

Young Researchers Seminar 2009

Torino, Italy, 3 to 5 June 2009

Audio and Speech signal processing for security and safety application in public transport

Sodoyer David, Ambellouis Sébastien, Flancquart Amaury



Table of content

- Framework & objectives
- Scene analysis
- Audio enhancement/separation
- Localisation
- Experimentation
- Conclusions
- Perspectives



- **Framework & objectives**
- Scene analysis
- Audio enhancement/separation
- Localisation
- Experimentation
- Conclusions
- Perspectives

Framework & objectives

- **Automatic surveillance system in transport**

Detect

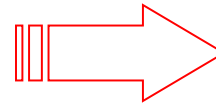
Identify

Localise

abnormal or critical situations

- Context : embedded areas (**Train**, bus or metro)

- Acoustic scene analysis



Audio

enhancement,
separation

- Framework & objectives
- **Scene analysis**
- Audio enhancement/separation
- Localisation
- Experimentation
- Conclusions
- Perspectives



Scene analysis

- Embedded video system drawbacks:



- Variances of luminance
- Limitation of the visual field of the video
- Obstruction by a crowd

Scene analysis

- Embedded video system drawbacks:



- Variances of luminance

- Obstruction by a crowd

- Limitation of the visual field of the video

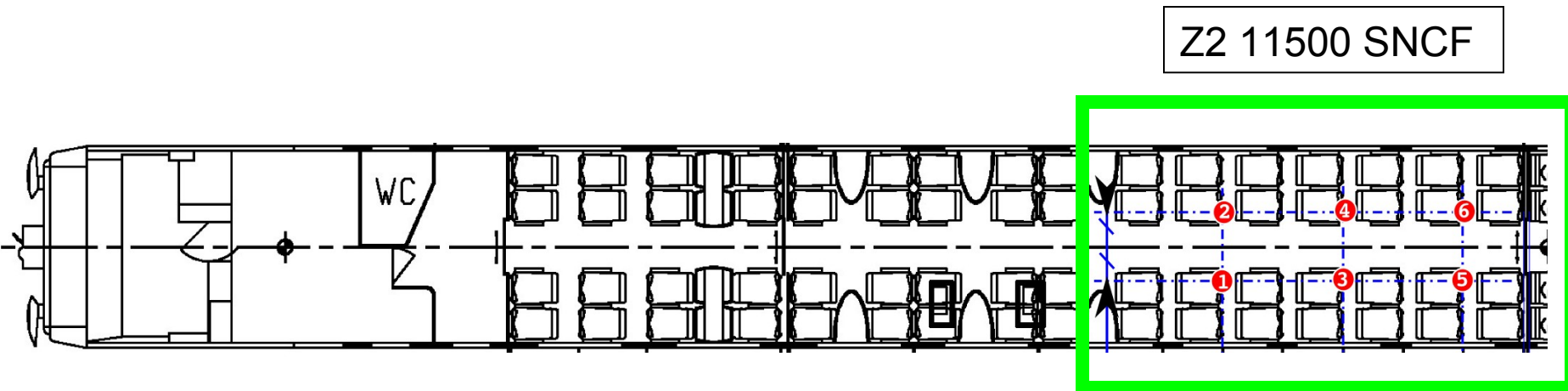
Sound analysis

is independent of the luminance

covers all space of the train coach

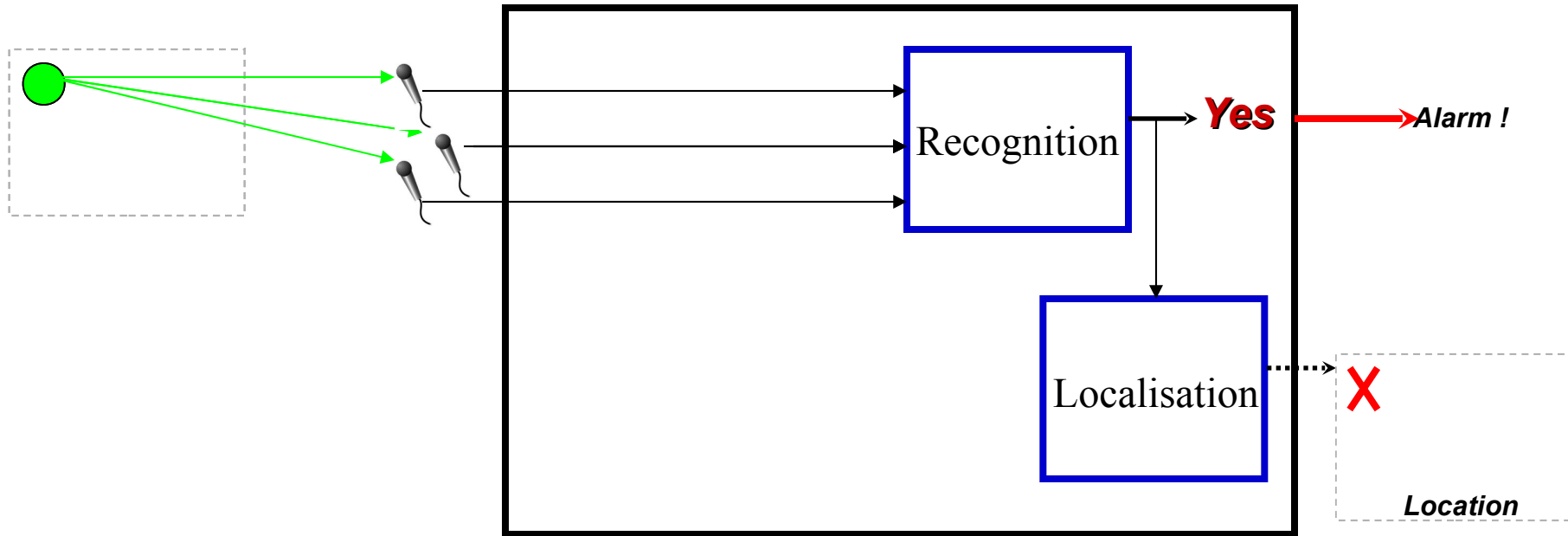
Scene analysis

- Works context



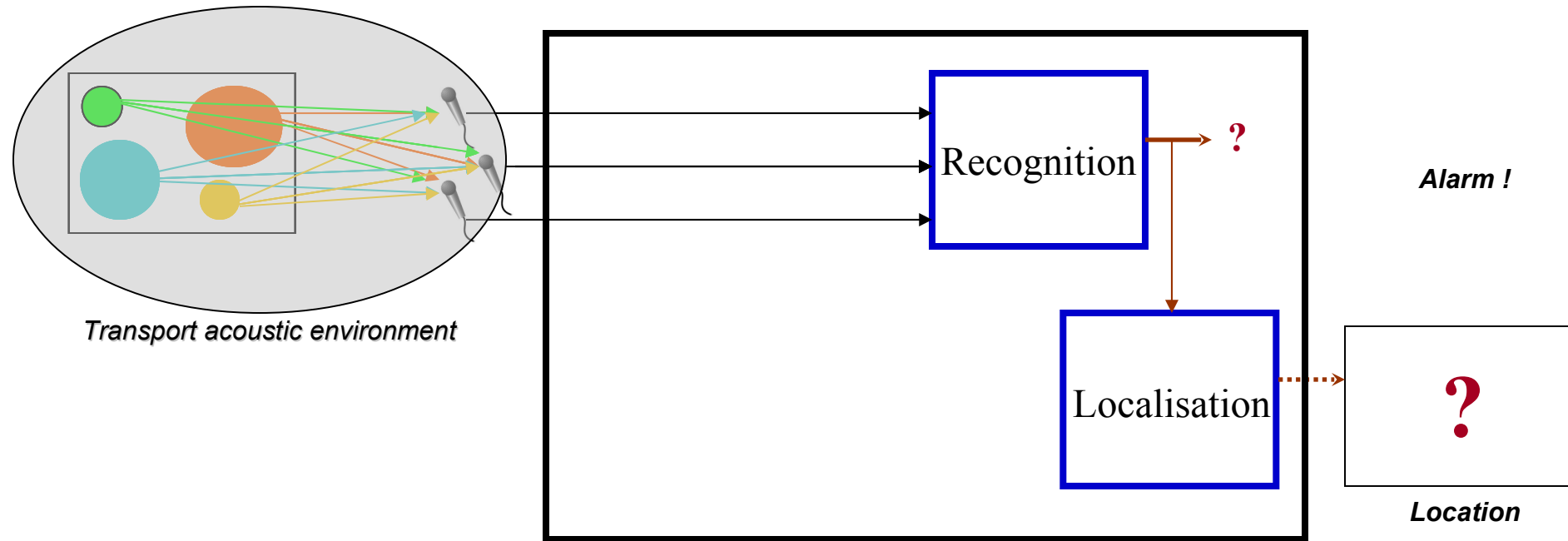
6 microphones

Scene analysis



- Automatic audio scene analysis in low constrained environment
 - Acoustic pattern recognition
 - Location
 - Speech analysis

Scene analysis



• Automatic audio scene analysis in transport environment

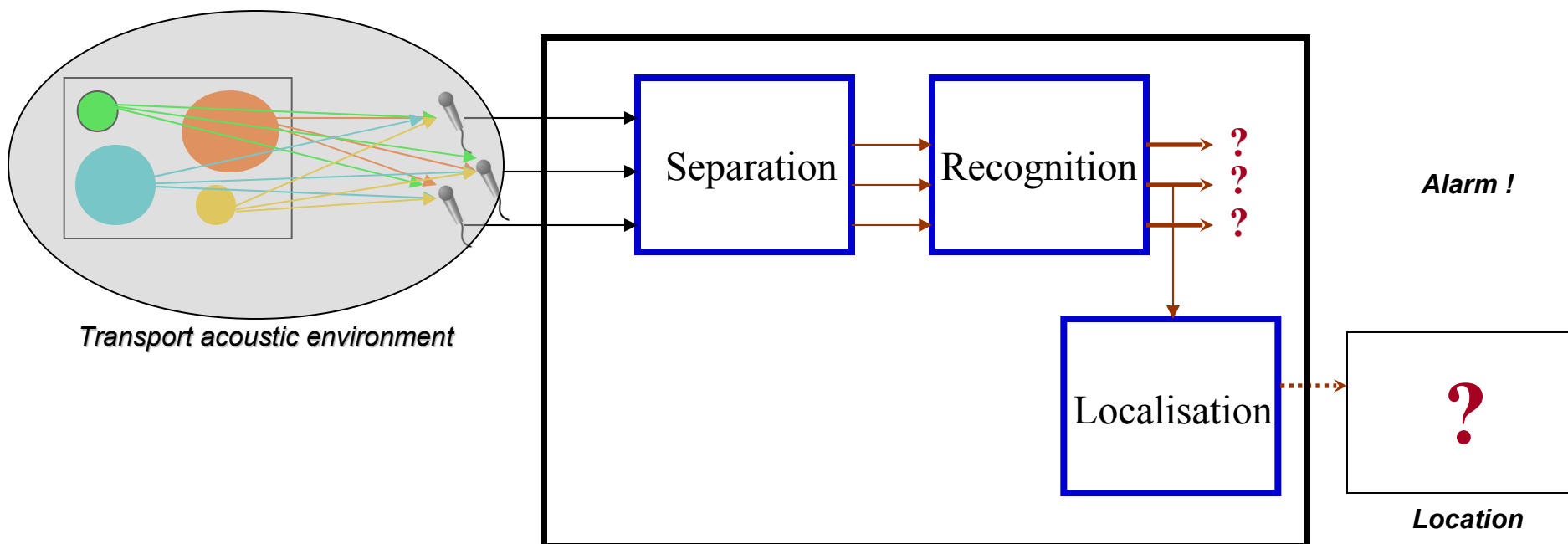
- Several speakers
- Train acoustic noise
- Variable sources number

