A Source Localization/Separation/Respatialization System Based on Unsupervised Classification of Interaural Cues

Joan Mouba and **Sylvain Marchand**

SCRIME – LaBRI, University of Bordeaux 1
firstname.name@labri.fr

Outline

Outline

Overview

- Given binaural audio mixtures, the system
  - detects more than 4 sources;
  - localizes each source (azimuth);
  - reconstructs each source.
- Given a mono audio source, the system:
  - generates a stereo source;
  - positions the source at any location.

based on

Interaural Cues (ILD, ITD)

Expectation Maximization approach

Outline

Motivation

Why?

- Binaural manipulation of source in mix
- Underdeterminated (degenerated) case

Applications

- Virtual reality, hearing aids, live music. . .

CASA-EM

- Subject independent
- Automatic processing
- Time-frequency processing

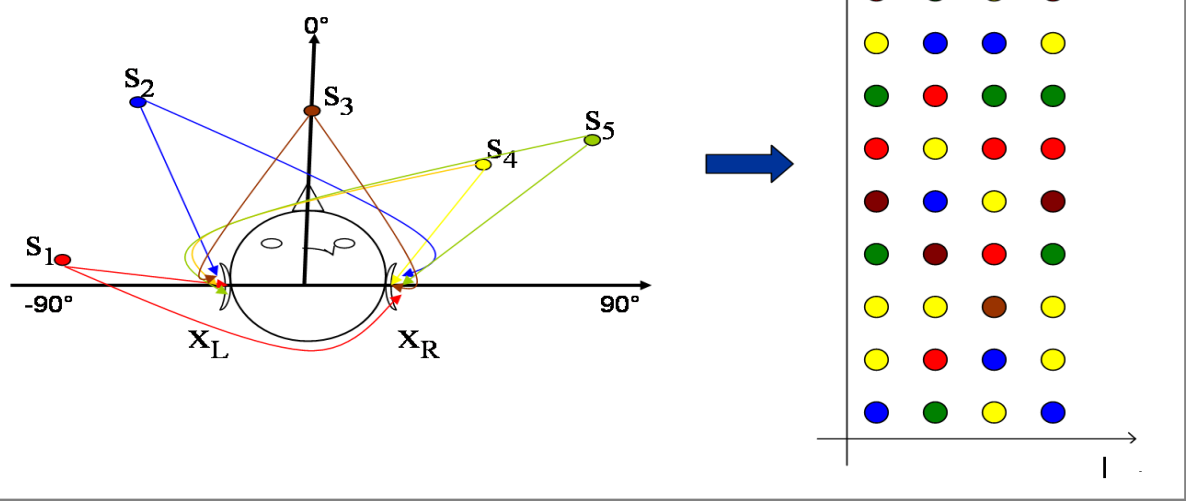
Problem Statement I

Hypothesis

- Sources do not overlap in the t-f plane

Windowed Disjoint Orthogonality

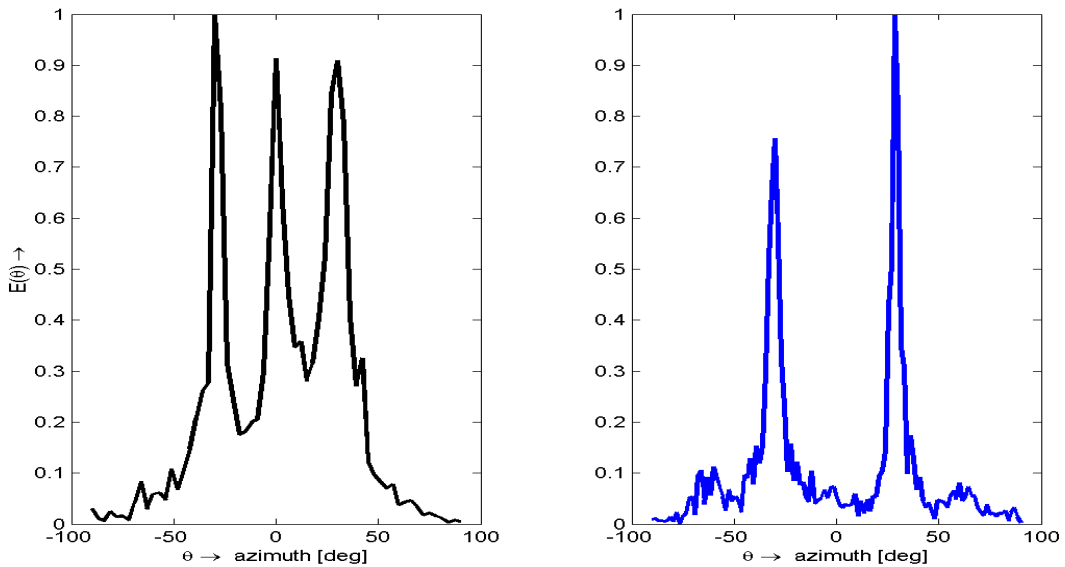
$$S_i(l,f) \cdot S_j(l,f) = 0 \qquad i,j = 1,\ldots,K \qquad i \neq j$$



