IRISA/D5 Thematic Seminar

Introduction to sound scene analysis
Source localization

Emmanuel Vincent

Inria Nancy – Grand Est

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

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.

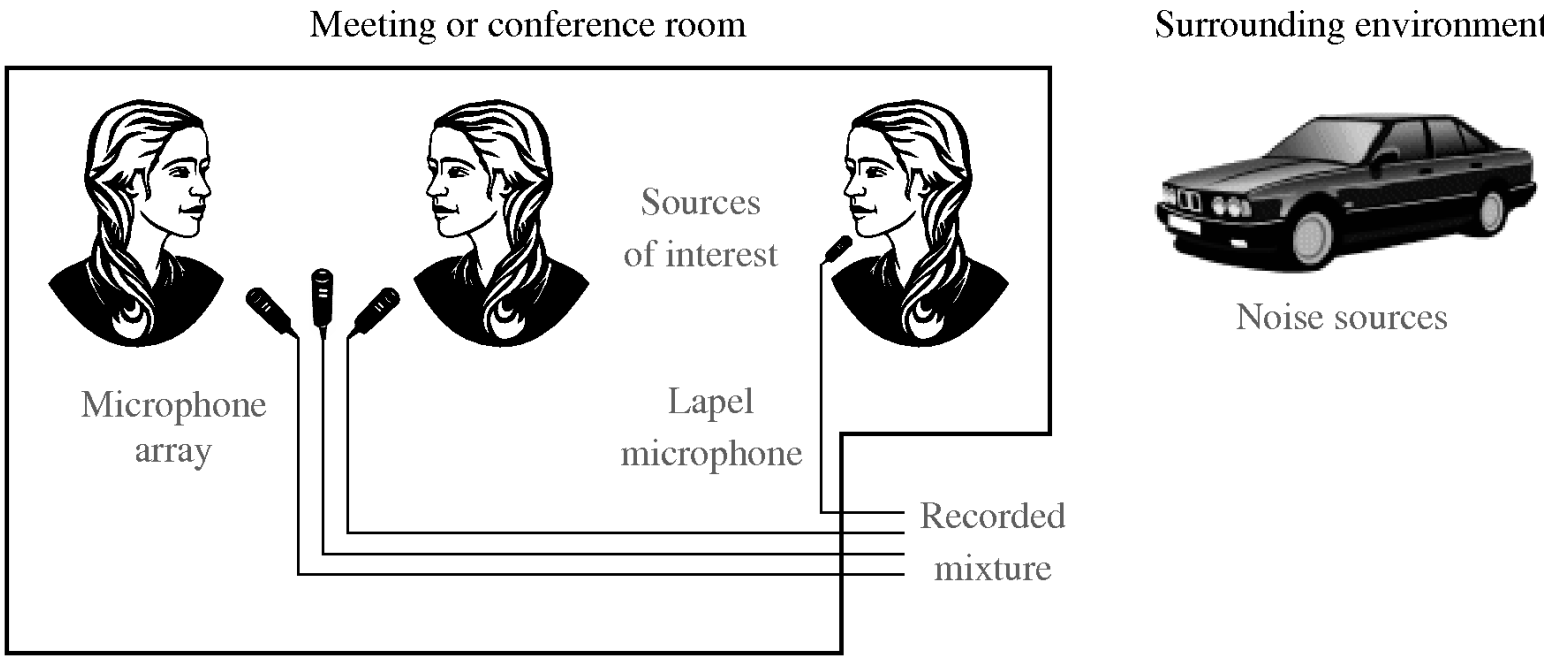
Humans are able to perform the three first tasks above in many situations.

This has applications such as:

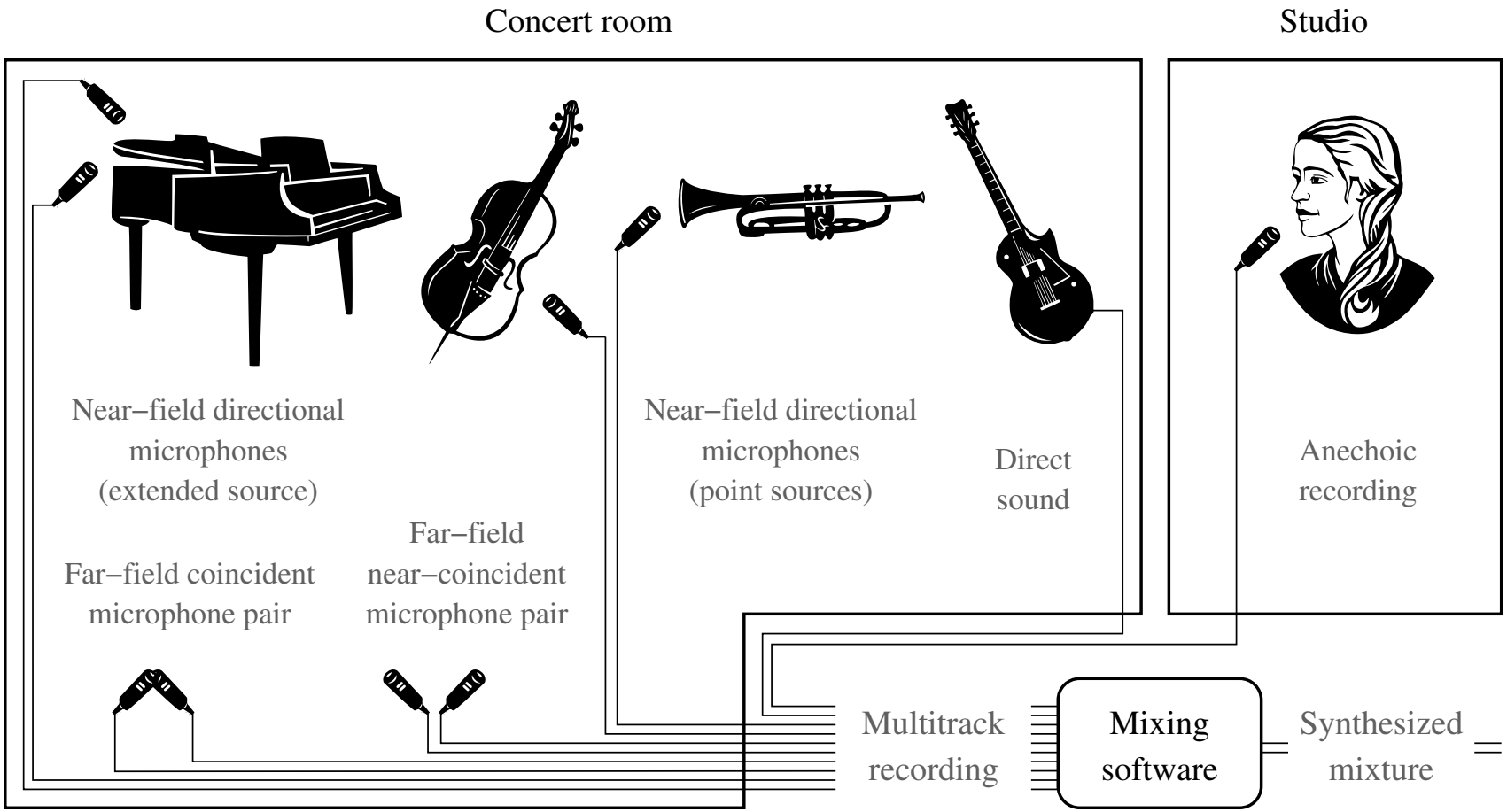
- hearing aids, denoising for handheld devices,
- post-production, remixing and 3D upmixing of music or movies,
- spoken/multimedia document retrieval, music information retrieval.

