Audio – video analysis and case study in a public transport context

Sébastien AMBELLOUIS, David SODOYER
IFSTTAR – LEOST
Sebastien.ambellouis@ifsttar.fr

Human Activity and Vision Summer School, INRIA Sophia Antipolis, october 3rd 2012

Content

Audio and video analysis: a brief state of the art

Case of study: audio and video processing system to improve security in a train coach

Audio and video analysis: a brief state of the art

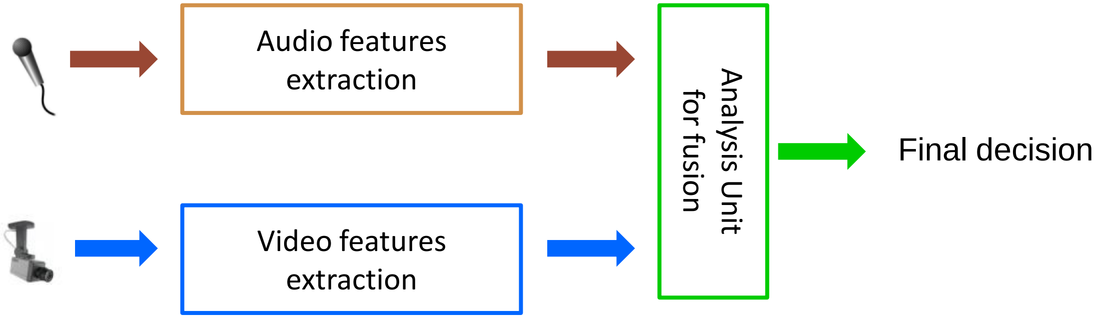
- In the context of the intelligent surveillance, automatic scene analysis and understanding often considered **visual** information.
- The audio modality can be a very interesting source of information in some cases
  - In bad lighting conditions where image processing fails at detecting a mobile object (a mobile emitting some sound);
  - A single image processing unit can fail at understanding a situation (a group of excited people are singing ? or are shouting to threat other people ?);

- P.K. Atrey and al., Multimodal fusion for multimedia analysis : a survey, Multimedia systems, 2010
  - Combine multiple modalities
  - for several tasks

- To fuse modalities at two levels
  - Low level
  - Decision level

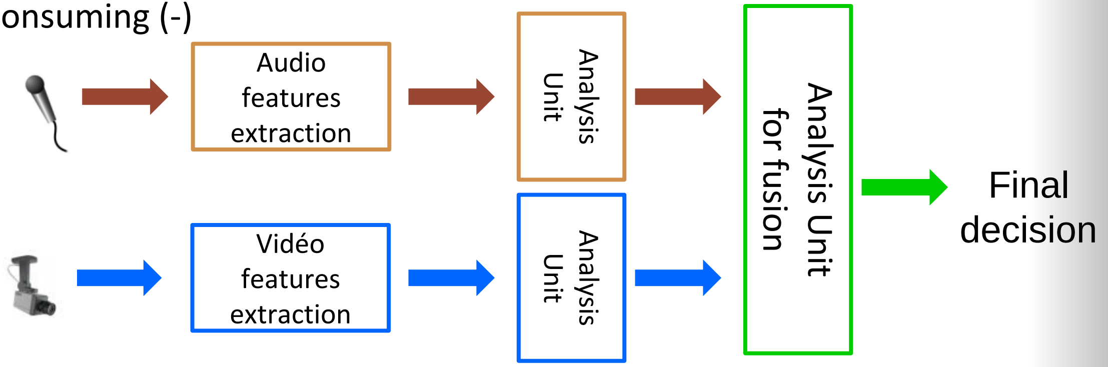
Audio and video analysis: a brief state of the art

- The analysis task is performed directly on the extracted modal features
  - To use high correlated features (+)
  - Features vector can have a high dimension (-)
  - It requires high synchronization between streams (-)



Low level fusion

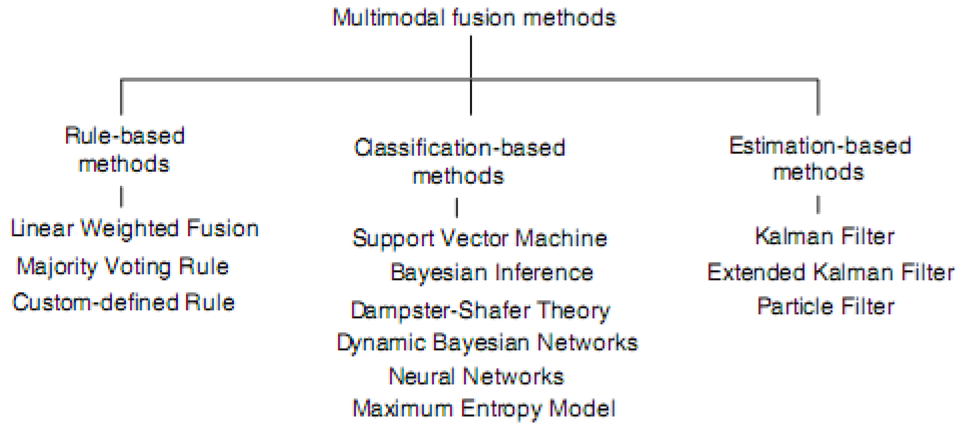
- Local decision are provided by analysing a single stream then local decisions are combined to obtain a final decision
  - Easier to manage multimodal streams (+)
  - It is possible to adapt the analysis method to each type of stream (+)
  - Learning process could be time consuming (-)



