A Source Localization/Separation/Respatialization System

Joan Mouba

advisors: Myriam Desainte-Catherine & Sylvain Marchand

SCRIME – LaBRI, University of Bordeaux 1

joan.mouba@labri.fr

Motivation

Why?

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

Applications

- Active listening, Live music, Virtual reality . . .

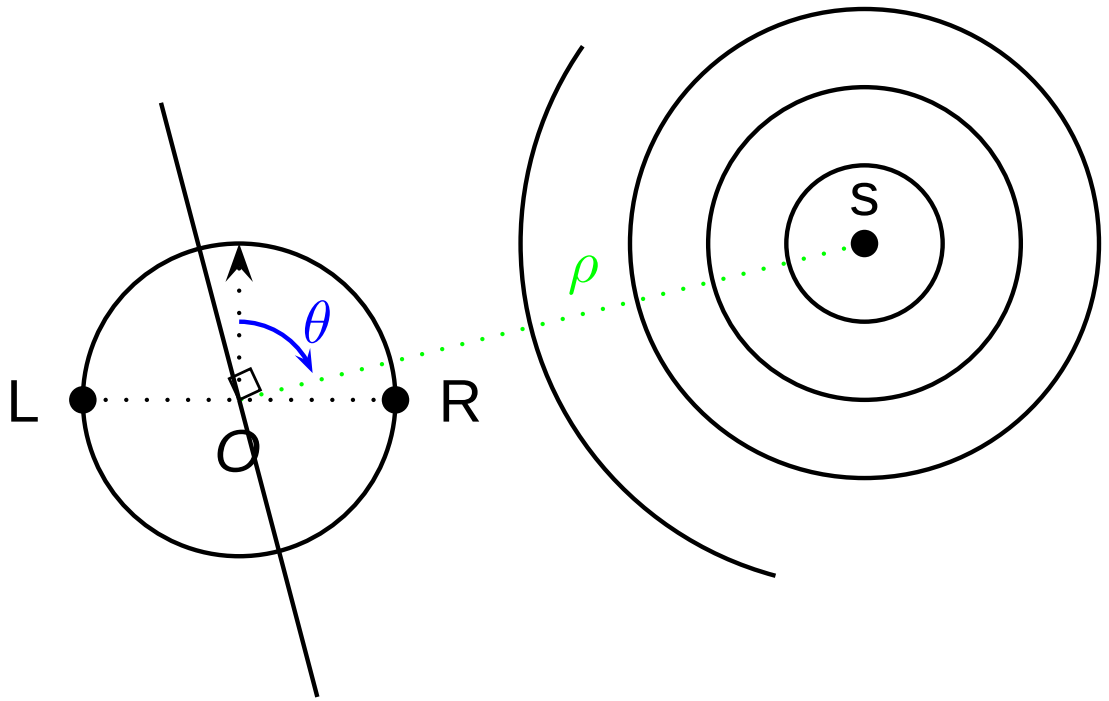
Objectives

- Detect, localize, separate, spatialize source
- Subject customizing with Interaural cues
- Efficient and automatic processing

Outline

- Binaural Head Model
- Binaural Source Localization
- Parametric Binaural Spatialization
- Spatial Demixing with a probabilistic Mask
- Summary and Future Works

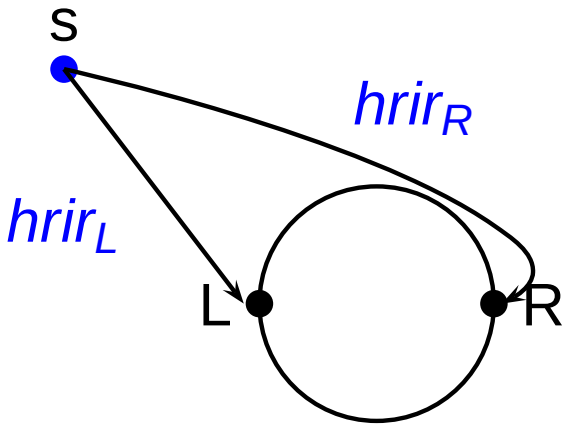
Horizontal plane



- θ azimuth
- ρ range
- $\phi = 0$ elevation nil

- *Left* / *Right* ears
- O Head center
- s sound source

Interaural Cues



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

CIPIC hrtf database (45 subjects)