Example recorded sound scene

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



Example synthetic sound scene



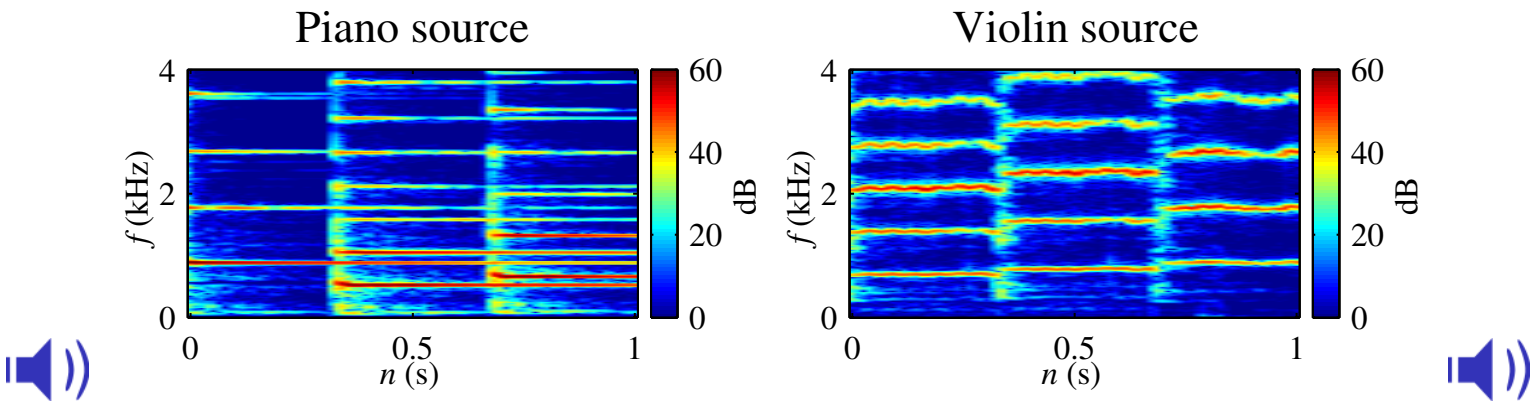
Source characteristics

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.

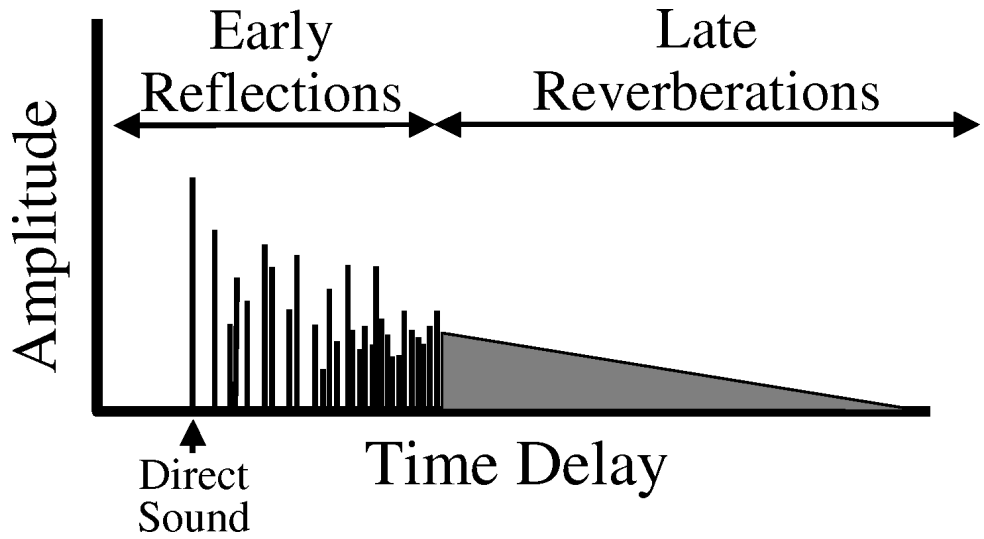
This results in a wide variety of sounds characterized by their:

- 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



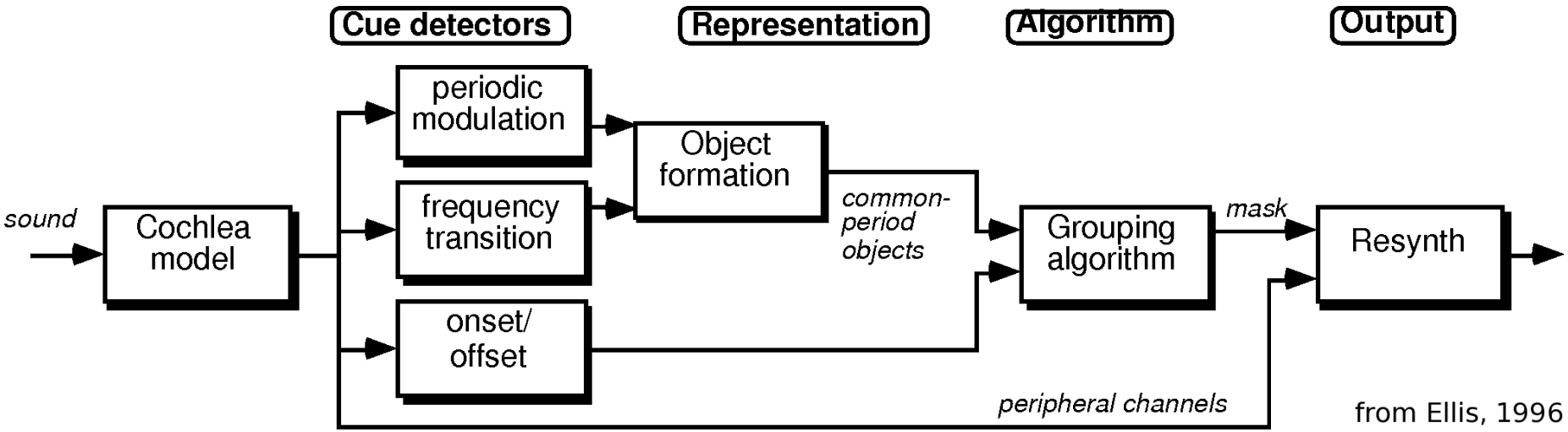
Software mixing has a similar effect.