Problem Statement II

Consequences

- Detection/Localization of phantom sources
- Cumulate energy spreading
- Interferences and distortions



Related Works

DUET: [Rickard (2002)]
- Computes $ILD(l, f)$, $ITD(l, f)$
- 2-dimensional power histogram (ITD$\times$ILD)

[Viste (2003,2004)]
- Estimates azimuth θ given interaural cues
- 1-dimensional power histogram (θ)

[Avendano (2003)]
- Interchannel metric: panning index
- Separation based on Gaussian window

[Kameoka (2004)]
- Spectrum density with tied Gaussian mixture
- Separation of harmonic structures

Head Model

ILD with shadow cast

$$\triangle L(\theta, f) = \alpha_f \frac{\sin\theta}{c}$$

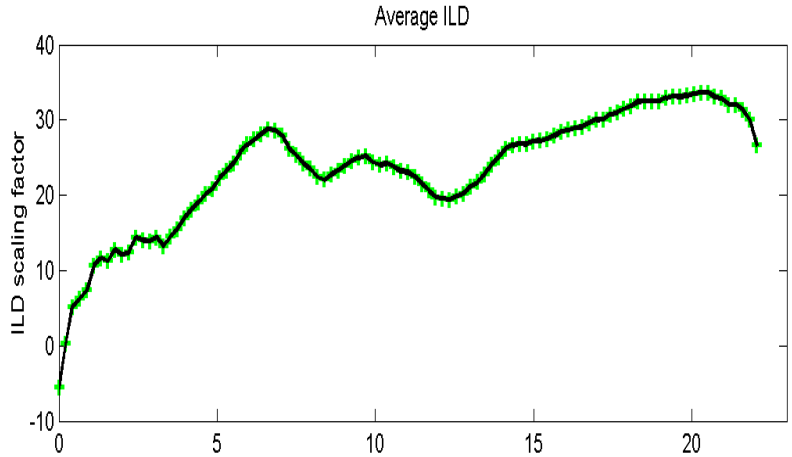
[Viste & Evangelista (2003)]

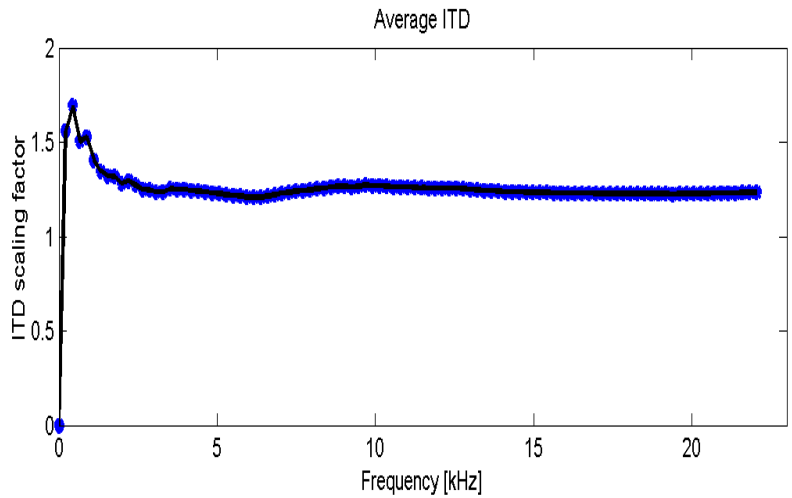
ITD with shadow cast

$$\triangle T(\theta, f) = \beta_f \frac{r(\sin\theta + \theta)}{c}$$

r: head radius c: sound celerity

[Viste & Evangelista (2003)]