**Specific and complex
acoustic environment**

- Framework & objectives
- Scene analysis
- **Audio enhancement/separation**
- Localisation
- Experimentation
- Conclusions
- Perspectives



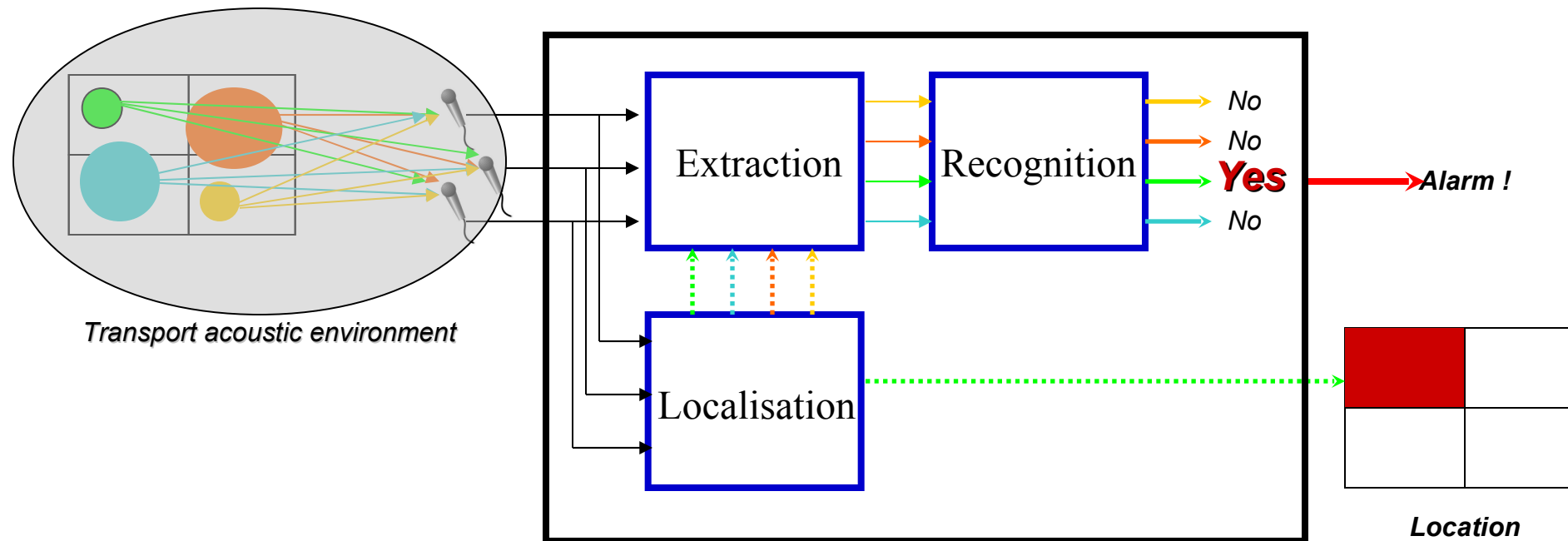
Audio enhancement/separation



- To isolate the sources : a sources separation problem.

- More sources than microphones
- Echoic environment

Audio enhancement/separation



- Localise to better separate/extract

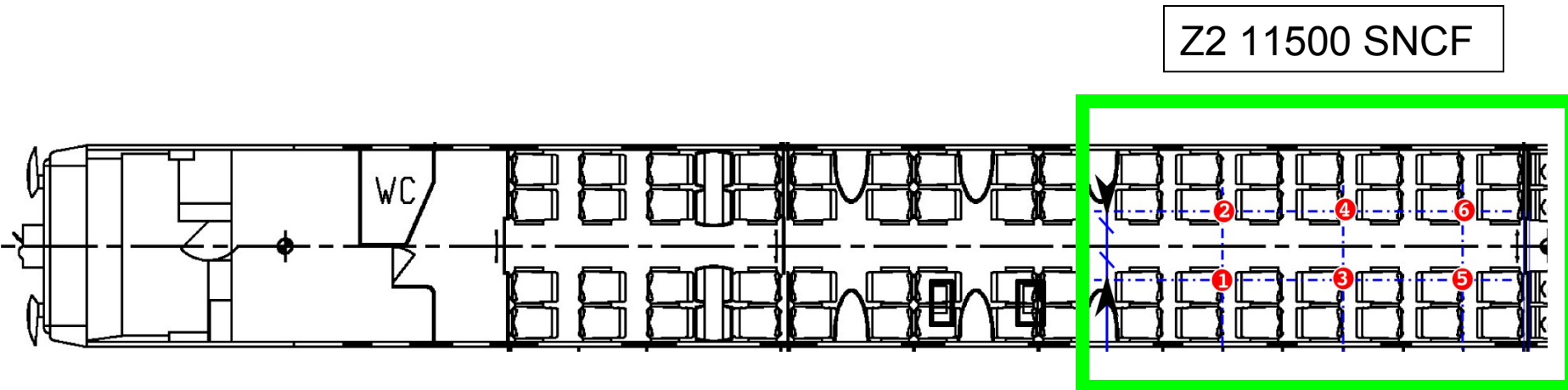
Separate

Enhance

submixtures of acoustic sources **located** at the same area of the train coach

Audio enhancement/separation

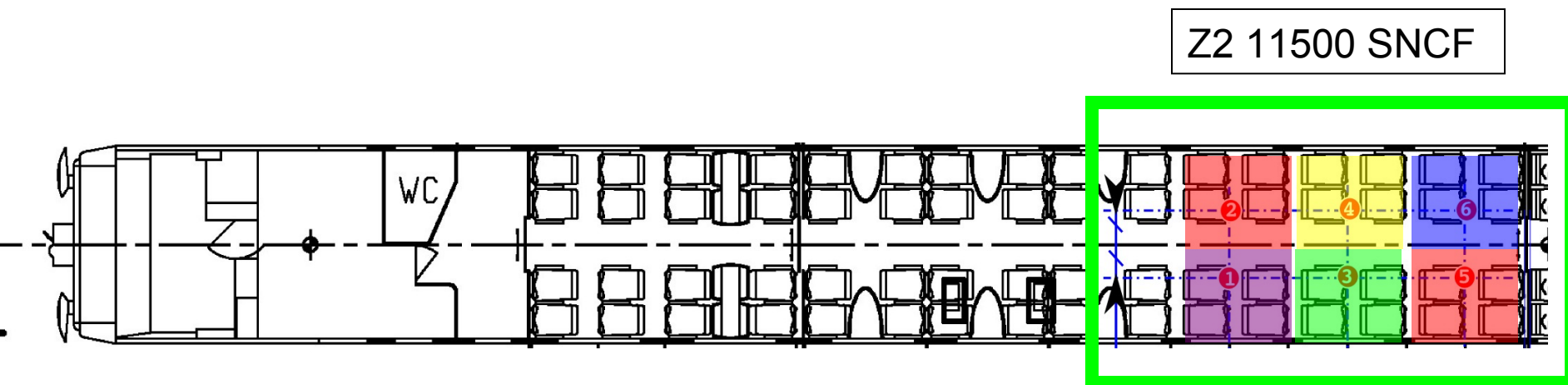
- Works context



6 microphones

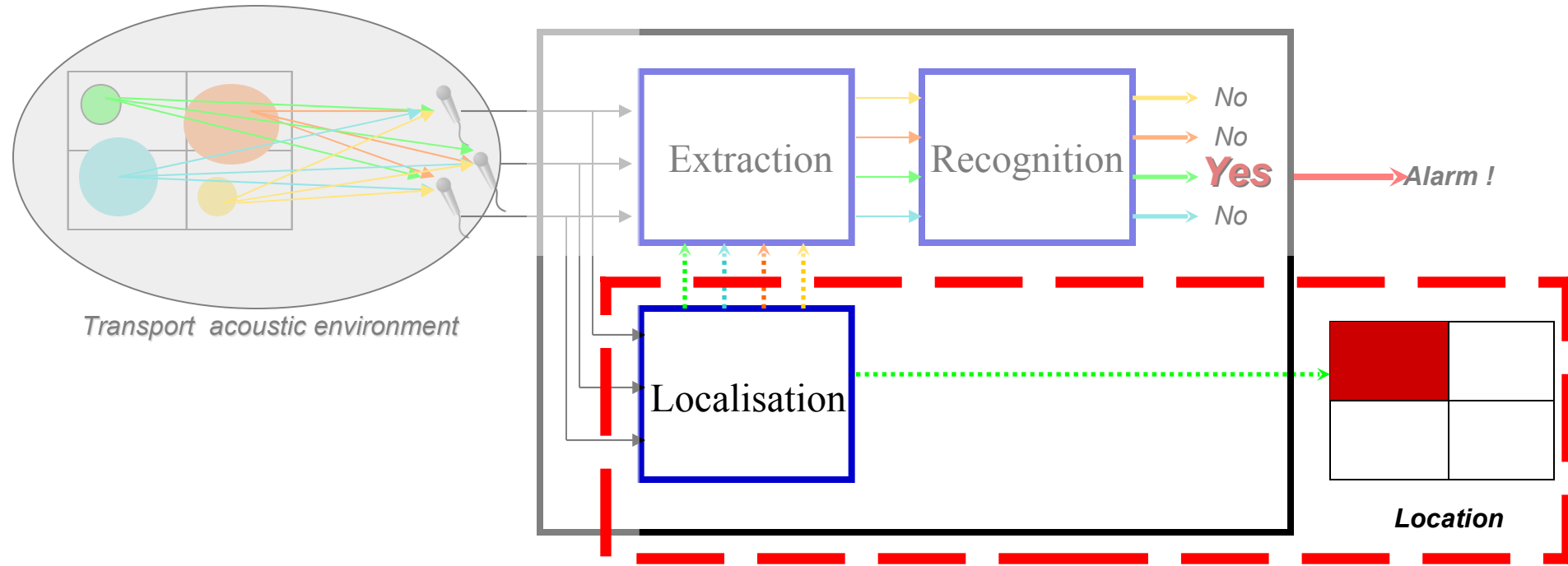
Audio enhancement/separation

- 6 areas have been defined in our context



- 6 microphones,
- focusing on speech sources

Audio enhancement/separation



- First phase:

Localisation system in transport environment

- Framework & objectives
- Scene analysis
- Audio enhancement/separation
- **Localisation** _____
- Experimentation
- Conclusions
- Perspectives