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.

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).



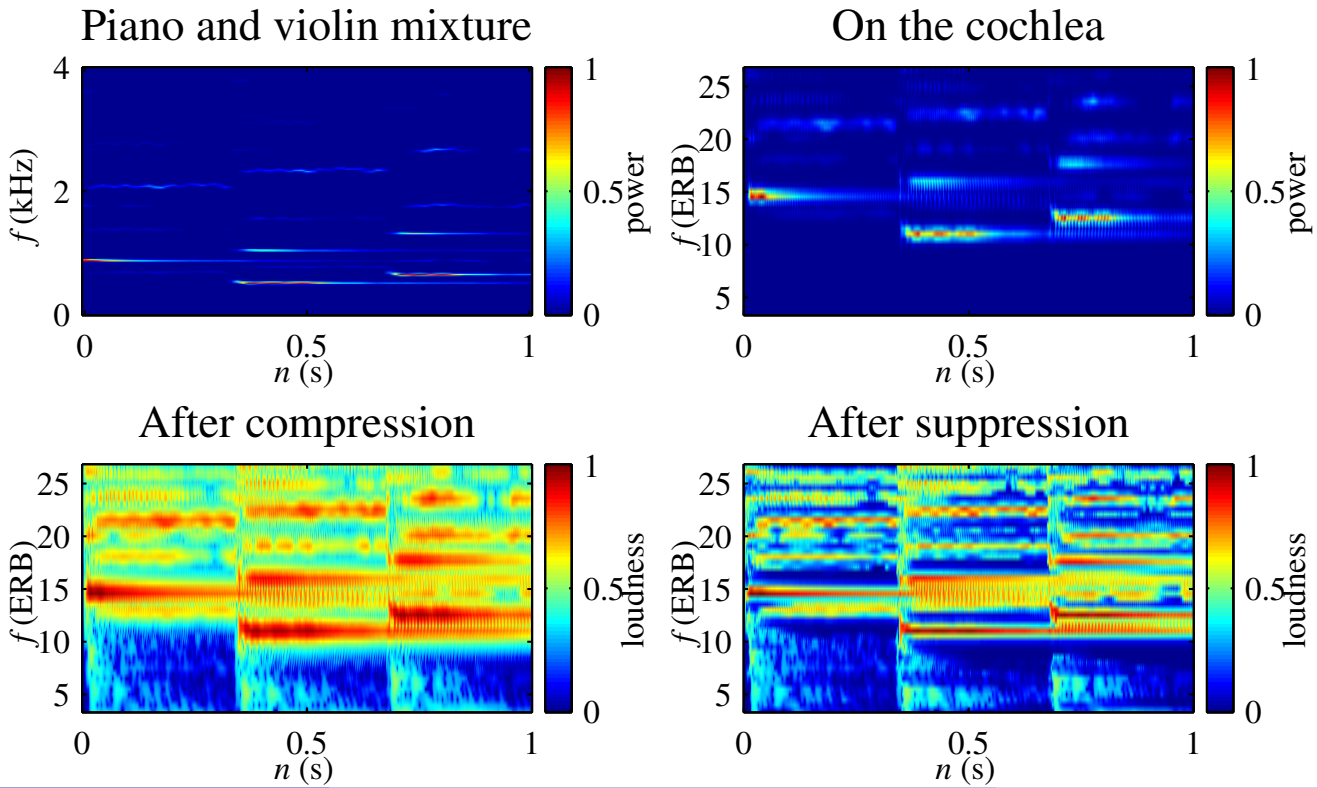
Source formation relies on the *Gestalt* rules of cognition:

- 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:

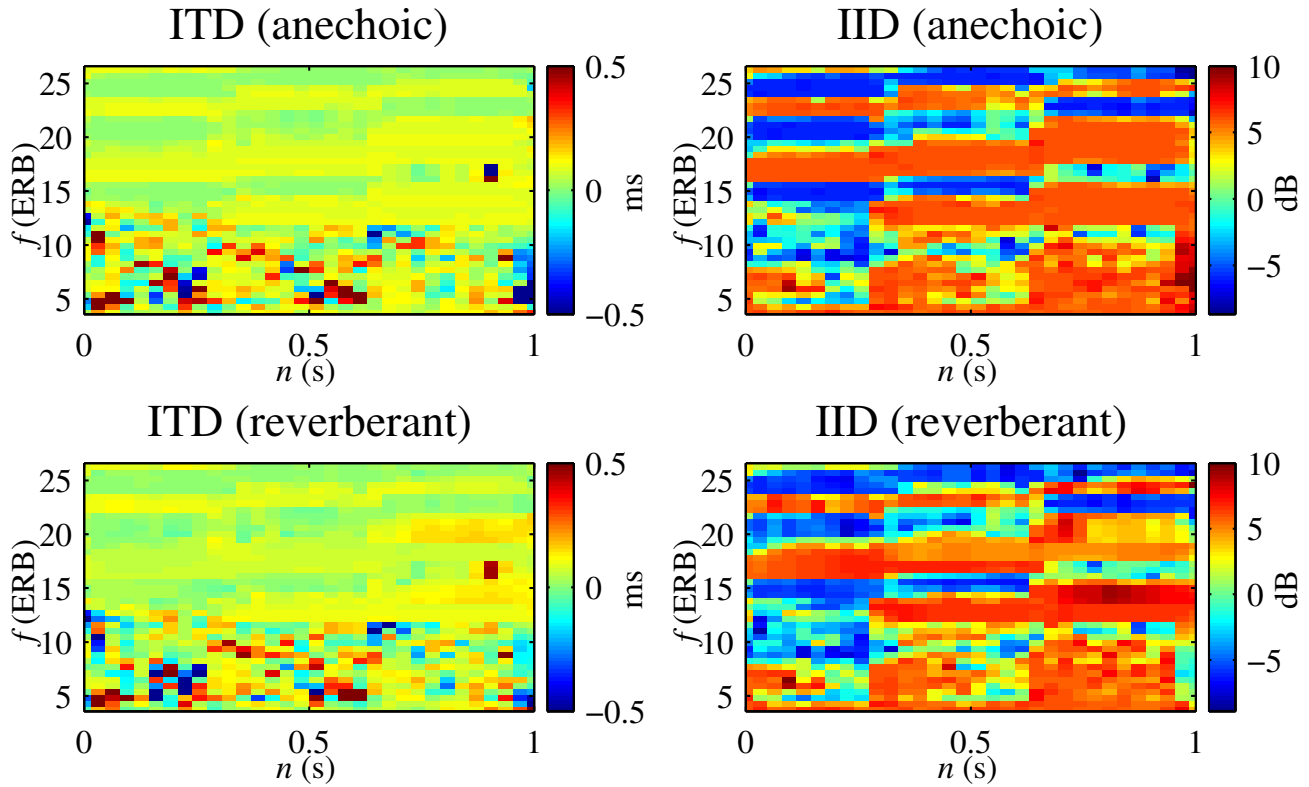
- outer- and middle-ear: filter
- cochlear traveling wave model: filterbank
- haircell model: halfwave rectification + bandwise compression + cross-band suppression



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).