Source Localization

Computes interaural cues:

$$\mathrm{ILD}(t,f) = 20\log_{10}\left|\frac{X_R(t,f)}{X_L(t,f)}\right|; \qquad \mathrm{ITD}_p(t,f) = \frac{1}{2\pi f}\left(\angle\frac{X_R(t,f)}{X_L(t,f)} + 2\pi p\right)$$

Computes azimuth based on ILD and ITD:

$$\theta_L(t,f) = \arcsin\left(\frac{c\cdot\mathrm{ILD}(t,f)}{\alpha_f}\right); \quad \theta_{T,p}(t,f) = \Pi\left(\frac{c\cdot\mathrm{ITD}_p(t,f)}{r\cdot\beta_f}\right)$$

$$\text{with } \Pi(x) = 0.50018\,x + 0.009897\,x^3 + 0.00093\,x^5 + O(x^5)$$

Finds p that minimizes:

$$\theta(t,f) = \theta_{T,m}(t,f) \text{ with } m = \mathrm{argmin}_p\left|\theta_L(t,f) - \theta_{T,p}(t,f)\right|$$

Cumulates the power in a histogram using a binary mask:

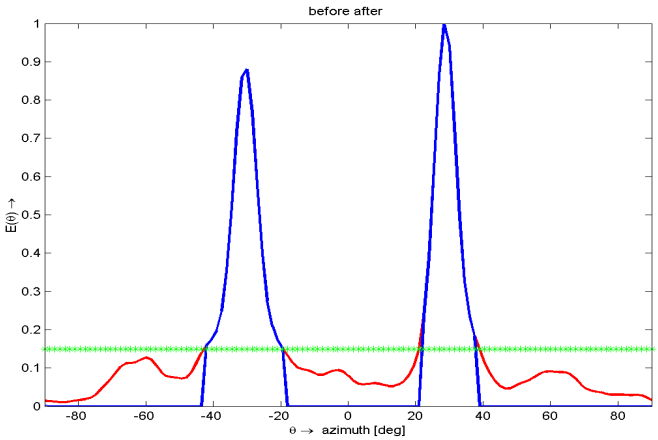
$$h(\theta) = \sum_f |M_\theta(t,f)X_L(t,f)X_R(t,f)|$$

Outline

Source Localization/Separation Method

- Build histogram $h(\theta)$
- Binomial smoothing and thresholding
- Local maxima search
- Outputs

  - Mixture order estimate (K)
  - Locations of detected sources ($\theta_1, \theta_2, \cdots \theta_K$)



Example

2-source mixture

- K = 6, before threshold
- K = 2, after threshold

Gaussian Mixture Model (GMM)

- $\Theta = \{\theta_1, ..., \theta_N\}$
- Each source associated to a Gaussian
- Gaussian Mix: $\{\Gamma\} = \{\mu_j, \sigma_j, \pi_j \mid j = 1, ..., K\}$:
  mean, *standard deviation*, *weight* for source j

$$f_K(\theta|\Gamma) = \sum_{j=1}^{K} \pi_j \, \phi_j(\theta|\gamma_j)^{h(\theta)} \qquad \textit{with} \qquad \sum_{j=1}^{K} \pi_j = 1$$

Find Γ that best matches the data:

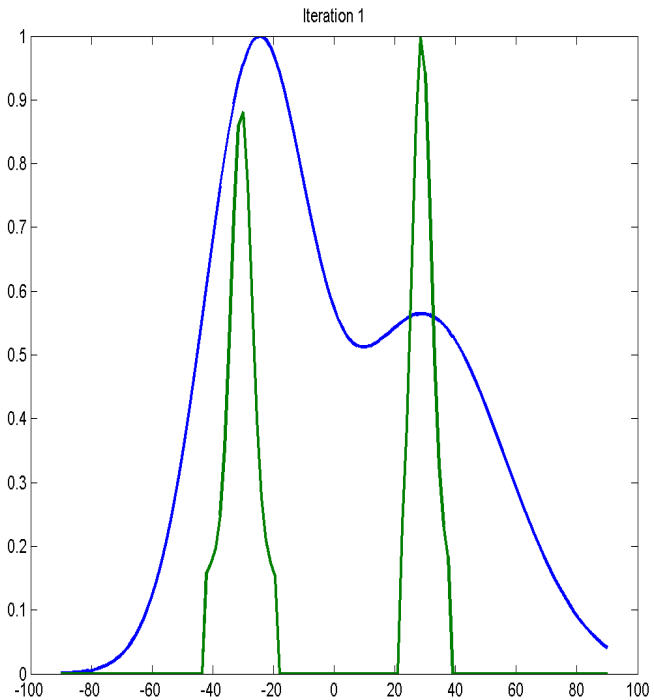
Maximum Likelihood-Expectation Maximization

objective: $\Gamma^{(t+1)} = \text{argmax}_\Gamma \, \mathcal{L}(\Gamma|\Theta) - \mathcal{L}(\Gamma^{(t)}|\Theta).$

EM Updates

- 2-order mix

s	θ_{ori}	θ_{est}	$\lvert\theta_{err}\rvert$
s_1	-30	-30.63	0.63
s_2	30	29.31	0.69



EM Updates

$$P_K(k|\theta,\Gamma) \leftarrow \frac{P_K(\theta,k|\Gamma)}{P_K(\theta|\Gamma)}$$

$$\pi_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)}$$

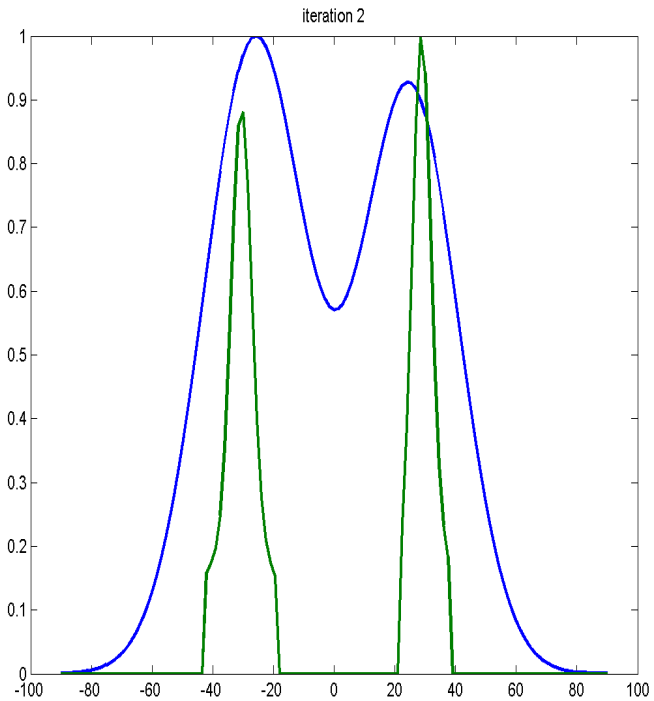
$$\mu_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, \theta\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

$$\sigma_k^2 \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, (\theta-\mu_k)^2\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

EM Updates

- 2-order mix

s	θ_{ori}	θ_{est}	$\lvert\theta_{err}\rvert$
s_1	-30	-30.63	0.63
s_2	30	29.31	0.69



EM Updates

$$P_K(k|\theta,\Gamma) \leftarrow \frac{P_K(\theta,k|\Gamma)}{P_K(\theta|\Gamma)}$$

