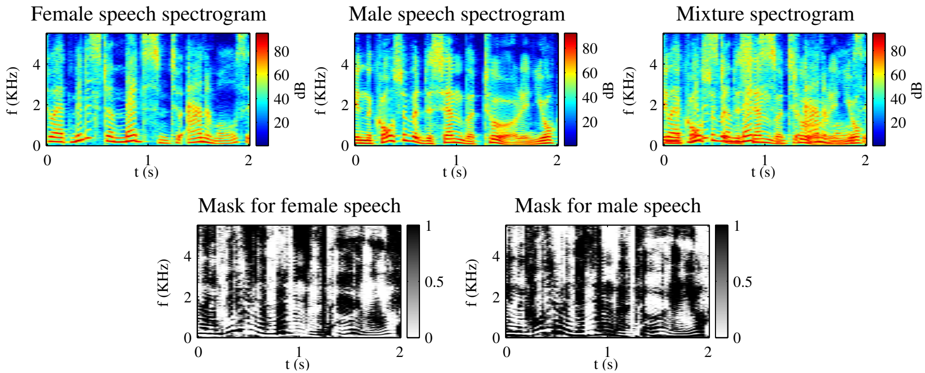
The most popular masking rules are adaptive Wiener filtering

$$\hat{S}_{\mathrm{img}\,ij}(n,f) = \frac{|S_{\mathrm{img}\,ij}(n,f)|^2}{\sum_{k=1}^{J} |S_{\mathrm{img}\,ik}(n,f)|^2} X_i(n,f)$$

and binary masking

$$\hat{S}_{\mathrm{img}\,ij}(n,f) = \begin{cases} X_i(n,f) & \text{if } |S_{\mathrm{img}\,ij}(n,f)| = \max_k |S_{\mathrm{img}\,ik}(n,f)|, \\ 0 & \text{otherwise.} \end{cases}$$

Time-frequency masking (2)



Female speech spectrogram

Male speech spectrogram

Mixture spectrogram

Mask for female speech

Mask for male speech

Comparison of beamforming *vs.* time-frequency masking

The computation of optimal demixing filters/time-frequency masks shows that

- beamforming does not generate pipe noise artifacts but is limited to determined instantaneous or slightly convolutive mixtures,
- time-frequency masking works potentially better for under-determined or strongly convolutive mixtures but generates artifacts due to time-frequency overlap of the sources.

BSS_ORACLE toolbox: `http://www.irisa.fr/metiss/bss_oracle/`.

Multi-channel identification based on sparsity

Sparsity forms the core of separation methods for simple multichannel mixtures.

Time-frequency sparsity allows to separate sources in each subband based on spatial diversity. Complementary assumptions are needed to link together subband signals belonging to the same source.

Common methods include

- ICA,
- DUET-like methods.

Independent Component Analysis (1)

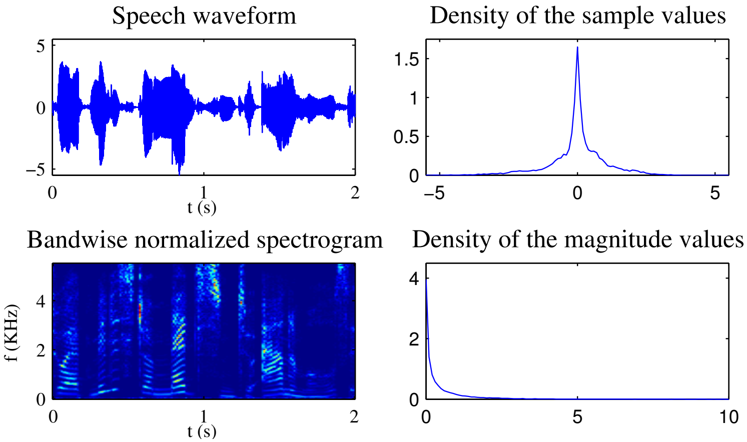
ICA for over-determined instantaneous mixtures models the source samples $(s_j(t))$ as independent draws from a known distribution.

Audio signals are often modelled by sparse distributions such as the generalized exponential

$$P(s_j(t)) \propto e^{-\beta|s_j(t)|^R} \quad \text{with } 0 \leq R < 2.$$

The sources are estimated up to a permutation using beamforming. The demixing gains (w_{jk}) are estimated using a maximum *a posteriori* criterion or equivalently a minimum mutual information criterion.

Independent Component Analysis (2)



Independent Component Analysis (3)

ICA is extended to over-determined convolutive mixtures by modeling the STFT values $(S_j(n,f))$ as independent draws from a sparse distribution and applying standard ICA in each subband f to recover $(\hat{S}_{\mathrm{img}\,ij}(n,f))$.