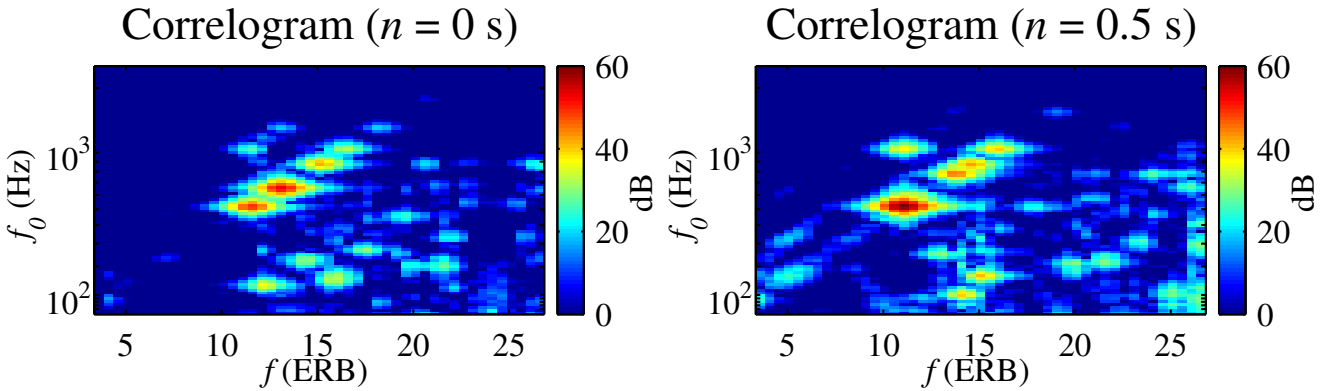


Spectral cues

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

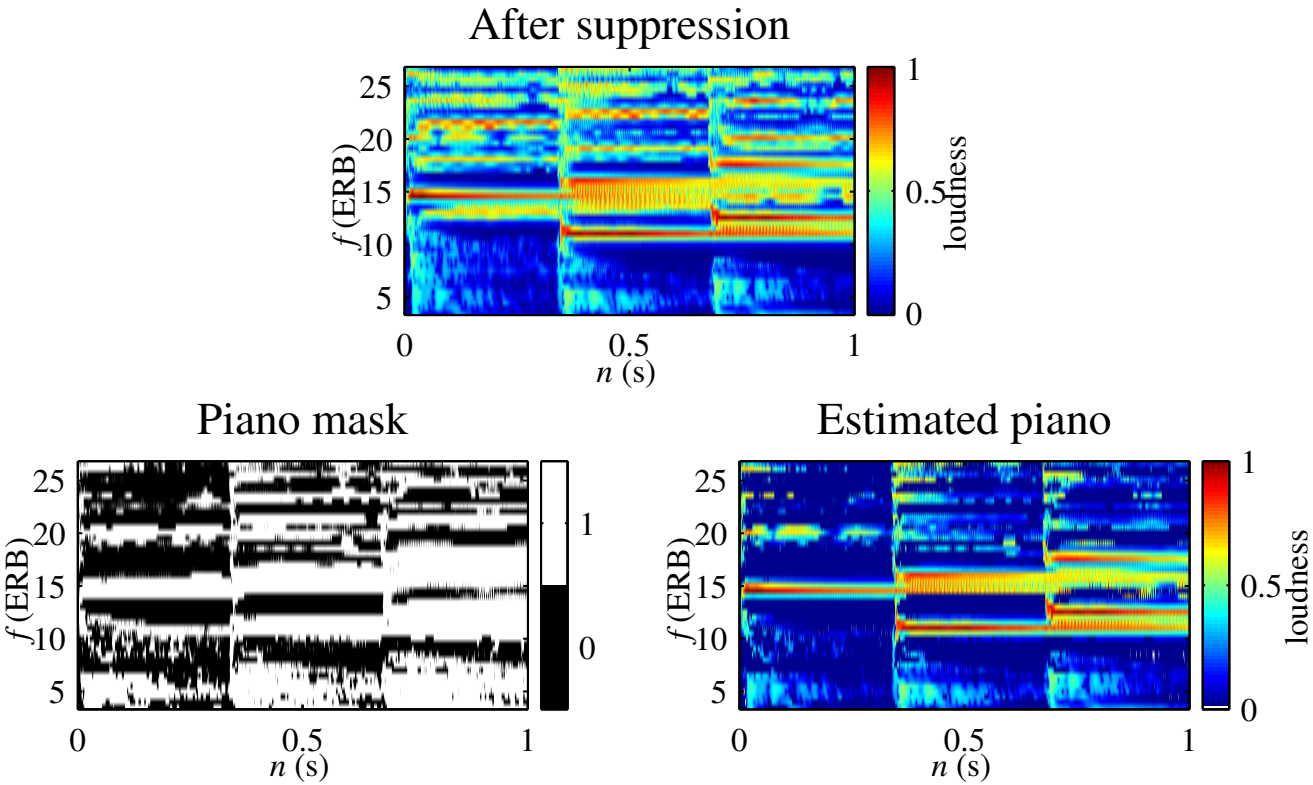
In addition to the primitive cues above, the human auditory system exploits learned cues:

- 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.



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.

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),

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

$\mathbf{x}(t,f)$: STFT of the mixture channels
N: number of sources
$s_n(t,f)$: STFT of source n
$\mathbf{b}(t,f)$: echoes, reverberation and noise

where

$$\mathbf{d}(f,\tau_n) = \begin{pmatrix} 1 \\ g_n e^{-2i\pi f \tau_n} \end{pmatrix}$$

is the steering vector associated with the ITD τ_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) θ, elevation δ 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 θ and δ 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

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

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.