[Algazi et al (2001)]

- ILD : Interaural Level Differences (from pressure level)
- ITD : Interaural Time Differences (from signal phase)
- ILD, ITD : most important cues in binaural localization
- ITD prominent at low f, ILD crucial at high f
  Duplex theory [Lord Rayleigh (1907)]

Related Works

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

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

ITD with shadow cast

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

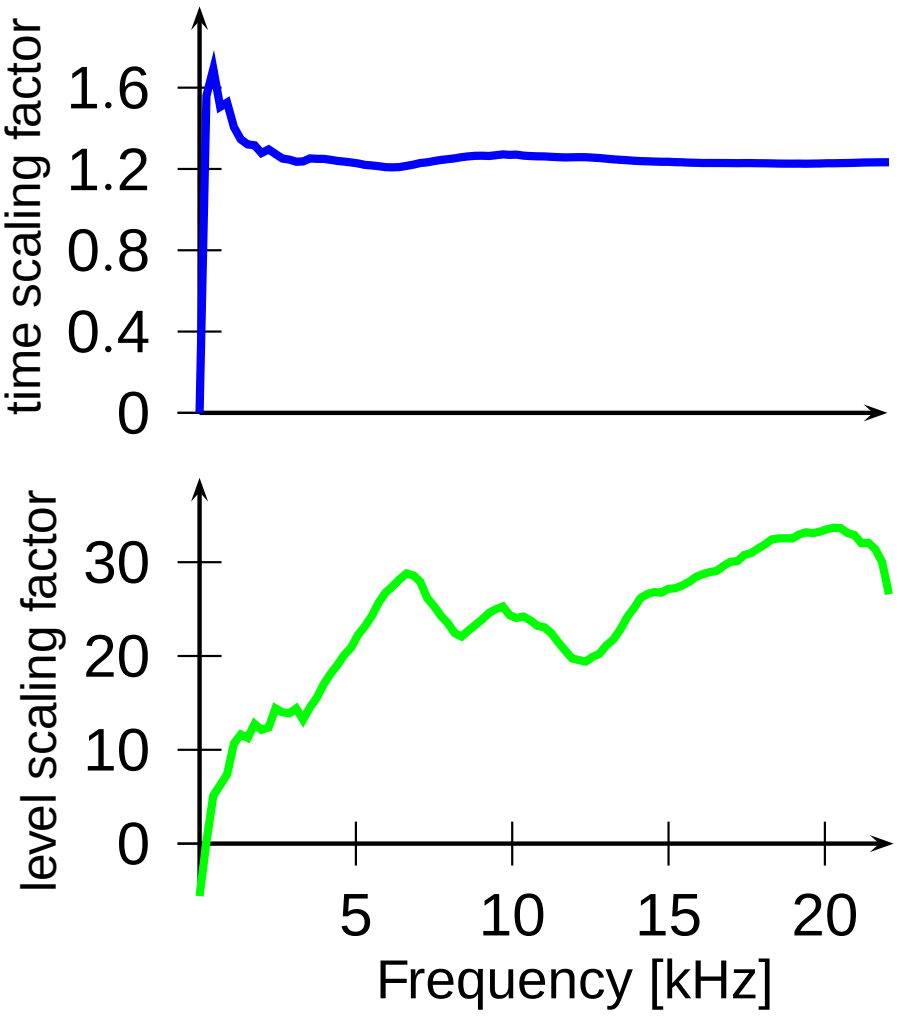
r: head radius c: celerity

[Viste & Evangelista (2003)]

ILD with shadow cast

$$ILD(\theta, f) = \alpha_f \sin(\theta)$$

[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:

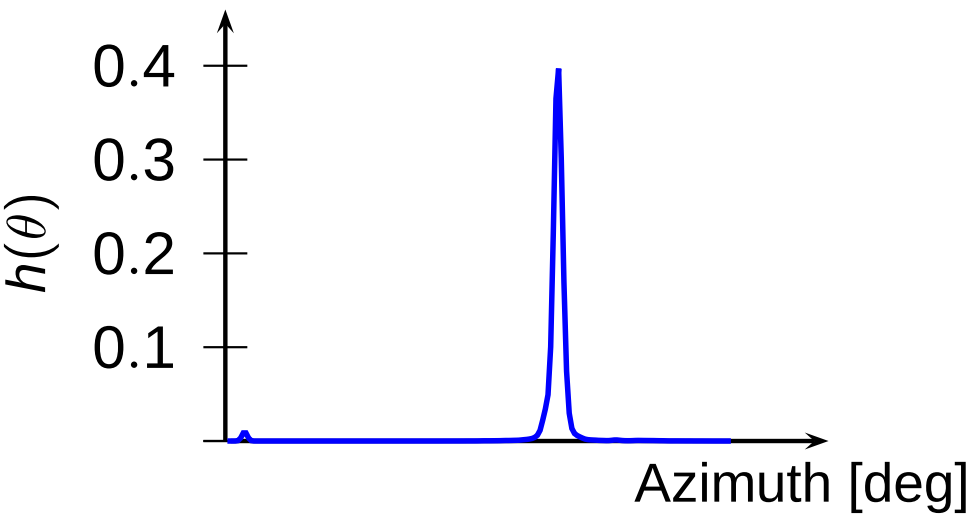
$$\theta(t,f) = \theta_{T,m}(t,f) \text{ with } m = \mathrm{argmin}_p\left|\theta_L(t,f) - \theta_{T,p}(t,f)\right|$$

spread the cumulated power with a gaussian function:

$$h(\theta) = \left(\sum_f |M_\theta(t,f)X_L(t,f)X_R(t,f)|\right)\cdot g(\theta,\sigma^2 = 1)$$

Source Localization

- Finds Local Maximum
- Outputs is the detected source location $\hat{\theta}$



Example

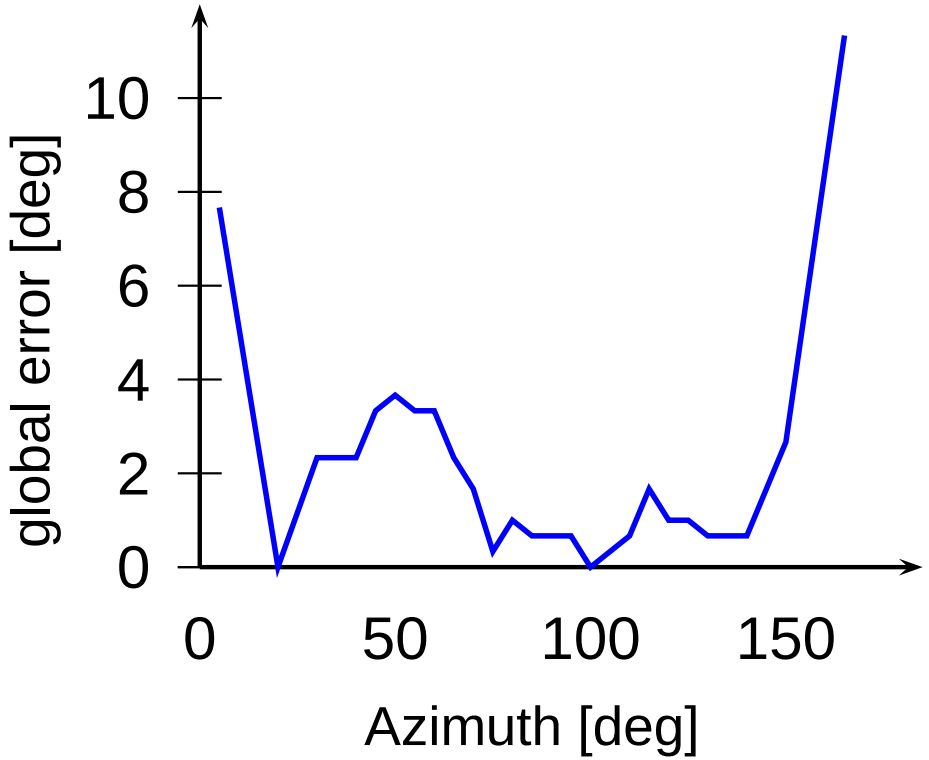
1-singing voice at $\theta = -30°$

- $\hat{\theta} = -28.42°$
- $error = 1.58°$

Gaussianity test: NCC factor above 0.92

Localization Results

azimuthal error on some TIMIT audio files



$FFTlength = 2048$ $overlap = 1024$ $Resolution = 360$

Overall Error less than 3° between 15° and 165°.