Subbands are grouped after solving the permutation ambiguity using

- source directions defined from interchannel phase difference by $\mathrm{IPD}_j(n,f) = \angle\hat{S}_{\mathrm{img}\,2j}(n,f) - \angle\hat{S}_{\mathrm{img}\,1j}(n,f) \approx 2\pi f \frac{d}{c} \sin\hat{\theta}_j \mod 2\pi$,
- correlation of the source subband magnitudes.

Reverb	Mixture	ICA	
6 ms	🔊	🔊	🔊
20 ms	🔊	🔊	🔊
90 ms	🔊	🔊	🔊
370 ms	🔊	🔊	🔊

Davies 02, Parra 02, Sawada 03

DUET-like methods (1)

In instantaneous mixtures, if source j is predominant in (n, f) then the interchannel intensity difference $\mathrm{IID}(n, f) = 20 \log_{10}(|X_2(n, f)|/|X_1(n, f)|)$ is close to the relative mixing gain $20 \log_{10}(a_{2j}/a_{1j})$.

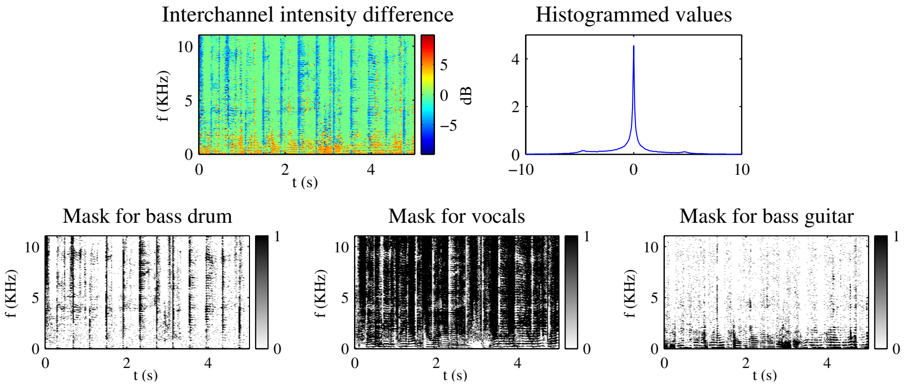
DUET-like methods separate under-determined mixtures as follows:

- compute IID (or similar quantities),
- find the relative mixing gains from the IID histogram,
- associate each time-frequency point with the closest source and perform binary masking.

The interchannel phase difference is used for convolutive mixtures instead.

Jourjine 00, Viste 03, Roman 03

DUET-like methods (2)



Mixture 🔊 Spatial masking 🔊 🔊 🔊

Single-channel identification based on advanced models

The time-frequency sparsity assumption is insufficient for the separation of single-channel mixtures.

Advanced signal models represent the short term magnitude spectrum of the mixture and take into account the discrete structure of the sources along with periodicity, spectral envelope and temporal continuity characteristics.

Common methods include

- factorial HMMs,
- spectral decomposition,
- CASA.

Factorial hidden Markov models (1)

Each source is modelled by a HMM

$$\log |S_j(n,f)|^2 = \Psi_{h_j(n)}(f) + \epsilon_j(n,f),$$

where $(h_j(n))$ is a series of hidden states following a Markov prior, $\Psi_{h_j(n)}$ is the expected spectrum and ϵ_j is a Gaussian noise.

The mixture is represented by a factorial HMM whose states are $(h_1(n), \ldots, h_J(n))$.

Under particular assumptions,

$$\log |X(n,f)|^2 \simeq \max_{j=1\ldots J} \Psi_{h_j(n)}(f) + \epsilon(n,f).$$

Factorial hidden Markov models (2)

Single-channel speech separation with factorial HMMs comprises 3 steps:

- train one HMM per source on solo excerpts beforehand,
- decode the MAP mixture states,
- set $\log|\hat{S}_j(n,f)|^2 = \Psi_{\hat{h}_j(n)}(f)$ and perform time-frequency masking.

Mixture 🔊 Vocals 🔊 Music 🔊 (from A. Ozerov)

Roweis 00, Ozerov 05, Reyes-Gomez 03

Spectral decomposition models (1)

The state space for music sources is large. This raises overlearning issues.

Music sources are better represented as a superposition of several components

$$|X(n,f)|^2 = \sum_{j=1}^{J} \sum_{k=1}^{K_j} e_{jk}(n) \Phi_{jk}(f) + \epsilon(n,f),$$

where (Φ_{jk}) and $(e_{jk}(n))$ are the basis spectra and time-varying weights for source j and ϵ is the total residual.