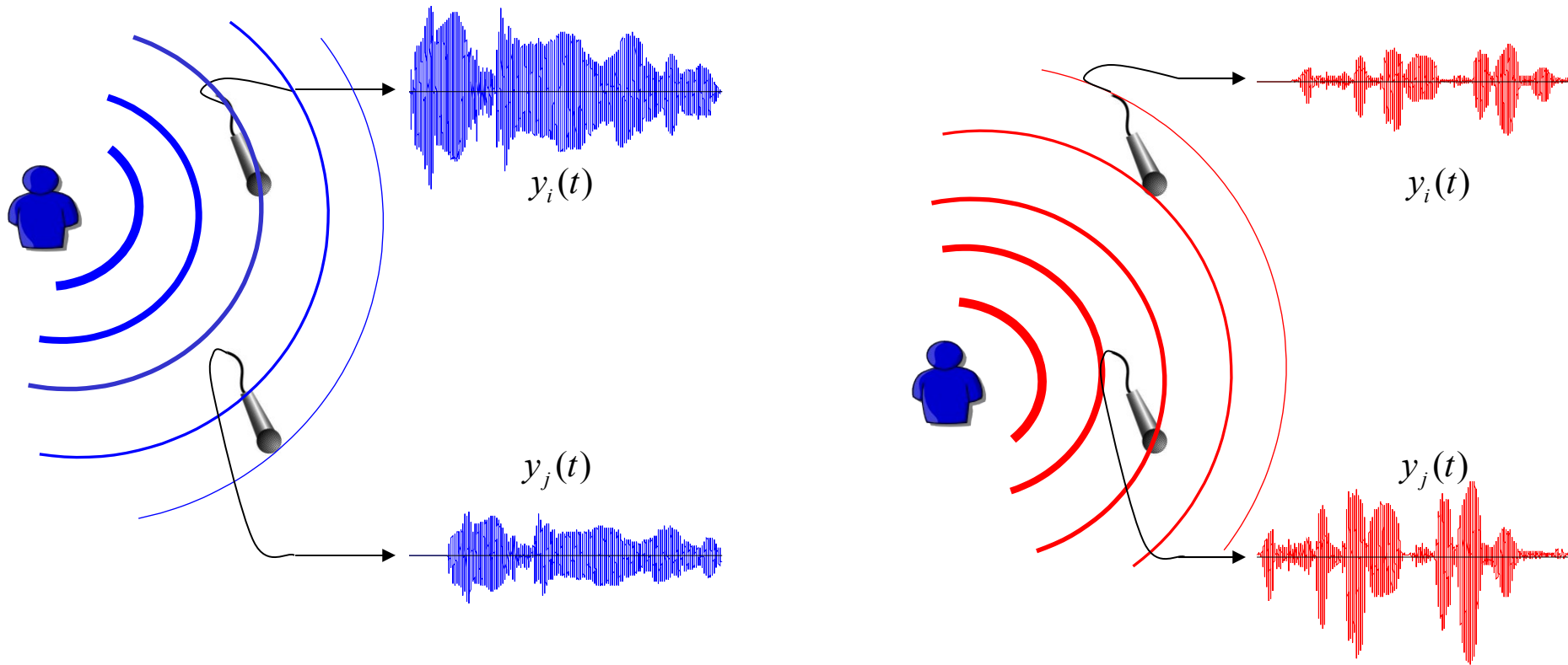
Principles

Position model

Decision process

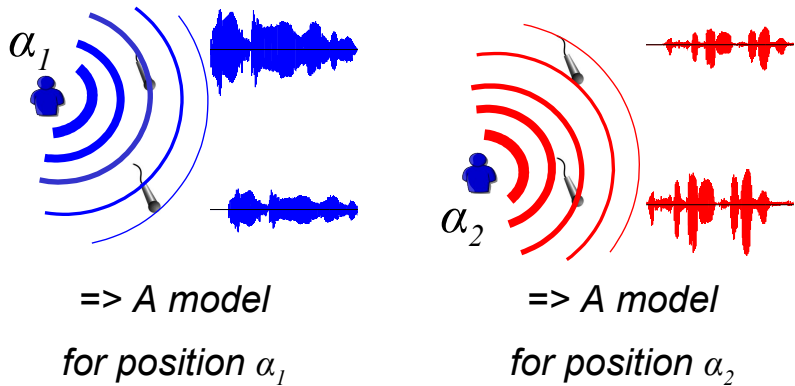


Localisation : Principles



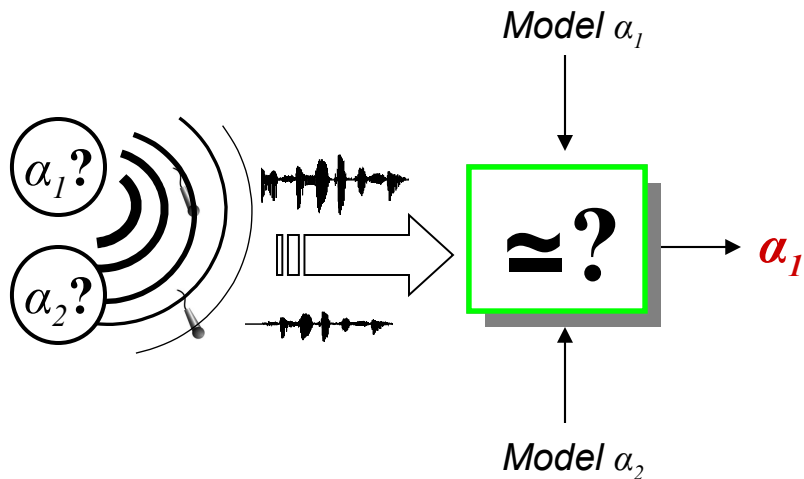
- Audio sources propagation is function of their positions

Localisation : Principles



- Step 1: **Learning**

Learn propagation characteristics for each position thanks to the signal received by each microphone

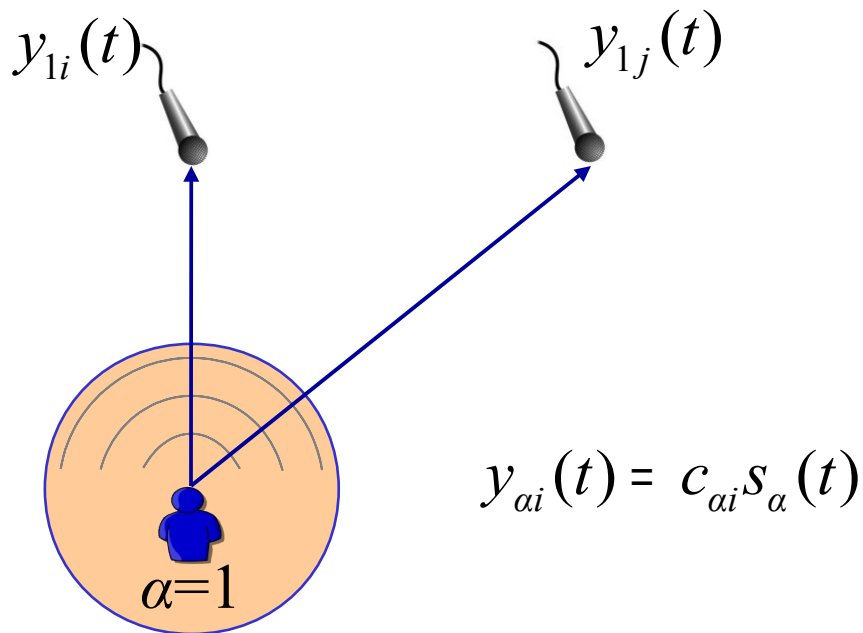


- Step 2: **Localisation**

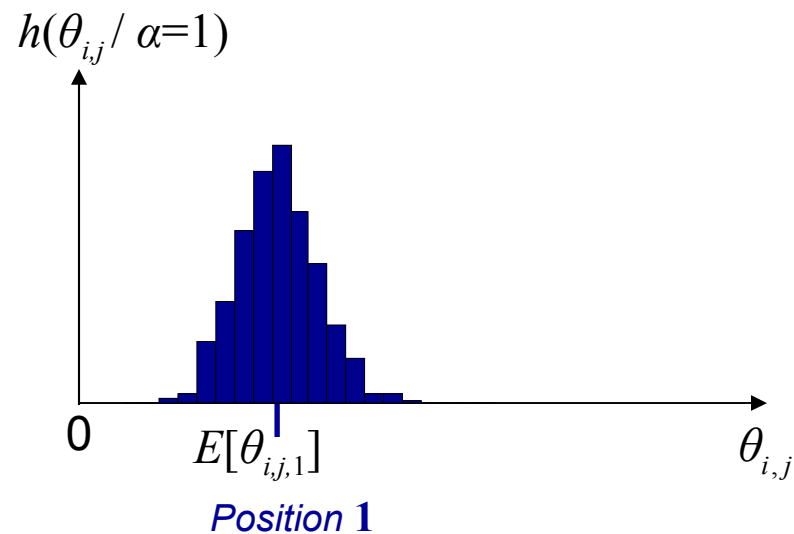
Find the position of an unknown source by checking the « better model »

Localisation : Position model

- Simple case

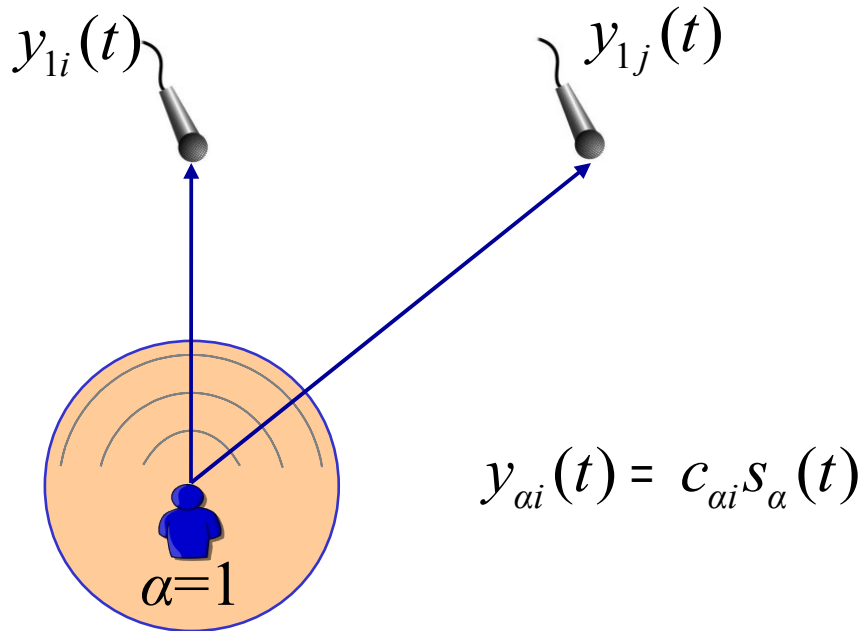


$$\theta_{i,j\alpha}(t) = \frac{c_{ai}}{c_{aj}} \approx \frac{|y_{ai}(t)|}{|y_{aj}(t)|}$$