Parametric Binaural Spatialization

proposed Spatialization

$$
\begin{aligned}
X_L &= X \cdot 10^{+\Delta_a(f)/2} e^{+j\Delta_\phi(f)/2} \\
X_R &= X \cdot 10^{-\Delta_a(f)/2} e^{-j\Delta_\phi(f)/2} \\
&\textit{with} \\
\Delta_a(f) &= \mathrm{ILD}(\theta, f)/(20\mathrm{dB}) \\
\Delta_\phi(f) &= \mathrm{ITD}(\theta, f) \cdot 2\pi f
\end{aligned}
$$

Disk space

- Array of 202 reals
- Geometical interpolation

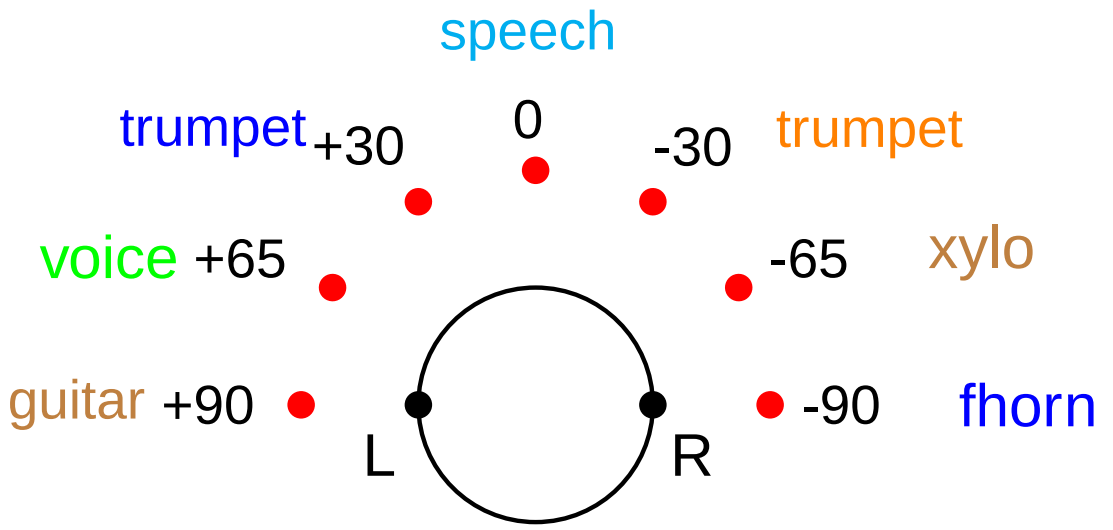
common Spatialization

$$
\begin{aligned}
x_L &= s * \text{mean-hrir}_L(\theta) \\
x_R &= s * \text{mean-hrir}_R(\theta)
\end{aligned}
$$