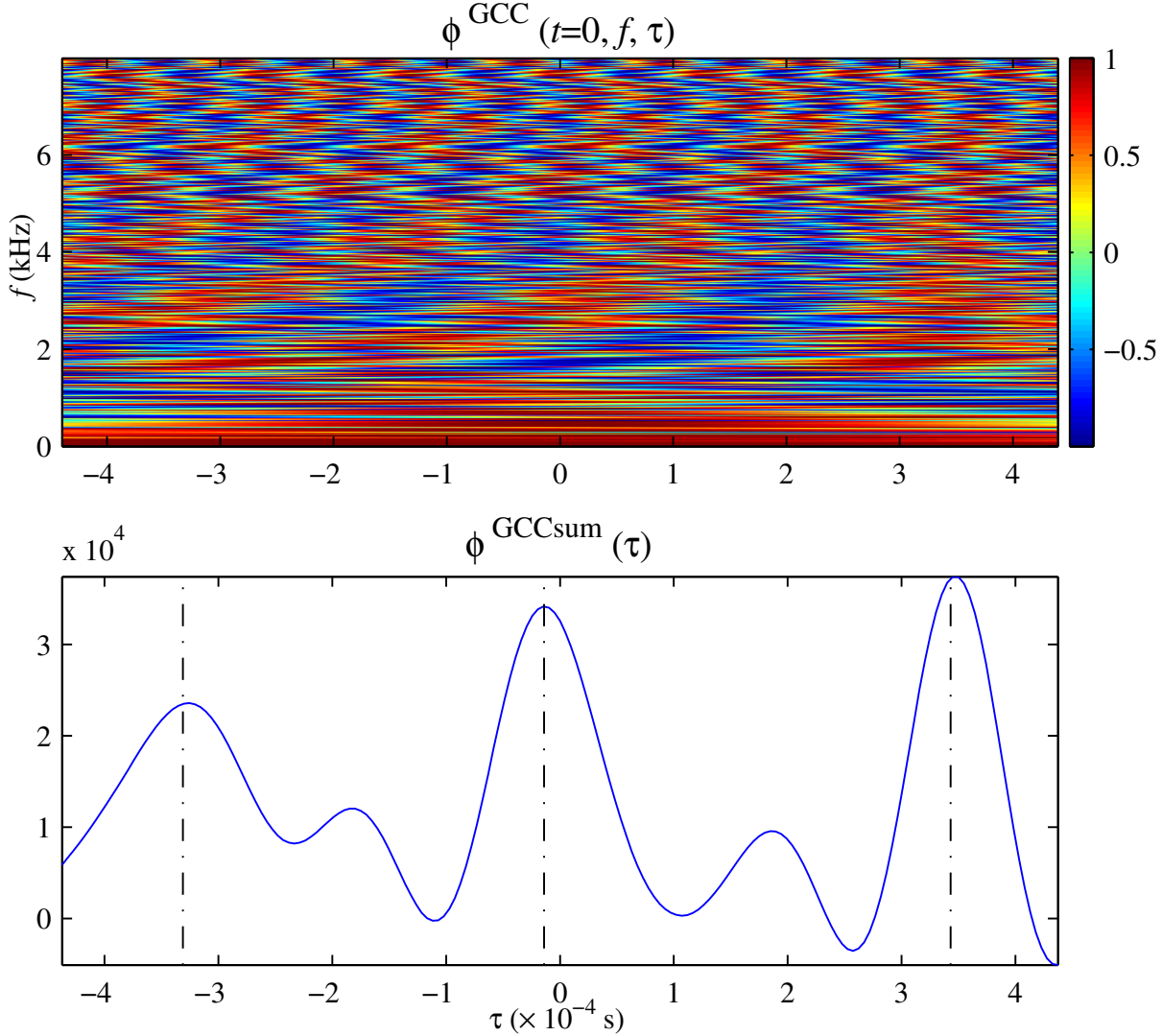
Note that this function exploits ITD only.

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

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

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

GCC-PHAT and its variants (2)



$N = 3$, $r = 50$ cm, $d = 15$ cm, $\mathrm{RT}_{60} = 500$ ms

MUSIC (1)

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

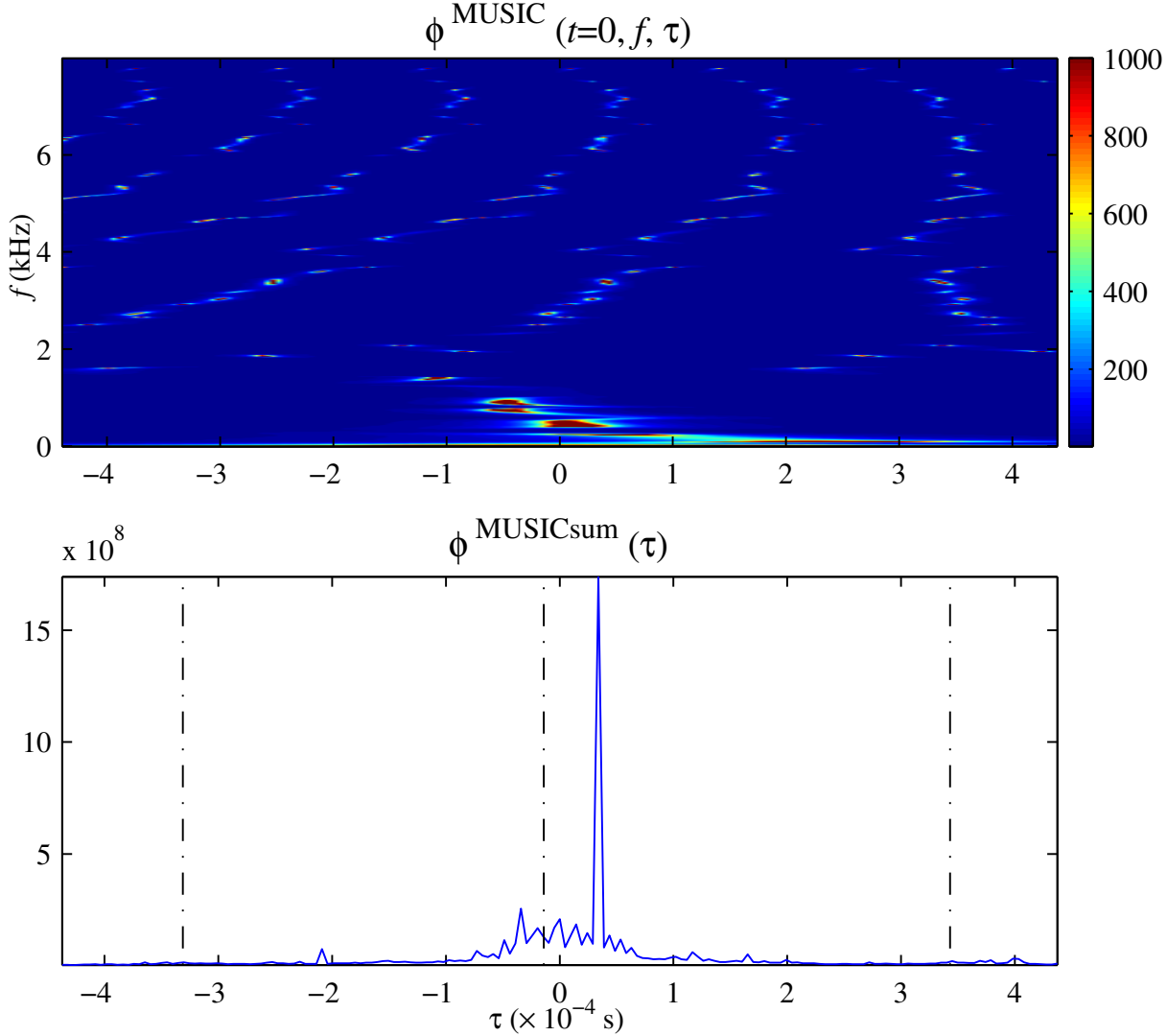
$$\widehat{\mathbf{R}}_{\mathbf{xx}}(t,f) = \frac{\sum_{t',f'} w(t'-t, f'-f)\mathbf{x}(t',f')\mathbf{x}(t',f')^H}{\sum_{t',f'} w(t'-t, f'-f)}$$

The local angular spectrum is defined by

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

where $\mathbf{v}(t,f)$ is the principal eigenvector of $\widehat{\mathbf{R}}_{\mathbf{xx}}(t,f)$.