Spectral decomposition models (2)

Spectral decomposition is often used in a data-driven fashion. Spectra and weights are derived from the mixture STFT using

- sparsity,
- non-negativity,
- various residual models.

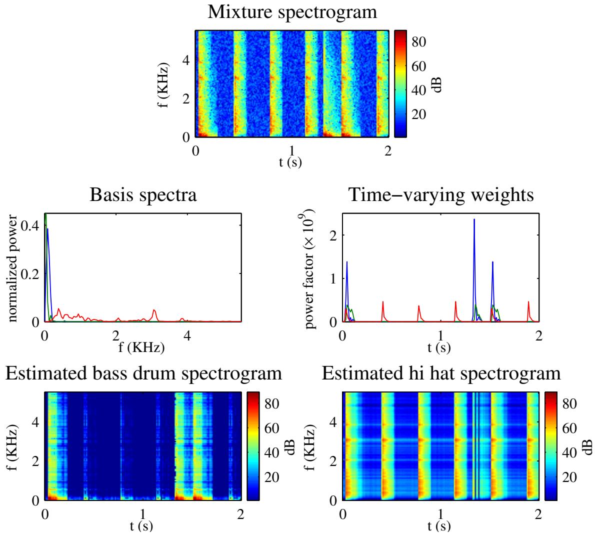
Then the components are clustered into sources exploiting

- residual dependencies between weight series,
- instrument-specific features.

Finally source STFTs are recovered as $|S_j(n,f)|^2 = \sum_{k=1}^{K_j} e_{jk}(n)\Phi_{jk}(f)$ and time-frequency masking is performed.

Casey 00, Virtanen 03, Smaragdis 03, Abdallah 04, Benaroya 03

Spectral decomposition models (3)



MIDI percussion example

Mixture 🔊

NMF 🔊 🔊

Music example with 32 components and manual clustering (from B. Wang)

Mixture 🔊

NMF 🔊 🔊

Computational auditory scene analysis

CASA groups time-frequency zones into objects and streams these objects into sources using experimental rules:

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

Blackboard CASA methods resemble spectral decomposition except that

- components are constrained (“wefts”, “transients”, “noise clouds”),
- advanced timbre features (onset duration, *vibrato*) are exploited.

Ellis 96, Kashino 95, Kinoshita 99

Hybrid identification techniques

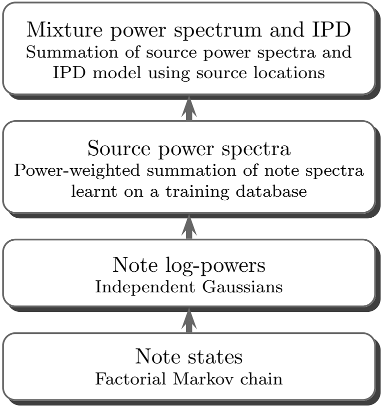
The previous methods suffer some limitations:

- algorithms exploiting spatial diversity only cannot deal with reverberation or sources with close directions,
- algorithms based on fundamental frequency and spectral envelope dissimilarity cannot separate instruments from the same class.

Performance can be improved using advanced models to represent spatial and spectro-temporal cues jointly.

An example of hybrid model

CASA algorithms deal with stereo by adding a spatial proximity rule.
A recent approach combines spectral decomposition, factorial HMMs and IPD models into a single probabilistic generative model for music.



Mixture power spectrum and IPD
Summation of source power spectra and IPD model using source locations

Source power spectra
Power-weighted summation of note spectra learnt on a training database

Note log-powers
Independent Gaussians

Note states
Factorial Markov chain

Mixture	🔊
ICA	🔊 🔊
Spatial TF masking	🔊 🔊
Hybrid method	🔊 🔊

Nakatani 02, Sakuraba 03, Vincent 05

Conclusion

The BASS problem remains far from being solved.

Realistic audio mixtures are better separated by DUET-like methods or hybrid models. However the quality remains insufficient for demanding applications (hearing aids, karaoke).

Several issues account for this:

- reverberation,
- time-frequency overlap of the sources,
- possibly source movements, similarities between source characteristics or proximity of source directions.

Promising approaches

Promising approaches include

- special recording techniques reducing reverberation and interferences,
- hybrid filtering techniques coupled with advanced source models,
- advanced models of the source waveforms.

Sanchis 04, Katayama 04, Viste 02, Virtanen 03, Every 05, Blumensath 04

The Blind Audio Source Separation evaluation database

The database can be used to create realistic mixtures from

- multitrack music recordings (acoustic pop, metal, electro...),
- realistic mixing filters (Ardour plugins, HRTFs, room impulse responses, "shoebox" room acoustics toolbox).

Insert your algorithm and its results!

BASS-dB: `http://www.irisa.fr/metiss/BASS-dB/`.