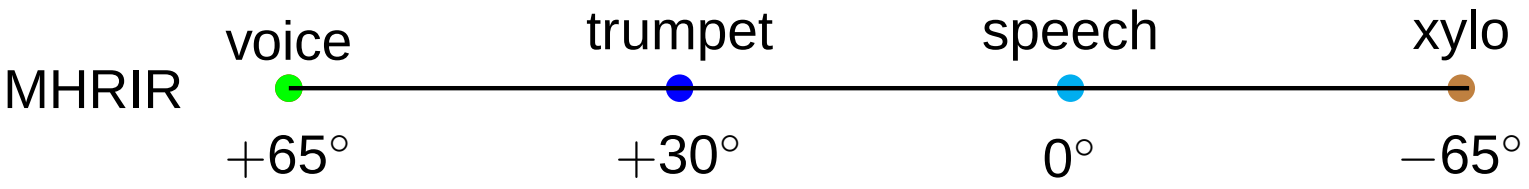
Disk space

-two tables of $25 * 101$ reals
- Interpolation not trivial

Spatialization Examples



azimuthal
moving voice



very good quality - no artifacts
MHRIR sounds more natural

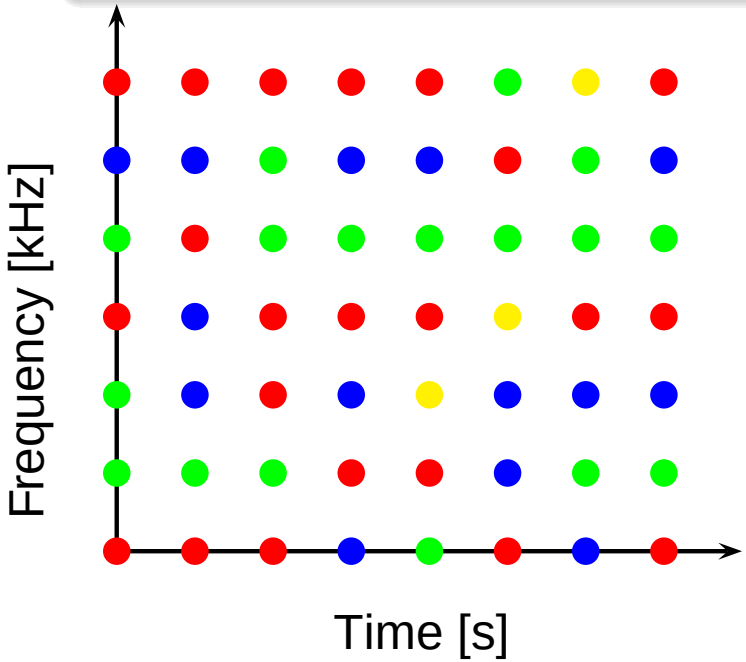
Multi-Source Mixture

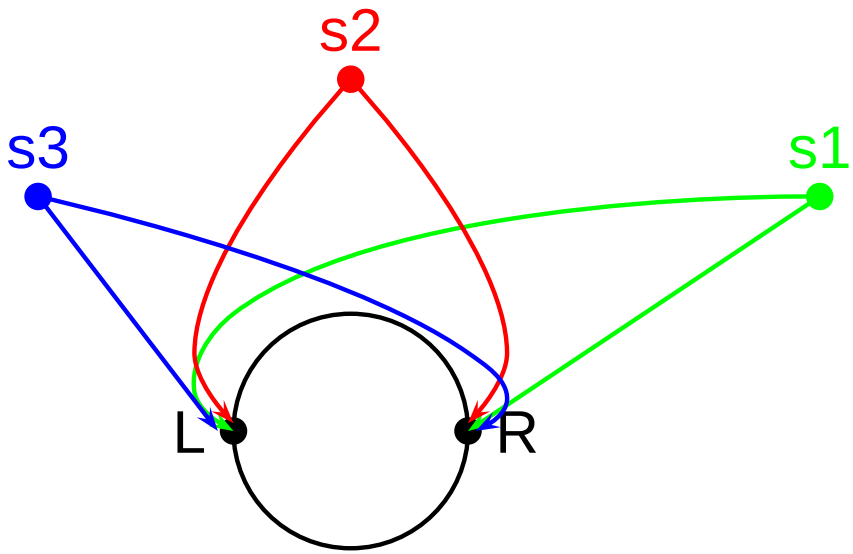
Hypothesis

- Sources do not overlap in the t-f plane
- In practice, speech approximately WDO

Windowed Disjoint Orthogonality [Rickard & Yilmaz (2002)]