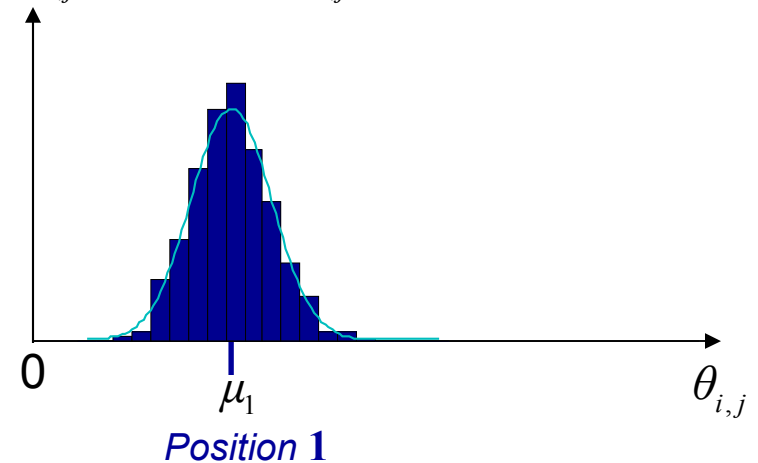
Localisation : Position model

- Simple case



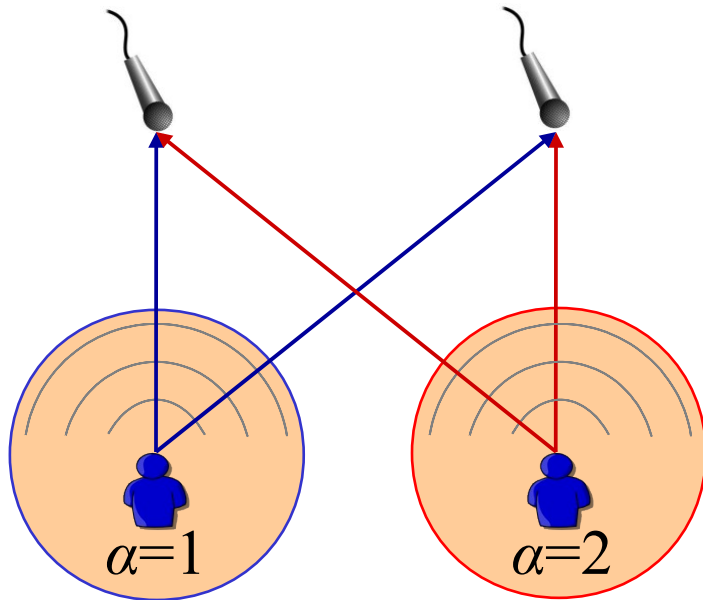
$$\theta_{i,j\alpha}(t) = \frac{c_{ai}}{c_{aj}} \approx \frac{|y_{ai}(t)|}{|y_{aj}(t)|}$$

$$p(\theta_{ij} / \alpha=1) = \mathcal{N}(\theta_{ij}; \mu_1, \sigma_1)$$

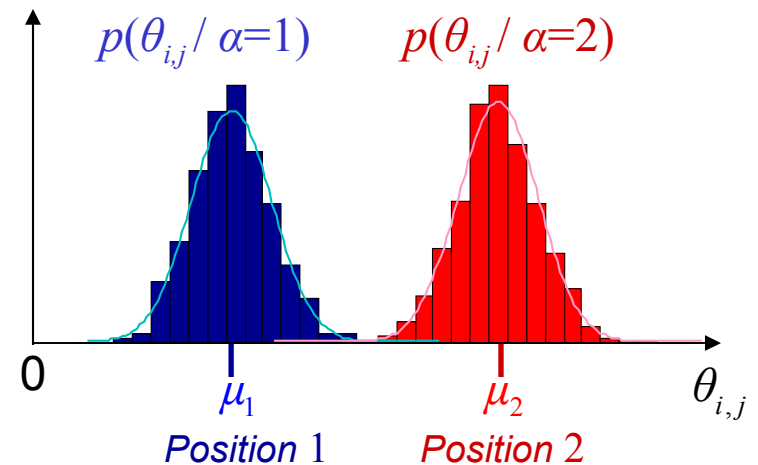


Localisation : Position model

- Simple case multi position

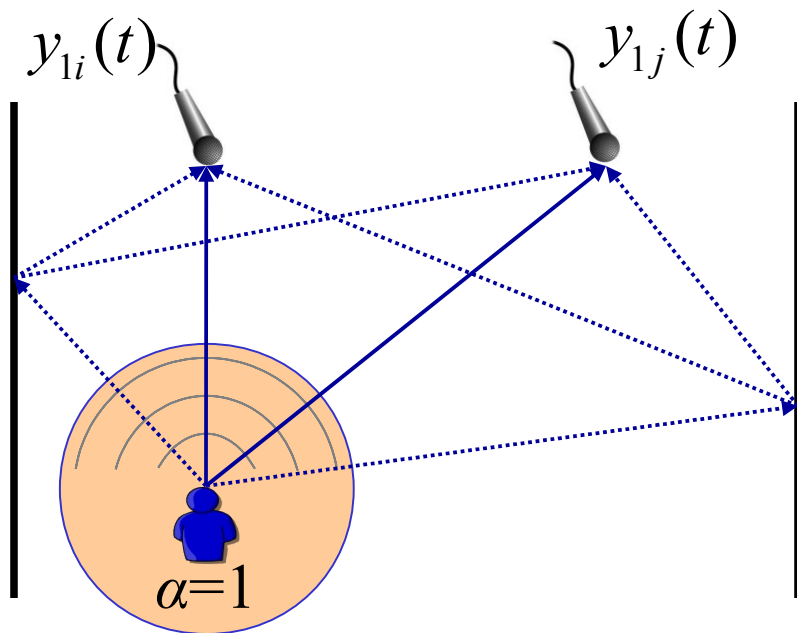


$$\theta_{i,j\alpha}(t) = \frac{c_{ai}}{c_{aj}} \approx \frac{|y_{ai}(t)|}{|y_{aj}(t)|}$$



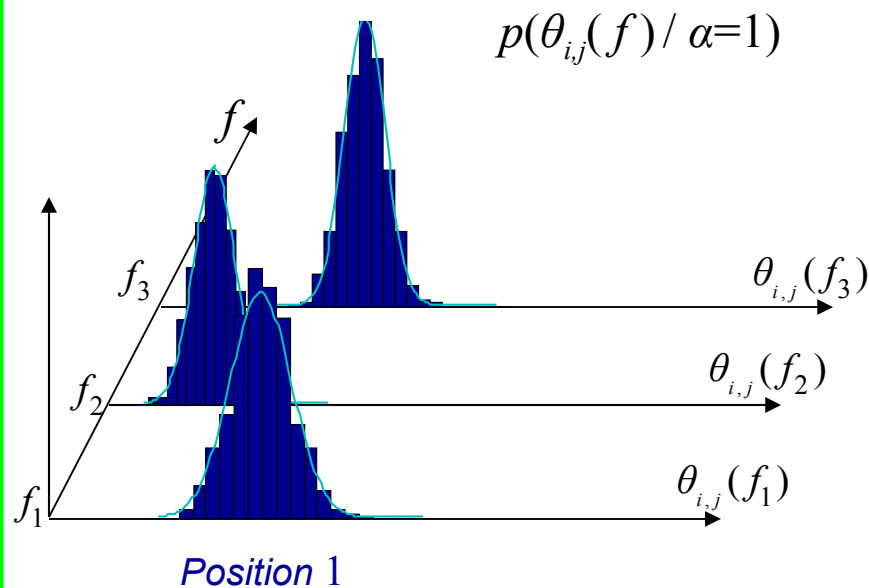
Localisation : Position model

- Reverberant case



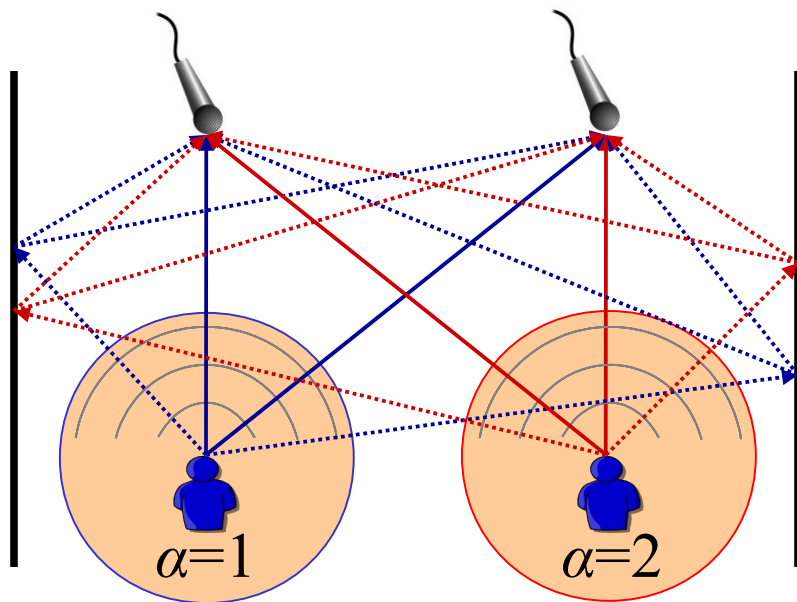
$$y_{\alpha i}(t, f) = c_{\alpha i}(f) s_{\alpha}(t, f)$$

$$\theta_{i,j\alpha}(f) = \frac{c_{\alpha i}(f)}{c_{\alpha j}(f)} \simeq \frac{y_{\alpha i}(t, f)}{y_{\alpha j}(t, f)}$$