MUSIC (2)



$N = 3$, $r = 50$ cm, $d = 15$ cm, $\mathrm{RT}_{60} = 500$ ms

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

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

where

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

is the interchannel coherence.

SNR-based angular spectra (2)

Valin et al. use the following weighting instead:

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

where

$$\gamma_i^{\mathrm{MCRA}}(t,f) = \frac{\mathrm{SNR}_i}{1+\mathrm{SNR}_i}$$

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)

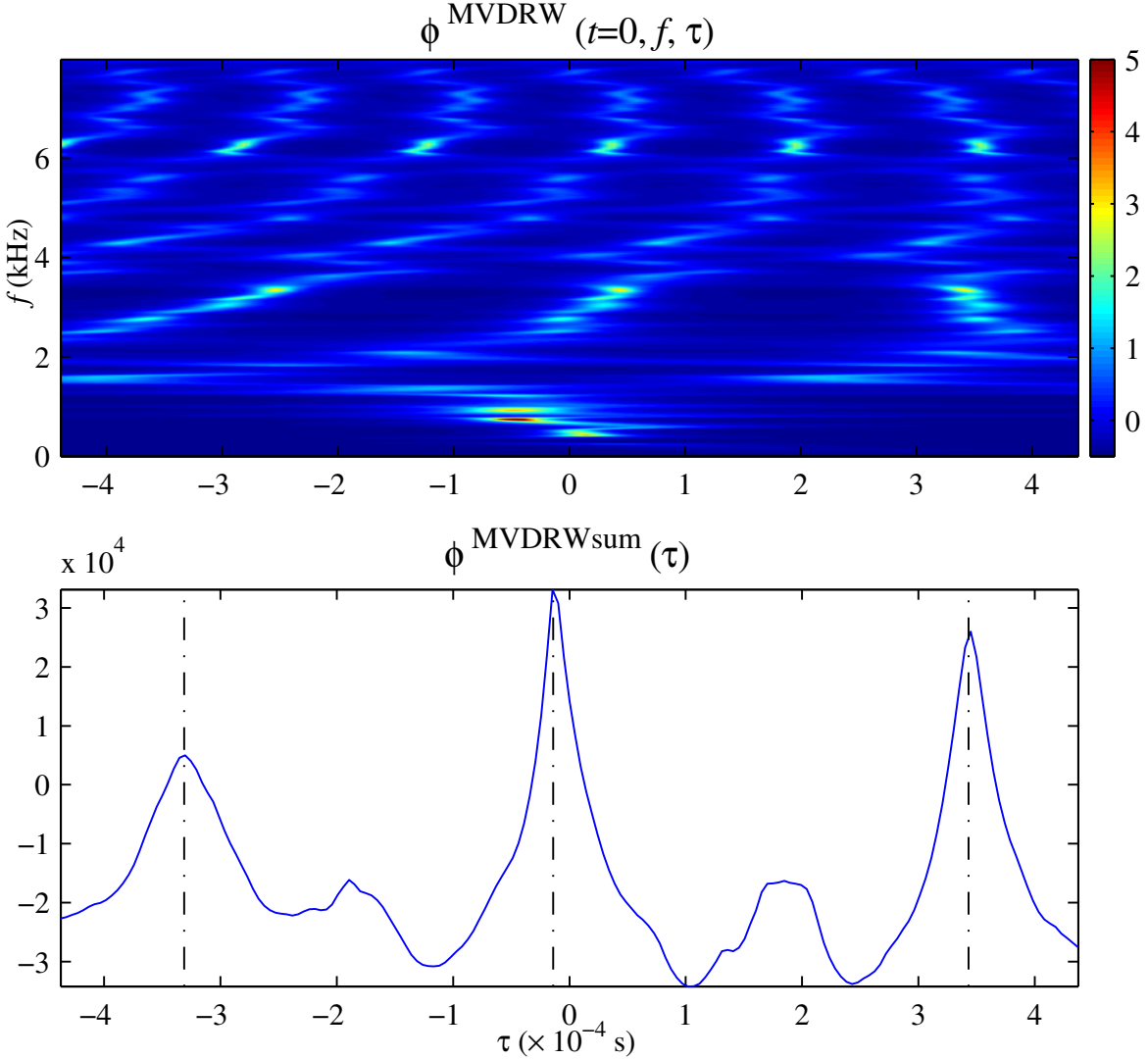
One may also define the angular spectrum to be the SNR itself.

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

$$\phi^{\mathrm{MVDR}}(t,f,\theta,\delta) = \frac{\left(\mathbf{d}(f,\theta,\delta)^H \widehat{\mathbf{R}}_{\mathbf{xx}}(t,f)^{-1} \mathbf{d}(f,\theta,\delta)\right)^{-1}}{\frac{1}{2}\,\mathrm{tr}\left(\widehat{\mathbf{R}}_{\mathbf{xx}}(t,f)\right) - \left(\mathbf{d}(f,\theta,\delta)^H \widehat{\mathbf{R}}_{\mathbf{xx}}(t,f)^{-1} \mathbf{d}(f,\theta,\delta)\right)^{-1}}$$

We will consider in the following a frequency-weighted version of this criterion called MVDRW.

SNR-based angular spectra (4)



$N = 3$, $r = 50$ cm, $d = 15$ cm, $\mathrm{RT}_{60} = 500$ ms

cSCT

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 talk).

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 τ_1 and τ_2 are the TDOAs of the two predominant sources in this neighborhood.

The state coherence transform (SCT) spectrum is defined by

$$\phi^{\mathrm{cSCT}}(t,f,\tau) = \sum_{n=1}^{2} \rho\left(\frac{1}{2}\left|e^{-2i\pi f\tau(\theta,\delta)} - a_n(t,f)\right|\right)$$

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 ± 7 frequency bands and ± 1 time frame
- linear TDOA search grid, corresponding to a resolution of 0.6° in the center and 1.3° on the sides

Only peaks differing by at least 5° are selected.