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

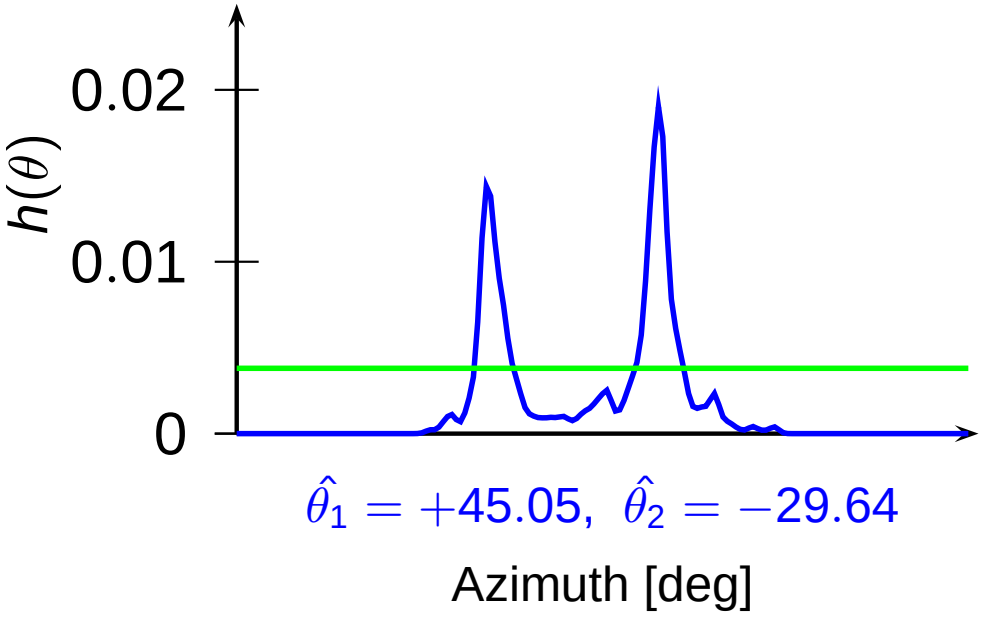




Multi Source Localization

- Finds Local Maxima after thresholding
- Outputs

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



Example

2-source mixture: $+45°$ $-30°$

- K = 8, 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*

$$f_K(\Theta|\Gamma) = \sum_{\theta=\theta_1}^{\theta_N} \sum_{j=1}^{K} \pi_j \ \phi_j(\theta|\gamma_j)^{h(\theta)} \quad \textit{with} \quad \sum_{j=1}^{K} \pi_j = 1$$

Find Γ that best match the data:

Maximum Likelihood-Expectation Maximization

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

EM Updates

EM Updates

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

Demixing 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

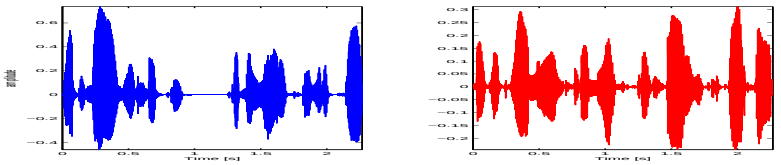
$$S_L(t,f) = M_k(t,f) \cdot X_L(t,f)$$
$$S_R(t,f) = M_k(t,f) \cdot X_R(t,f)$$

Source Separation Results

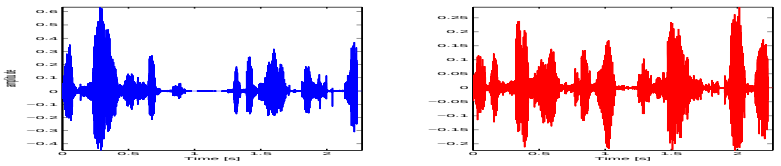
Example: 2-Source Mix ($+45°$ & $-30°$)

	mix_L	mix_R
Mix	●	●
Ori	ori_L ●	ori_R ●
post. Demix	$u1_L$ ●	$u2_R$ ●
ML Demix	$u1_L$ ●	$u2_R$ ●

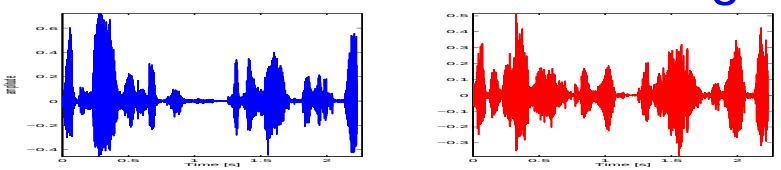
original Waveforms



posterior Demixing



maximum Likelihood Demixing



The posterior demix has less interferences than ML demix

The posterior demix has as much musical noise than ML demix

The posterior demix Mean Opinion Score = 3 on 5 levels

Summary

Summary

- Binaural localization of several sources
- Parametric binaural spatialization
- Source separation with GMM, spatial filtering, probabilistic mask

Future Works

- Distance localization
- Original source energy recovering
- From binaural towards multi-source and multi-diffusion.

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.

Danke

Danke für Ihre Aufmerksamkeit!