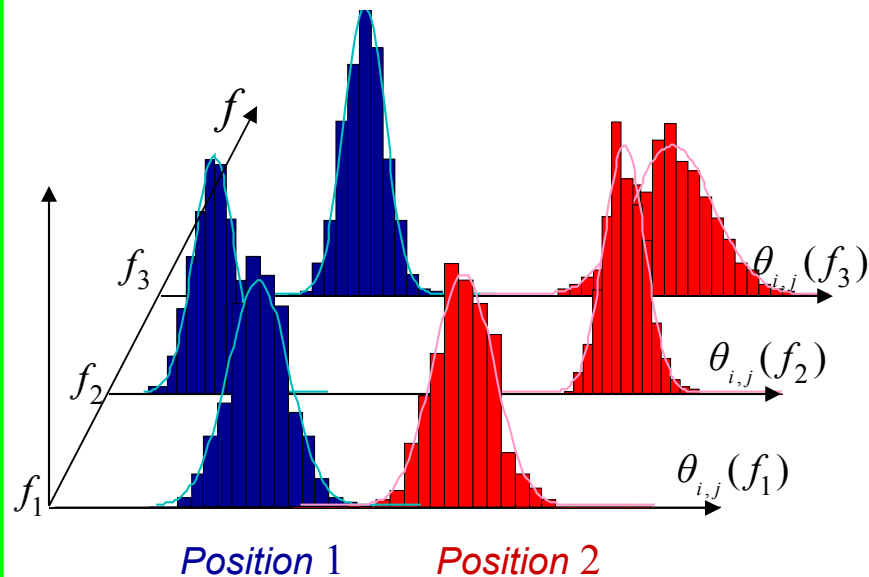


Localisation : Position model

- Reverberant case multi position

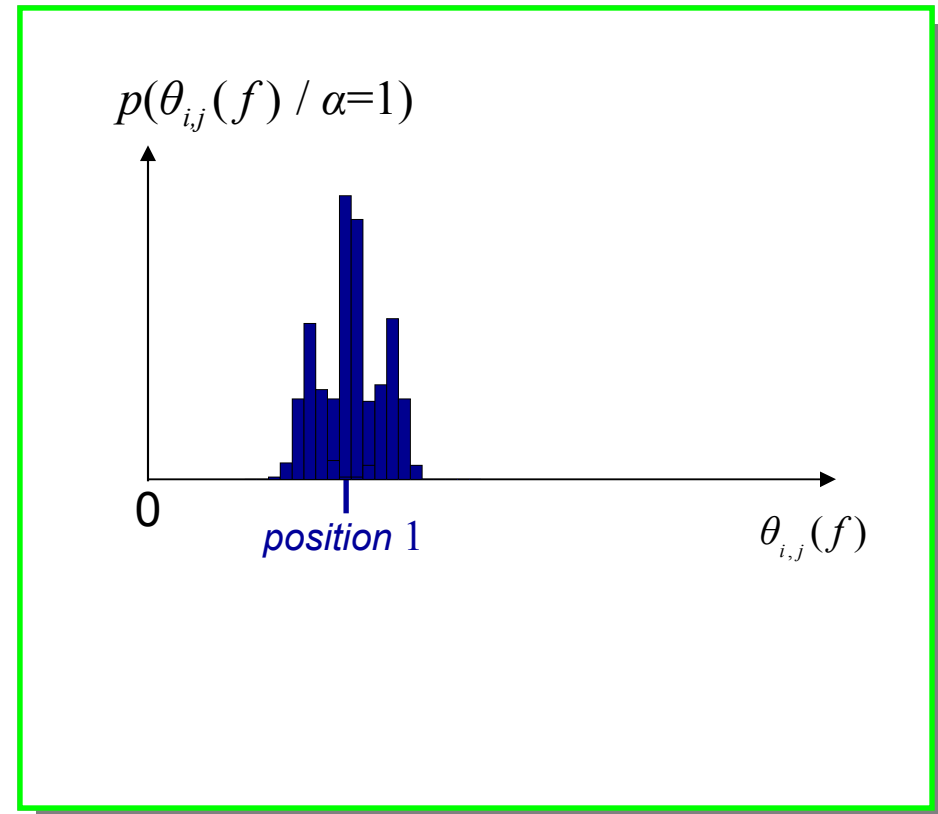
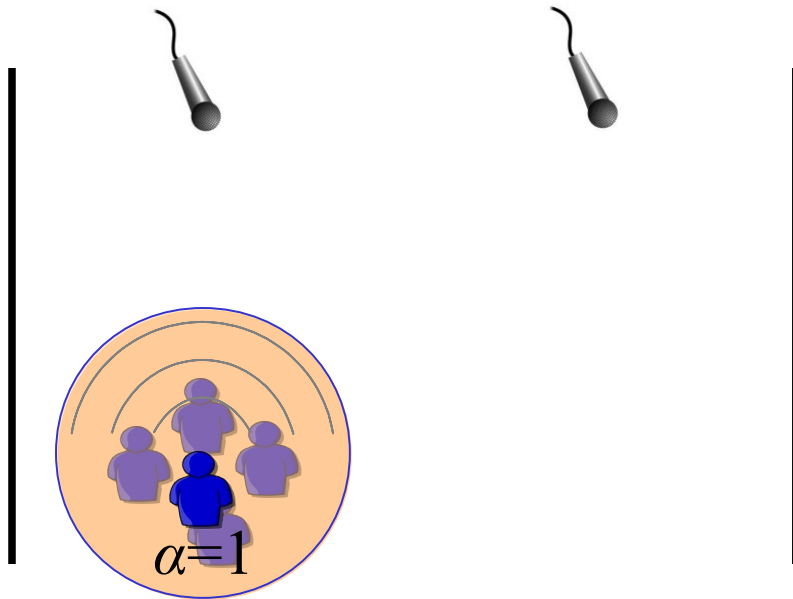


$$\theta_{i,j\alpha}(f) = \frac{c_{\alpha i}(f)}{c_{\alpha j}(f)} \simeq \frac{y_{\alpha i}(t, f)}{y_{\alpha j}(t, f)}$$



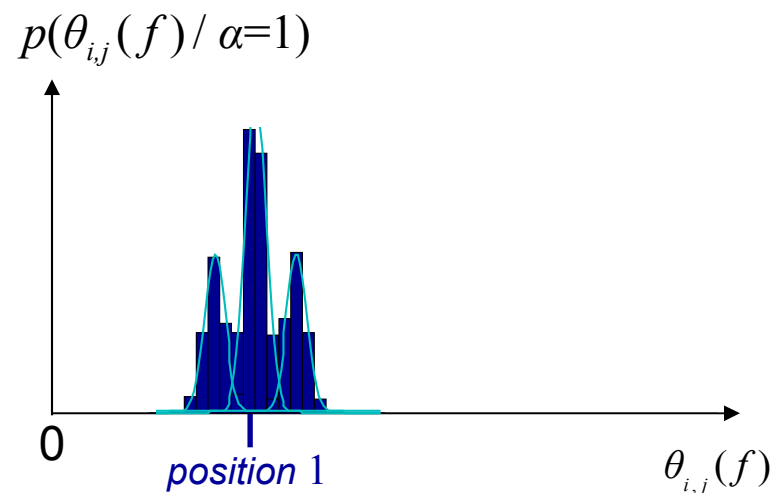
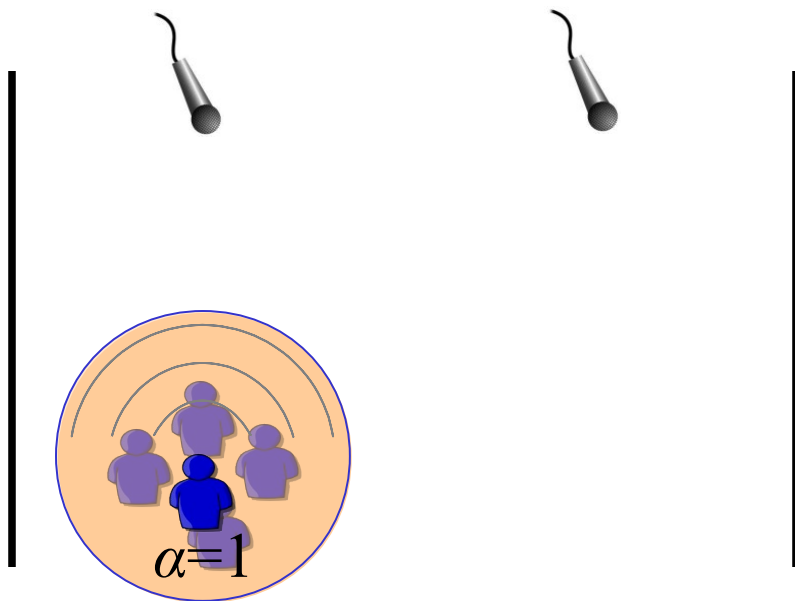
Localisation : Position model

- Variability of the position : non gaussian hypothesis



Localisation : Position model

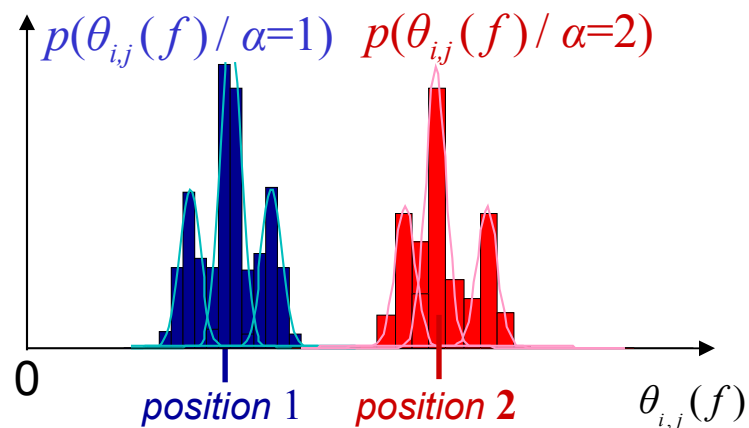
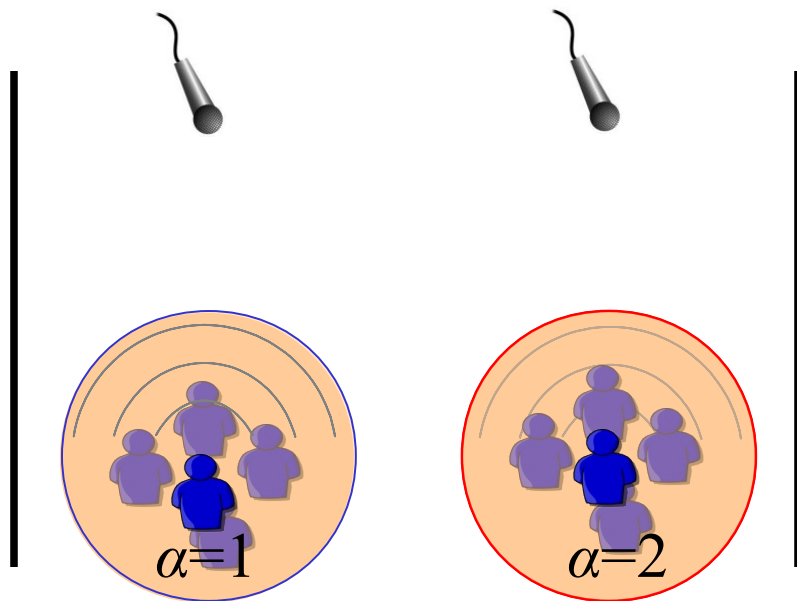
- Variability of the position : multi-gaussian solution



$$p(\theta_{i,j}(f) / \alpha) = \sum_{g=1}^G p(g) \mathcal{N}(\theta_{i,j}(f); \mu_{\alpha_g}, \sigma_{\alpha_g})$$

Localisation : Position model

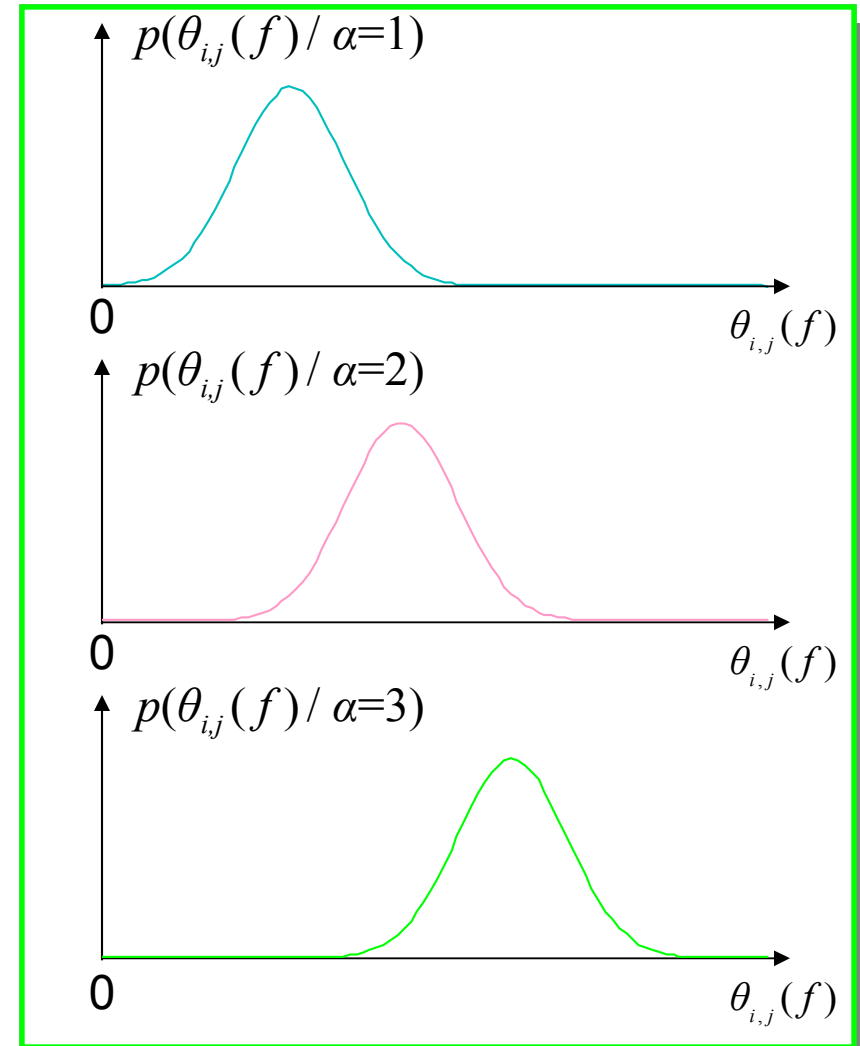
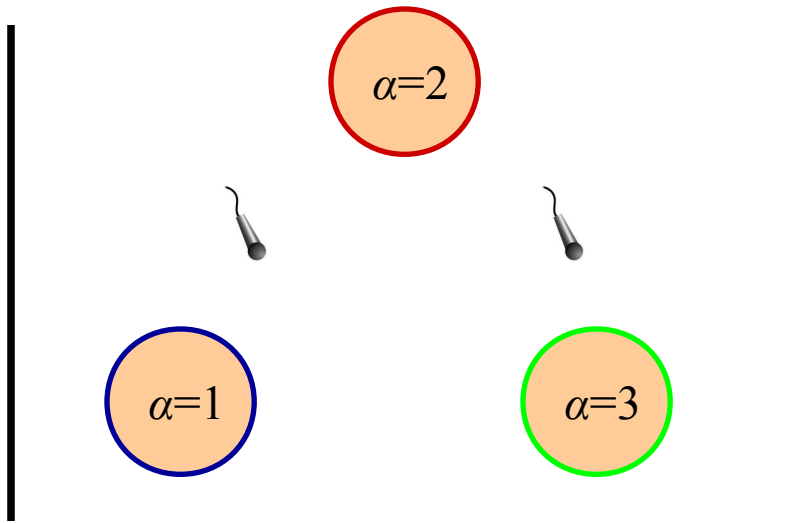
- Variability of the position : multi-gaussian solution



$$p(\theta_{ij}(f) / \alpha) = \sum_{g=1}^G p(g) \mathcal{N}(\theta_{ij}(f); \mu_{\alpha_g}, \sigma_{\alpha_g})$$