$$\pi_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)}$$

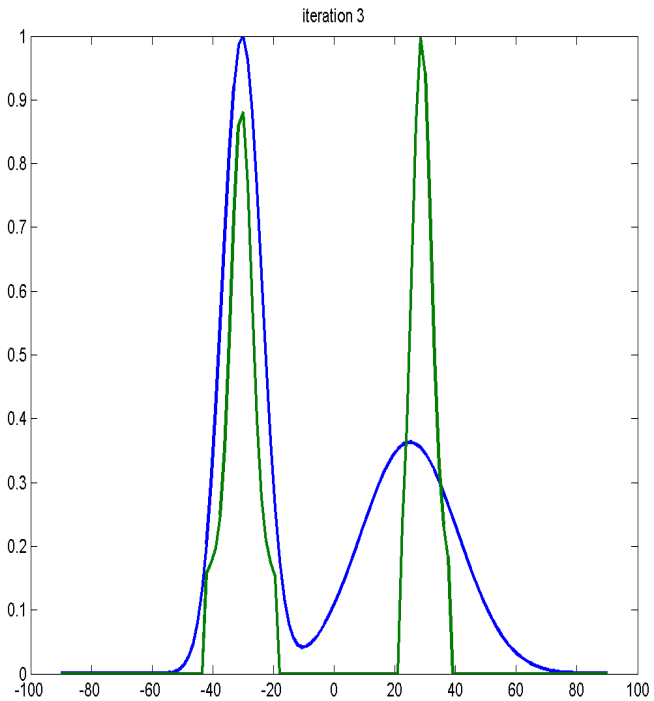
$$\mu_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, \theta\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

$$\sigma_k^2 \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, (\theta-\mu_k)^2\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

EM Updates

- 2-order mix

s	θ_{ori}	θ_{est}	$\lvert\theta_{err}\rvert$
s_1	-30	-30.63	0.63
s_2	30	29.31	0.69



EM Updates

$$P_K(k|\theta,\Gamma) \leftarrow \frac{P_K(\theta,k|\Gamma)}{P_K(\theta|\Gamma)}$$

$$\pi_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)}$$

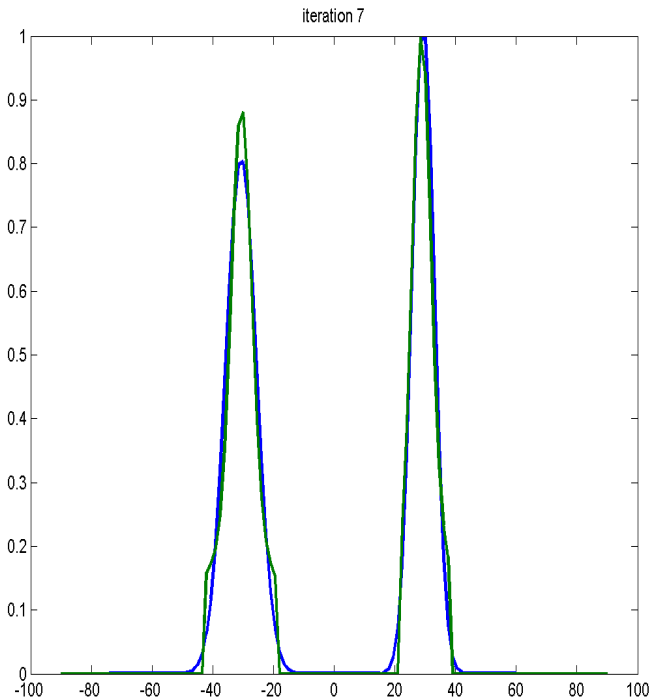
$$\mu_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, \theta\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

$$\sigma_k^2 \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\, (\theta-\mu_k)^2\, P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\, P_K(k|\theta,\Gamma)}$$

EM Updates

- 2-order mix

s	θ_{ori}	θ_{est}	$\lvert\theta_{err}\rvert$
s_1	-30	-30.63	0.63
s_2	30	29.31	0.69



EM Updates

$$P_K(k|\theta,\Gamma) \leftarrow \frac{P_K(\theta,k|\Gamma)}{P_K(\theta|\Gamma)}$$

$$\pi_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)}$$

$$\mu_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ \theta\ P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta,\Gamma)}$$

$$\sigma_k^2 \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ (\theta-\mu_k)^2\ P_K(k|\theta,\Gamma)}{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta,\Gamma)}$$