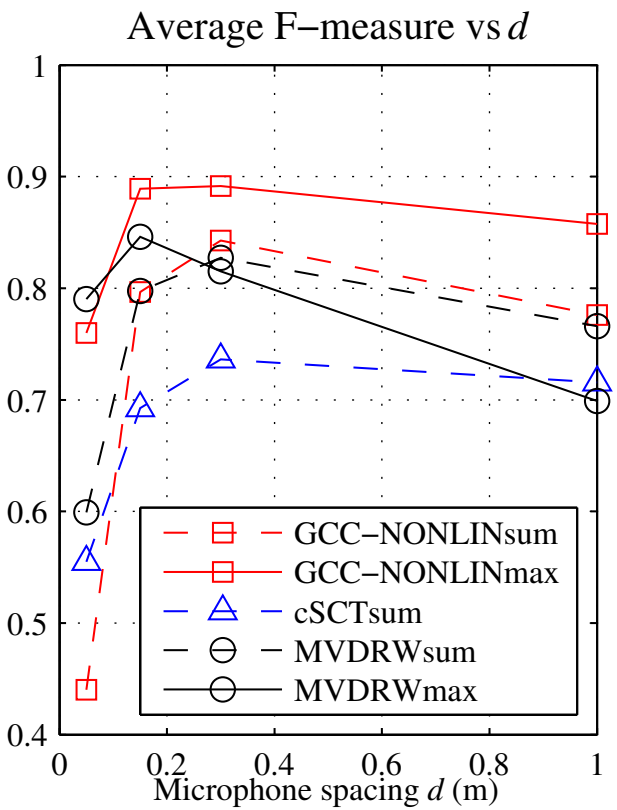
Localization performance is evaluated in terms of

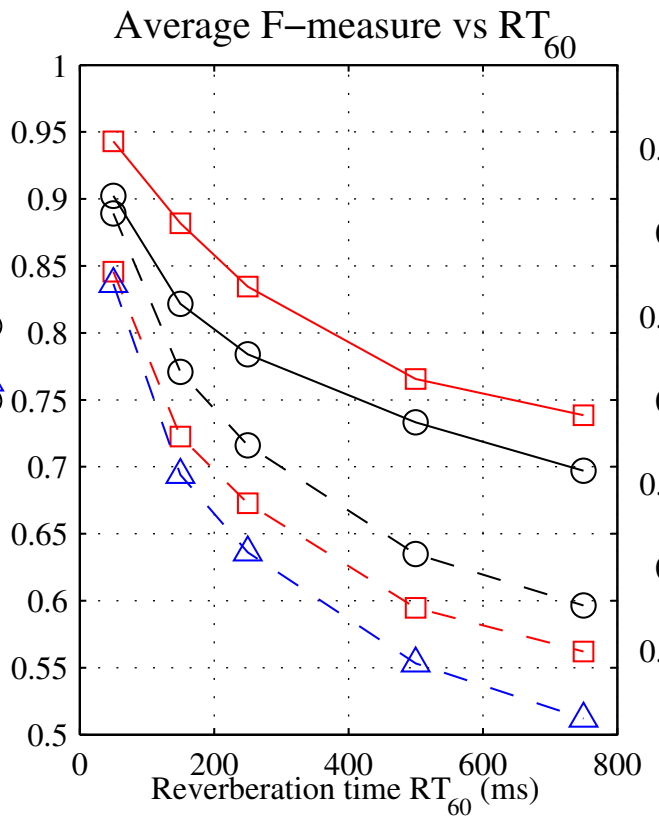
- detection of the correct source locations, as measured by the F-measure for a tolerance of 5° (optimal number of peaks known),
- precision of the correctly detected locations, as measured by the standard deviation for all methods with an F-measure larger than 0.7.

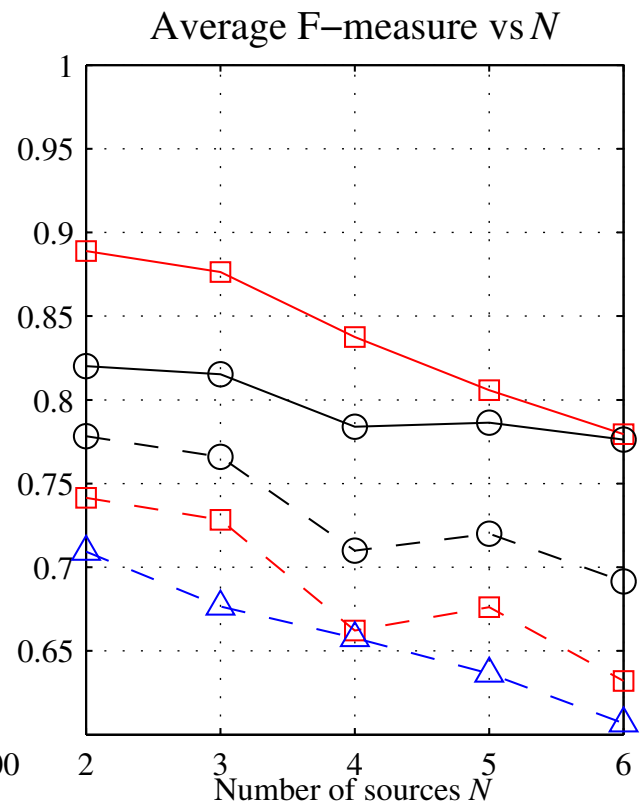
Evaluation (3)

Local angular spectrum function	F-measure (0 to 1)		Accuracy (degrees)	
	ϕ^{sum}	ϕ^{max}	ϕ^{sum}	ϕ^{max}
GCC-PHAT	0.65	0.84	-	**0.41**
GCC-NONLIN	0.69	**0.85**	-	0.43
MUSIC	0.43	0.41	-	-
MVDRW	**0.74**	0.80	0.54	0.46
cSCT	0.66	-	0.83	-

Evaluation (4)







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

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

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

The algorithm consists of:

- associating bin (t,f) with the source n minimizing $\|\tilde{\mathbf{x}}(t,f) - \mathbf{d}(f,\tau_n)\|$
- reestimating τ_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:

$$\mathbf{x}(t,f) = s_{n_{tf}}(t,f)\mathbf{d}(f,\tau_{n_{tf}}) + \mathbf{b}(t,f)$$

n_{tf}: predominant source
$\mathbf{b}(t,f)$: diffuse noise

where $s_{n_{tf}}(t,f)$ is a deterministic and $\mathbf{b}(t,f)$ follows a Gaussian distribution with covariance $v^b(t,f)\mathbf{\Psi}(f)$ where

$$\mathbf{\Psi}(f) = \begin{pmatrix} 1 & \mathrm{sinc}(2\pi f \frac{d}{c}) \\ \mathrm{sinc}(2\pi f \frac{d}{c}) & 1 \end{pmatrix}$$

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_{tf},
- M-step: update $s_{n_{tf}}(t,f)$, $v^b(t,f)$ and τ_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 $v^s(t,f)$.

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

- E-step: estimate the posterior probability of n_{tf},
- M-step: update $v^s(t,f)$, $v^b(t,f)$ and τ_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):

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

where $s_n(t,f)$ and $\mathbf{b}(t,f)$ follow Gaussian distributions with variance $v_n^s(t,f)$ and covariance $v^b(t,f)\mathbf{\Psi}(f)$.

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 $v_n^s(t,f)$, $v^b(t,f)$ and τ_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).