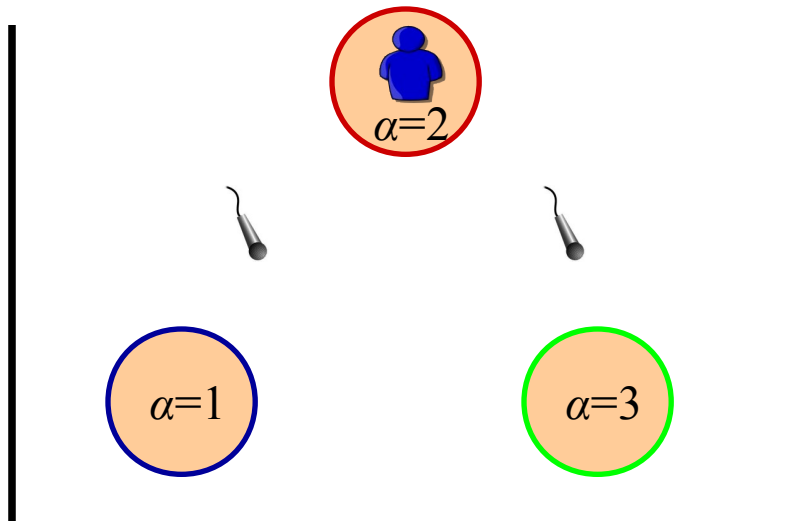
Localisation : Decision process

- Localisation : M.A.P
Maximum a posteriori

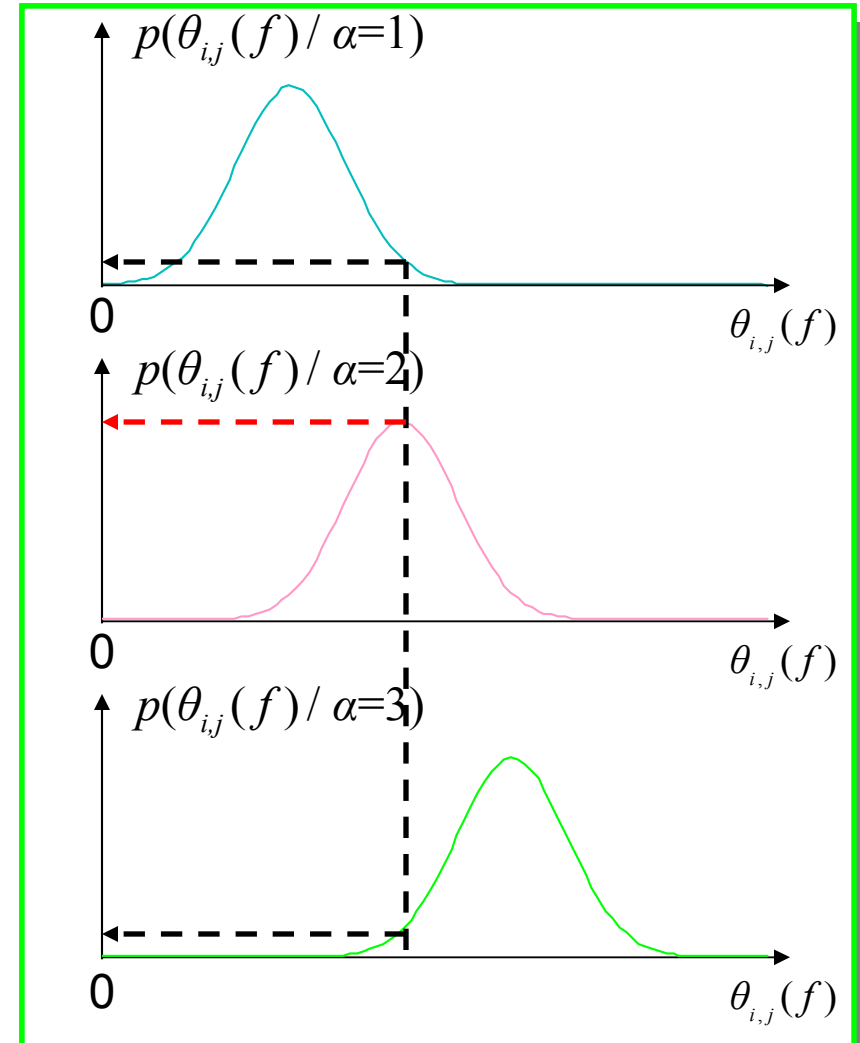


Localisation : Decision process

- Localisation : M.A.P
Maximum a posteriori



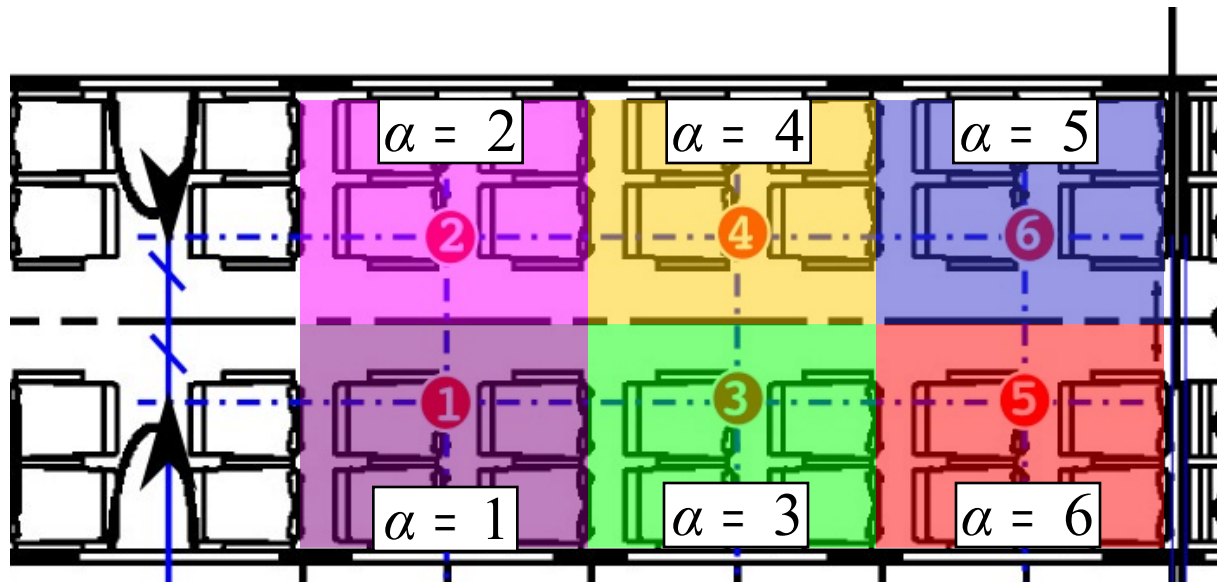
$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [1,2,3]} p(\theta_{i,j}(t, f) / \alpha)$$



- Framework & objectives
- Scene analysis
- Audio enhancement/separation
- Localisation
- **Experimentation** _____
- Conclusions
- Perspectives

Context
Learning step
Test detection
Evaluation
Results

Experimentation : Context



- 6 microphones
- 6 positions $\alpha \in [1 \dots 6]$ « centered » under each microphone

Experimentation : Learning step

- Speaker at each position turning on himself many times
- Data repartition randomly selected
 - 2/3 of data set for the learning step
 - 1/3 of data set for the test step
- Phase of manual labelling
- Learning model with the *E.M.* algorithm (*Expectation Maximisation*)
 - 3 Gaussians per position model
 - Max frequency used $F = 8000$ Hz (speech)
 - Frequency sample $F_s = 48000$ Hz
 - Estimation at every $t = 10$ ms

Experimentation : Test detection

- Decision made with several couples of microphones:

$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{c=1}^{N_c} p(\theta_{i_c, j_c}(t, f) / \alpha) \quad i, j \in [1 \dots 6], i \neq j$$

Experimentation : Test detection

- Decision made with several couples of microphones:

$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{c=1}^{N_c} p(\theta_{i_c, j_c}(t, f) / \alpha) \quad i, j \in [1 \dots 6], i \neq j$$

- Decision made on all frequencies

$$\hat{\alpha}(t) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{f=1}^F \prod_{c=1}^{N_c} p(\theta_{i_c, j_c}(t, f) / \alpha)$$

Experimentation : Test detection

- Decision made with several couples of microphones:

$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{c=1}^{N_c} p\left(\theta_{i_c, j_c}(t, f) / \alpha\right) \quad i, j \in [1 \dots 6], i \neq j$$

- Decision made on all frequencies

$$\hat{\alpha}(t) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{f=1}^F \prod_{c=1}^{N_c} p\left(\theta_{i_c, j_c}(t, f) / \alpha\right)$$

- Decision made on several consecutive time frames

$$\hat{\alpha}(t) = \arg \max_{\alpha \in [1 \dots 6]} \prod_{n=0}^{T-1} \prod_{f=1}^F \prod_{c=1}^{N_c} p\left(\theta_{i_c, j_c}(t-n, f) / \alpha\right)$$

Experimentation : Evaluation

- Performance evaluation of the test step

Comparing the position labels estimated with the reference positions labelled manually.