Unmixing with probabilistic t-f Mask

Philosophy

each t-f bin belongs to all K sources

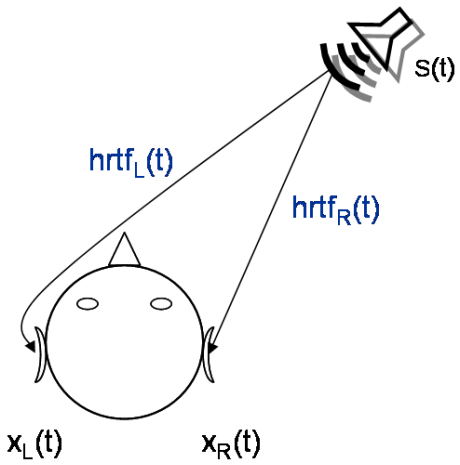
Build a probabilistic mask for each source k

$$M_k(t,f) = P_K(k|\theta(t,f),\Gamma)$$

Energy allocation according to posterior probability

$$\begin{aligned} S_L(t,f) &= M_k(t,f) \cdot X_L(t,f) \\ S_R(t,f) &= M_k(t,f) \cdot X_R(t,f) \end{aligned}$$

Binaural Spatialization Method 1



$hrtf_{subject}(\rho, \theta, \phi, f)$ depends on:
subject, position, frequency

CIPIC hrtf database (45 subjects)

[Algazi et al (2001)]

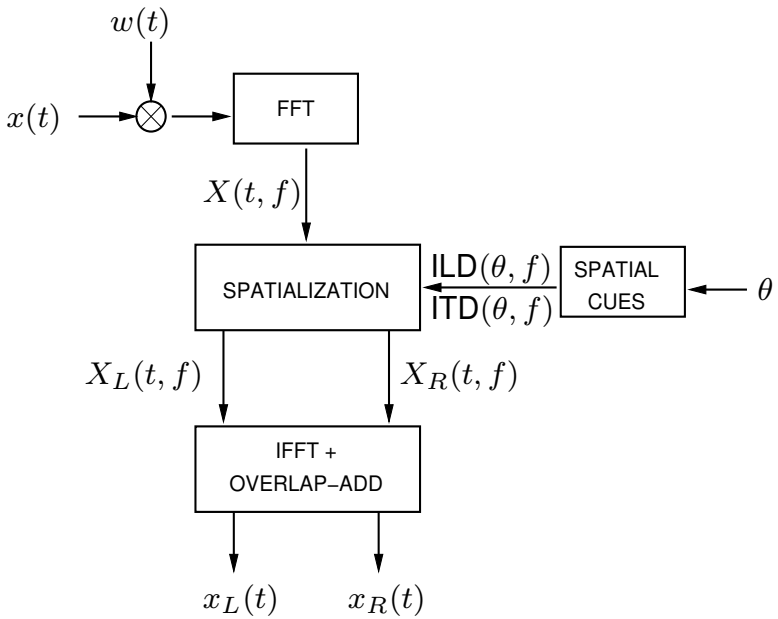
Disk space

- Table of 25×101 reals
- Interpolation not trivial...

Spatialization

$$x_L = s * \text{mean-hrtf}_L(\theta)$$
$$x_R = s * \text{mean-hrtf}_R(\theta)$$

Binaural Spatialization Method 2



Disk space

- Array of 202 reals
- Geometrical interpolation

Spatialization

$$X_L(t,f) = X(t,f) \cdot 10^{-\Delta_a/2} e^{-j\Delta_\phi/2}$$
$$X_R(t,f) = X(t,f) \cdot 10^{+\Delta_a/2} e^{+j\Delta_\phi/2}$$