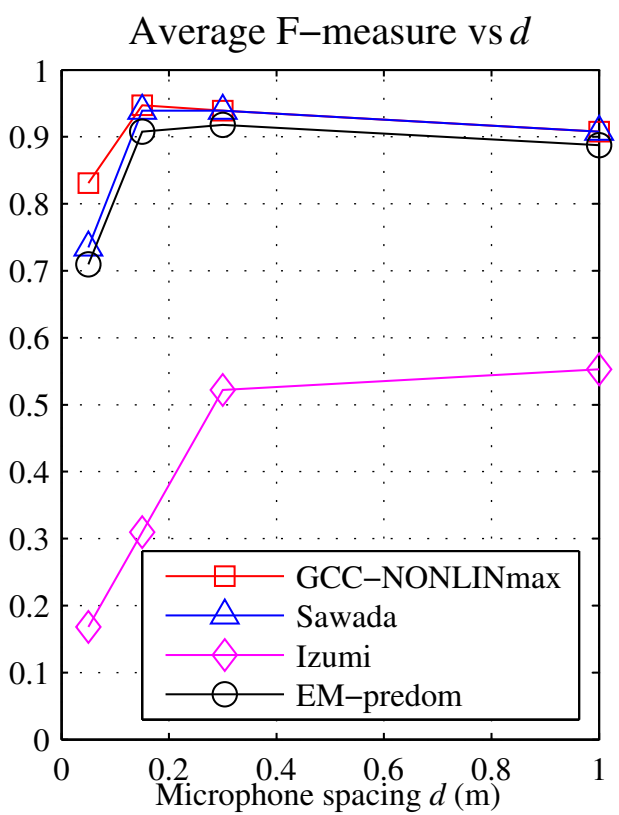
For all algorithms, the following parameter values were used for the computation of the STFT and the empirical covariance of the mixture:

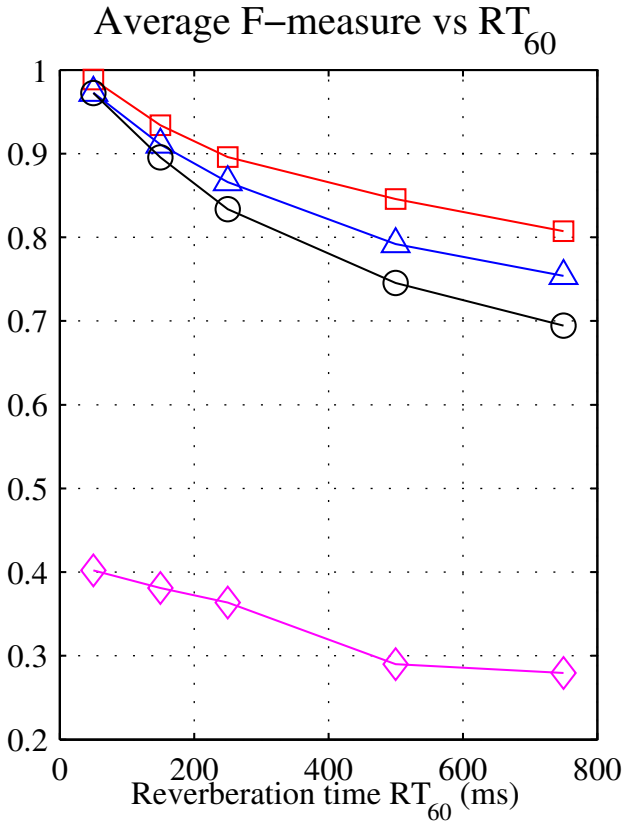
- half-overlapping Hanning windows of 1024 samples (64 ms)
- neighborhood of ± 1 frequency band and ± 1 time frame
- 100 iterations at most

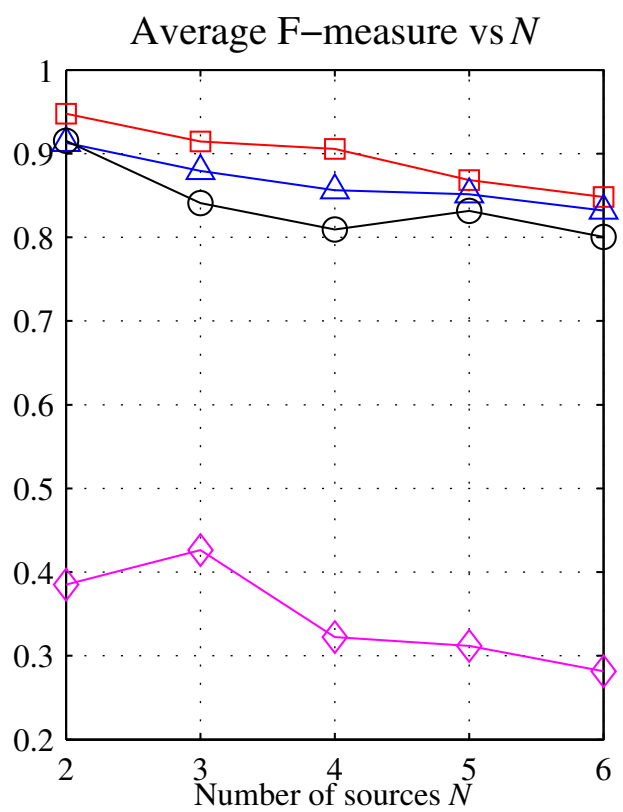
Evaluation (2)

Clustering algorithm	F-measure (0 to 1)		Accuracy (degrees)	
	rand	init	rand	init
None (GCC-NONLINmax)	-	**0.90**	-	**0.56**
Sawada *et al.*	0.49	0.87	-	0.75
Izumi *et al.*	0.30	0.35	-	-
EM-predom	**0.52**	0.85	-	0.66
EM-multi	0.32	**0.90**	-	**0.56**

Evaluation (3)







Multichannel localization

A pair of in-air omnidirectional microphones for 2D localization of a source in terms of azimuth θ and elevation δ.

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),
- $I > 3$ unaligned microphones (array).

In the first case, θ is found via the ITD and δ 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 $\widehat{\mathbf{R}}_{\mathbf{xx}}(t, f)$ and $\mathbf{d}(f, \tau)$.

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

SRP-PHAT

Following approach 2, GCC-PHAT becomes

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

where $\phi_{ii'}^{\mathrm{GCC-PHAT}}(t, f, \theta, \delta)$ 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)

$$\phi^{\mathrm{GCC-PHATref}}(t, f, \theta, \delta) = \sum_{i'} \phi_{ii'}^{\mathrm{GCC-PHAT}}(t, f, \theta, \delta)$$

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.

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 (see other talks in this Thematic Seminar).

References

C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering", Signal Processing, 92, pp. 1950–1960, 2012.

J. DiBiase, H. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms", in Microphone Arrays: Signal Processing Techniques and Applications, pp. 131-154, Springer, 2001.

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.

G. Lathoud, and I. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays", in Proc. Workshop on Statistical and Perceptual Audio Processing (SAPA), 2004.

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.

J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering", Robotics and Autonomous Systems, 55(3), pp. 216-228, 2007.

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/manyears/
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)