Rate of the real position $\alpha = n$ estimated as a position $\hat{\alpha} = m$

$$R_{\alpha=n}(\hat{\alpha} = m) = 100 \frac{\text{Number of positions } \hat{\alpha} = m}{\text{Total number of tested positions } \alpha = n}$$

Experimentation : Results

1 frame - 1 couple of micros

$$\theta_{3,4}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	38.5	26.2	0.00	0.00	13.9	21.5
α_2	22.1	43.0	0.00	0.00	19.8	15.1
α_3	0.00	0.00	79.0	0.00	19.4	1.61
α_4	1.75	1.75	0.00	84.2	5.26	7.02
α_5	12.0	11.0	0.00	0.00	55.5	21.5
α_6	22.4	10.5	0.00	0.00	20.9	46.3

1 frame - **6 couples of micros**

$$\theta_{1,2}(t, f) \quad \theta_{5,6}(t, f)$$

$$\theta_{1,3}(t, f) \quad \theta_{2,4}(t, f)$$

$$\theta_{3,5}(t, f) \quad \theta_{4,6}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	50.8	32.3	0.00	0.00	7.69	9.23
α_2	1.08	96.8	0.00	0.00	1.62	0.54
α_3	0.00	0.00	83.9	0.00	16.1	0.00
α_4	0.00	0.00	0.00	94.7	0.00	5.26
α_5	0.00	0.00	0.00	0.00	99.5	0.50
α_6	12.0	7.46	0.00	0.00	22.4	58.2

$$\alpha_p \rightarrow \alpha = p$$

Experimentation : Results

1 frame - 1 couple of micros

$$\theta_{3,4}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	38.5	26.2	0.00	0.00	13.9	21.5
α_2	22.1	43.0	0.00	0.00	19.8	15.1
α_3	0.00	0.00	79.0	0.00	19.4	1.61
α_4	1.75	1.75	0.00	84.2	5.26	7.02
α_5	12.0	11.0	0.00	0.00	55.5	21.5
α_6	22.4	10.5	0.00	0.00	20.9	46.3

5 frames - 1 couple of micros

$$\theta_{3,4}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	98.5	0.00	0.00	0.00	0.00	1.54
α_2	0.00	100	0.00	0.00	0.00	0.00
α_3	0.00	0.00	90.3	6.45	0.00	3.23
α_4	0.00	0.00	3.51	96.5	0.00	0.00
α_5	0.00	0.00	0.00	0.00	100	0.00
α_6	1.49	0.00	0.00	0.00	11.9	86.6

Experimentation : Results

1 frame - 1 couple of micros

$$\theta_{3,4}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	38.5	26.2	0.00	0.00	13.9	21.5
α_2	22.1	43.0	0.00	0.00	19.8	15.1
α_3	0.00	0.00	79.0	0.00	19.4	1.61
α_4	1.75	1.75	0.00	84.2	5.26	7.02
α_5	12.0	11.0	0.00	0.00	55.5	21.5
α_6	22.4	10.5	0.00	0.00	20.9	46.3

5 frames - 6 couples of micros

$$\theta_{1,2}(t, f) \quad \theta_{5,6}(t, f)$$

$$\theta_{1,3}(t, f) \quad \theta_{2,4}(t, f)$$

$$\theta_{3,5}(t, f) \quad \theta_{4,6}(t, f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	100	0.00	0.00	0.00	0.00	0.00
α_2	0.00	100	0.00	0.00	0.00	0.00
α_3	0.00	0.00	100	0.00	0.00	0.00
α_4	0.00	0.00	0.00	100	0.00	0.00
α_5	0.00	0.00	0.00	0.00	100	0.00
α_6	0.00	0.00	0.00	0.00	8.96	91.0

- Framework & objectives
- Scene analysis
- Audio enhancement/separation
- Localisation
- Experimentation
- **Conclusions**
- Perspectives

Conclusions

- The localisation results are quite promising particularly when:
 - Decision is made on **several consecutive temporal frames,**
 - Speakers are sensed by **several couples of microphones.**

- Framework & objectives
- Scene analysis
- Audio enhancement/separation
- Localisation
- Experimentation
- Conclusions
- **Perspectives**



Perspectives

- Localise when **several speakers appear**, considering the difference of the time-frequency signatures of the speakers (**temporal-frequency parsimony**).
- To extend this localisation method on **other kind of signals** (not only speech).
- **To use this localisation to separate and enhance** each audio source.

Young Researchers Seminar 2009

Torino, Italy, 3 to 5 June 2009

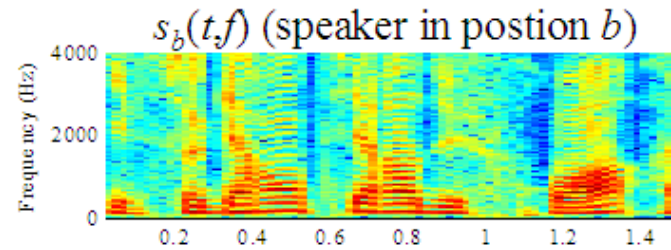
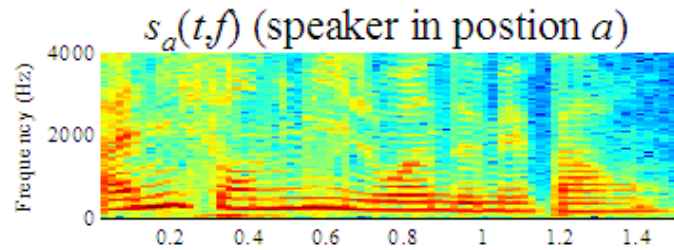
Thank you for your attention

**Audio and Speech signal processing for security
and safety application in public transport**

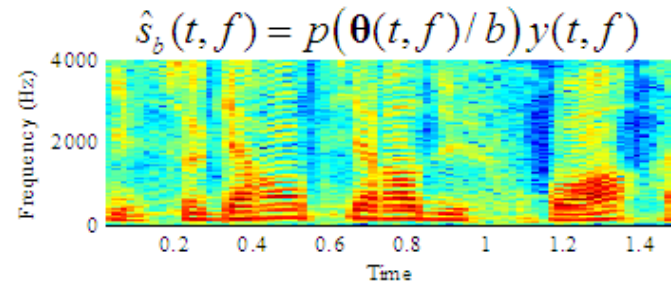
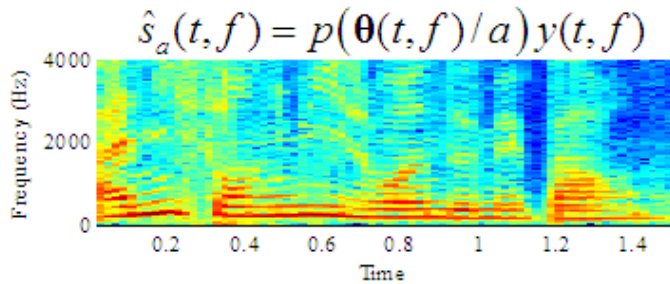
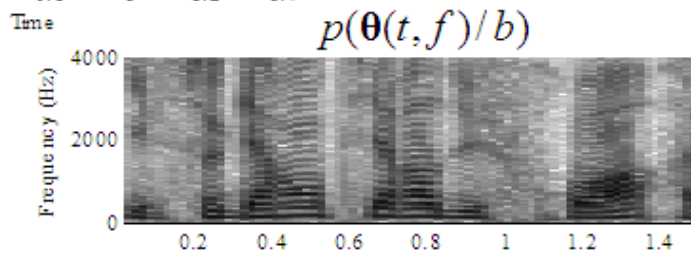
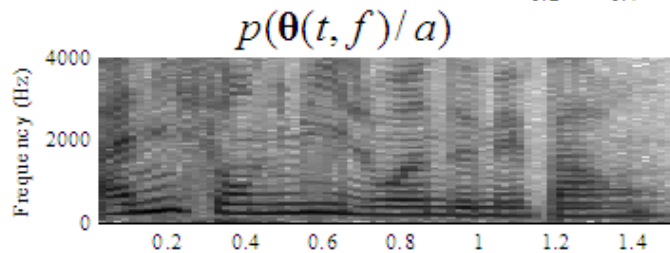
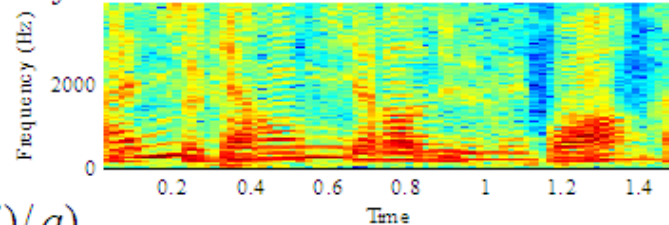
Sodoyer David, Ambellouis Sébastien, Flancquart Amaury



Perspectives



$$y_j(t,f) = s_a(t,f) + s_b(t,f) \quad (\text{Microphone } j)$$



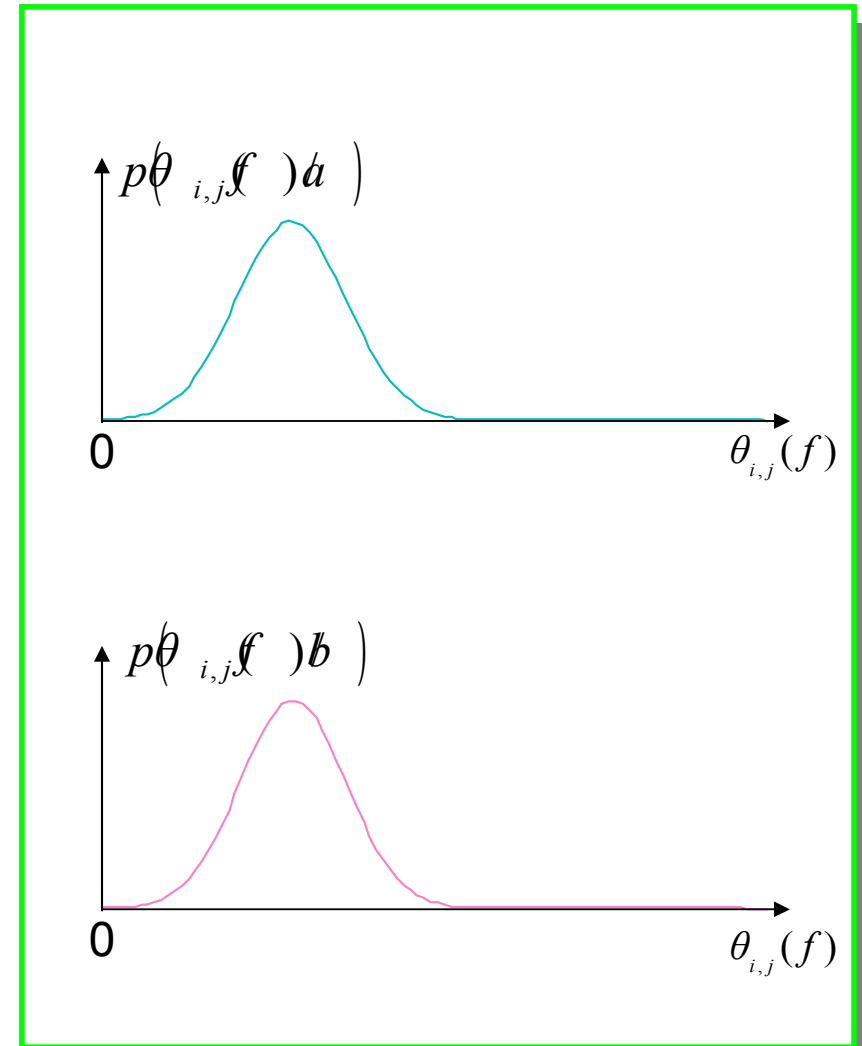
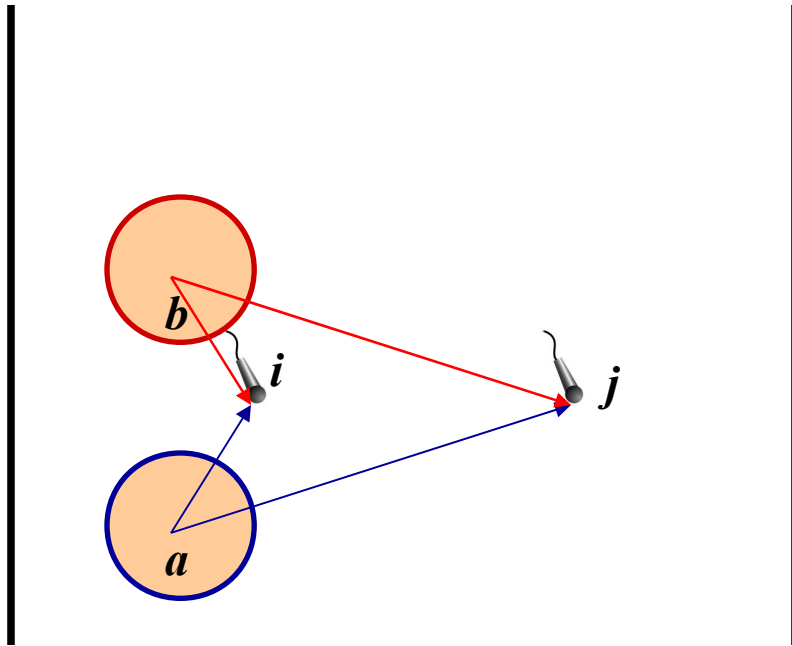
Audio and Speech signal processing for security and safety application in public transport