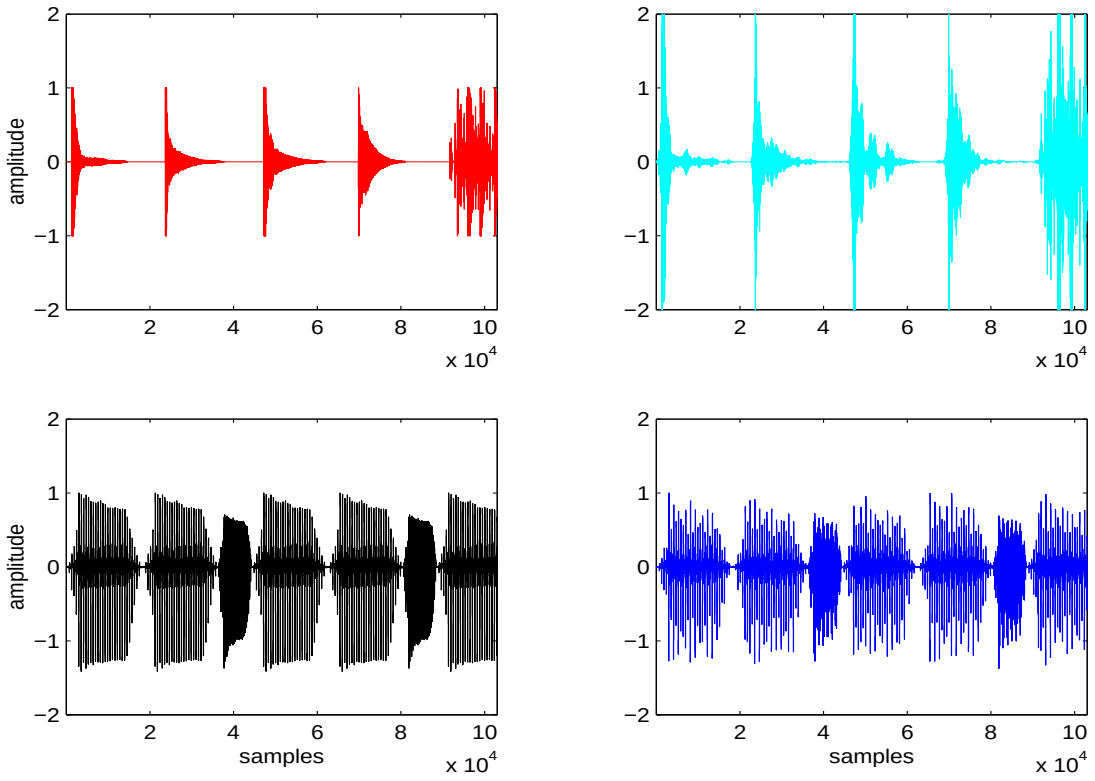
with

$$\Delta_a = \mathrm{ILD}(\theta,f)/(20\mathrm{dB})$$
$$\Delta_\phi = \mathrm{ITD}(\theta,f) \cdot 2\pi f$$

Outline

Source Separation Results: Signals

xylophone ($-55°$) (top) and horn ($30°$)



Rhythm respected Shape preserved

→ Unmix similar to original

Source Separation Results: Listening Tests

2-source mix

- Mix
- original eguitar -80°
- unmix eguitar
- original saxo 80°
- unmix saxo

3-source mix

- Mix
- original piano -30°
- unmix piano
- original xylo 0°
- unmix xylo
- original trumpet 30°
- unmix trumpet

Mean Opinion Score: 3 on 5 levels

Source Spatialization Results

ReSPA

- xylo -45°
- fhorn 80°
- saxo -30°
- tuba 0°
- eguitar -80°

Mean HRTF

- xylo -45°
- fhorn 80°
- saxo -30°
- tuba 0°
- eguitar -80°

MHRTF better lateralization SSPA good enough

MHRTF sounds more natural

Outline

Summary

Summary

- Source localization (azimuth)
- Source separation
- Source spatialization

Future Works

- Study the localization of moving sources
- Implement the system in real-time environment
- Improve source separation with processing inside each bin
- Study the brightness of spectra to weight distance
- Conduct further MOS listening tests for spatialization

References

- J. Blauert: *Spatial Hearing*, MIT Press, 1983.
- H. Viste, G. Evangelista: *Binaural Source Localization*, PhD Thesis, 2004.
- O. Yilmaz and S. Rickard: *Blind Separation of Speech Mixtures via Time-Frequency Masking*, IEEE Transactions On signal Processing, Vol.52, NO.7, July 2004.
- V.R. Algazi, R.O. Duda, D.P. Thompson: *The CIPIC HRTF database*, Proc. IEEE WASPAA01, NY, pp.99-102, 2002.
- A. Dempster, N. Laird and D. Rubin: *Maximum Likelihood from Incomplete Data via EM Algorithm*, Journal of the Royal statistical Society series B, vol. 39, no. 1, pp.1-38, 1977.