Sodoyer David, Ambellouis Sébastien, Flancquart Amaury



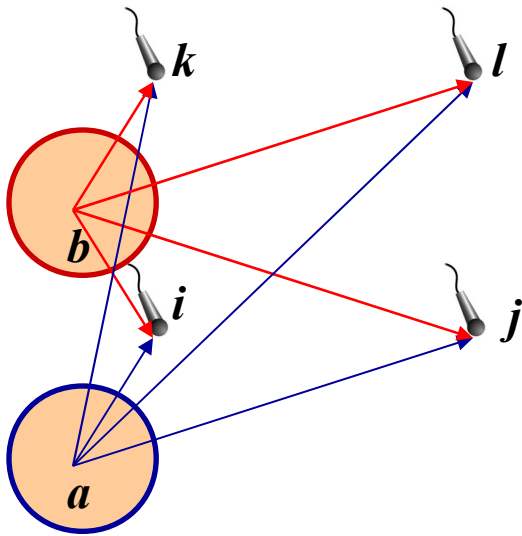
Localisation : Multi even microphones

- Ambiguity

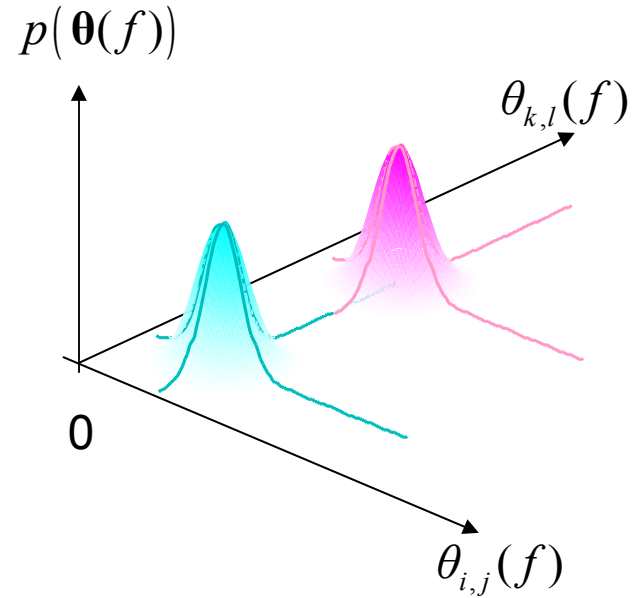


Localisation : Multi even microphones

- Ambiguity



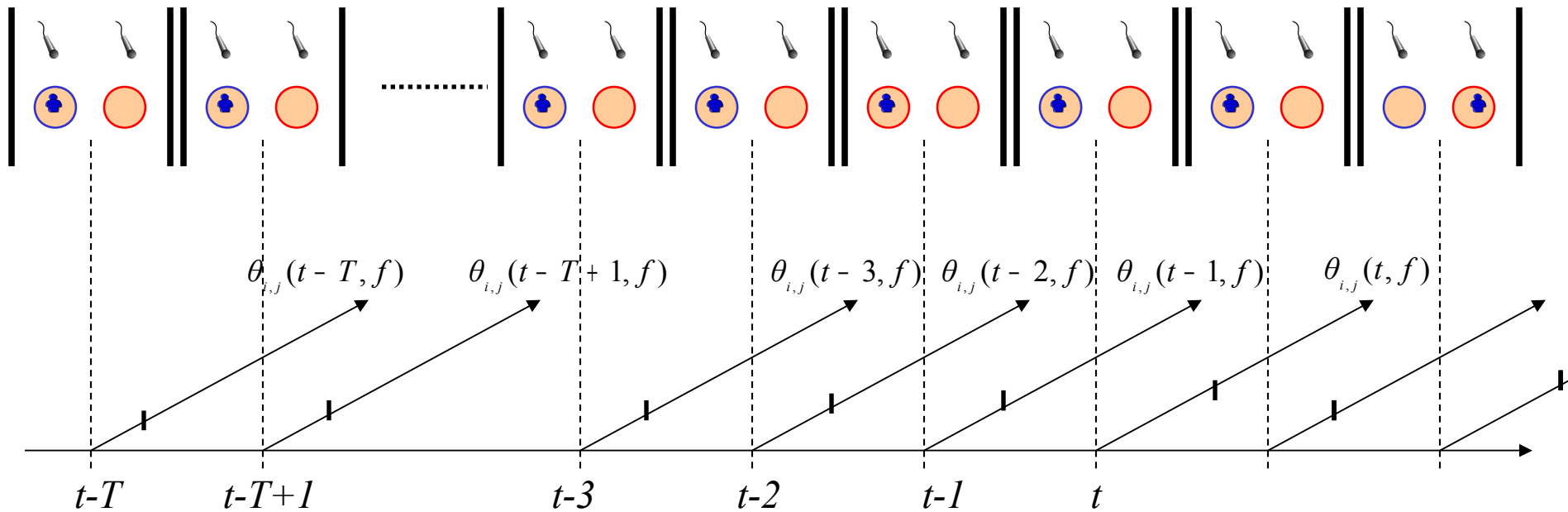
$$\boldsymbol{\theta}(f) = \begin{bmatrix} \theta_{i,j}(f) \\ \theta_{k,l}(f) \end{bmatrix}$$



$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [a,b,c]} p(\boldsymbol{\theta}(t, f) / \alpha)$$

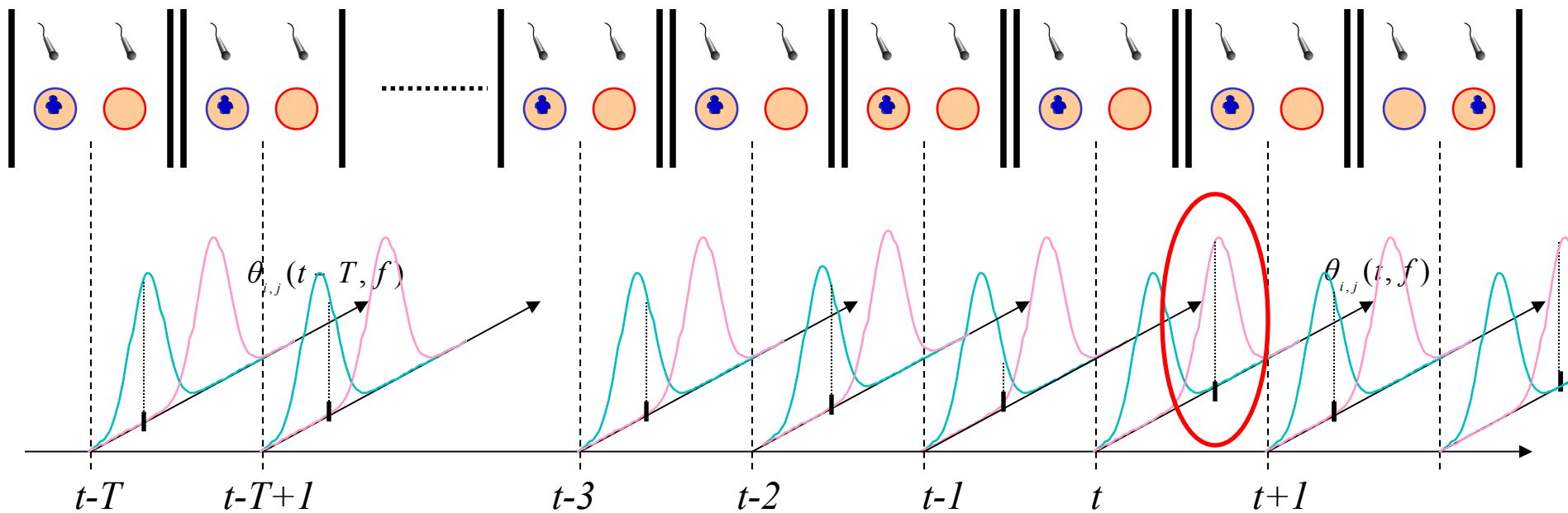
Localisation : Multi frames

- Detection at every time t



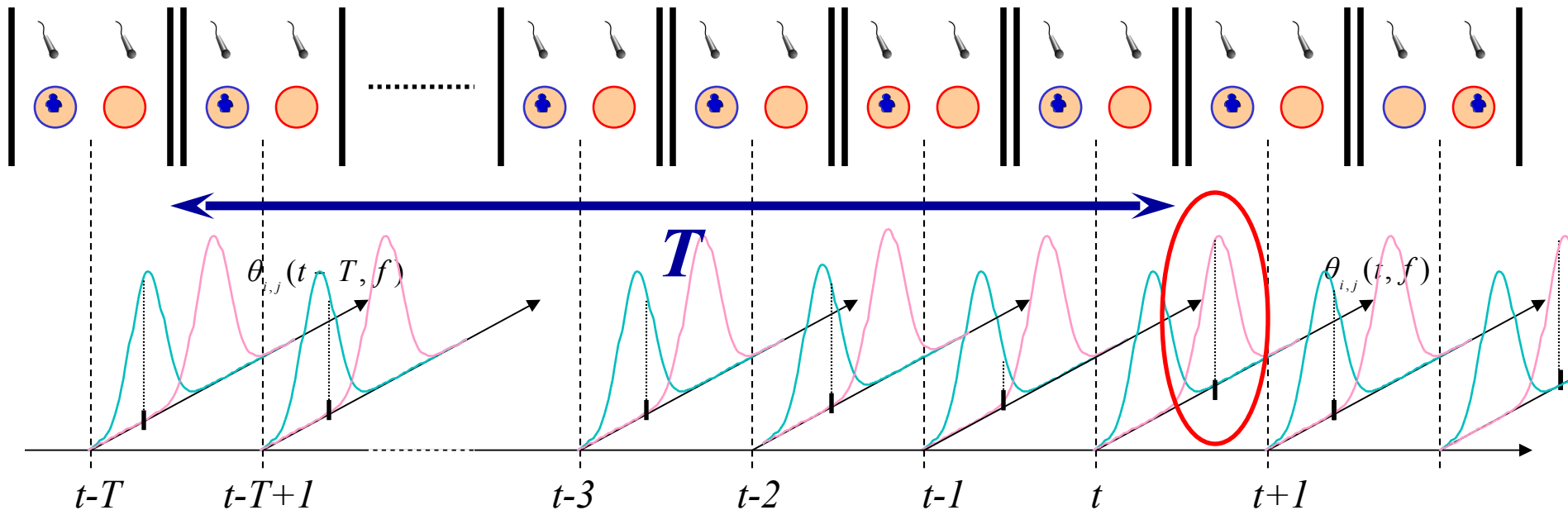
Localisation : Multi frames

- Lack of robustness



Localisation : Multi frames

- Lack of robustness



$$\hat{\alpha}(t, f) = \arg \max_{\alpha \in [a, b, c]} \prod_{n=0}^T p(\boldsymbol{\theta}(t - n, f) / \alpha)$$