Decision level fusion

Audio and video analysis: a brief state of the art

- Fusion unit are based on different methods that can be divided into the 3 categories: Rule-based, classification-based and estimation-based methods.
- Rule-based methods
  - High temporal alignment between modalities is required
  - Linear weighted fusion is the most popular method: face detection, speech and speaker recognition, person identification etc.
- Classification-based methods
  - SVM : good classification performance
  - Dynamic Bayesian Networks : good model temporal data
- Estimation-based methods
  - Adpated for tracking task
  - Kalman flter : good for linear model
  - Particle filter : adapted for non linear and non-Gaussian models



Audio and video analysis: a brief state of the art

- In the following we present 3 papers to illustrate the previous slide contents

Illustration 1: talking head detection

- Dongge Li and al., Multimedia Content Processing through Cross-Modal Association, ACM int. conf. on Multimedia, 2003

Illustration 2: Audio/video synchrony analysis

- M. Cristani and al., Audio-visual Event Recognition in Surveillance Video Sequence, IEEE trans. On Multimedia, Vol. 9, NO. 2, 2007

Illustration 3: hierarchical event detection

- P.K. Atrey and al., Information assimilation framework in multimedia surveillance systems, Multimedia Systems, 2006

Audio and video analysis: correlation analysis (illustration 1)

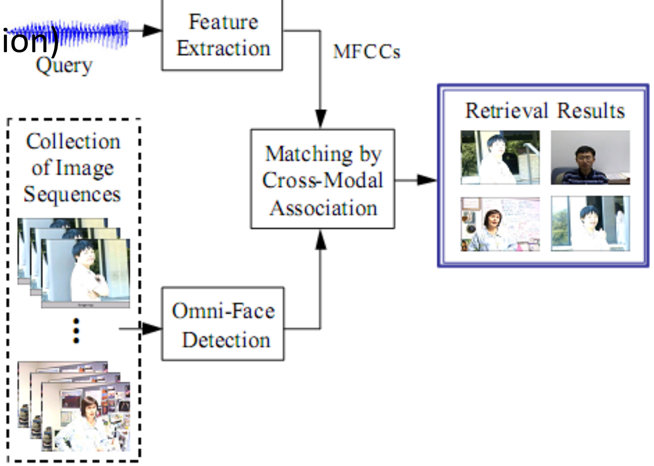
- Talking head detection
  - Audio features: 12 Mel-frequency cepstral coefficients
  - Video features: pixel intensities (or eigenface)
  - A supervised method

- Learning step: example of CFA (Cross-modal Factor Analysis)
  - X is an audio features vector and Y is a video features vector
    Features are extracted from video clip where video and audio streams are synchronized
    X and Y are coupled row-by-row
    Define a subspace where X and Y are closed to each other
    Learning step aims at computing the matrices A and B by minimizing

$$\|XA - YB\|_F^2 \quad \text{where} \quad \|M\|_F = \left(\sum_i \sum_j |m_{ij}|^2\right)^{1/2}$$

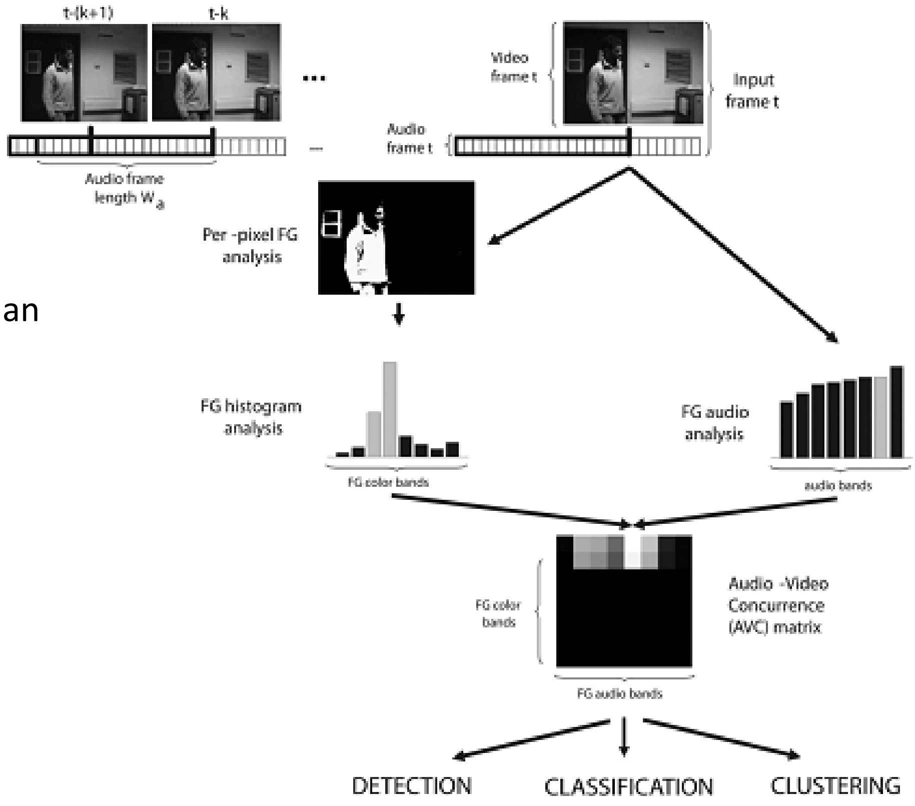
Frobenius norm

- Evaluation step
  - It is performed after applying the transformation matrices A and B to the features
  - The goal is to find the images (among a image sequence collection) related to a audio signal (the query)
  - Matches are evaluated by using Correlation Coefficient in the learned subspace
  - A face detection is applied to reduce the matching candidates

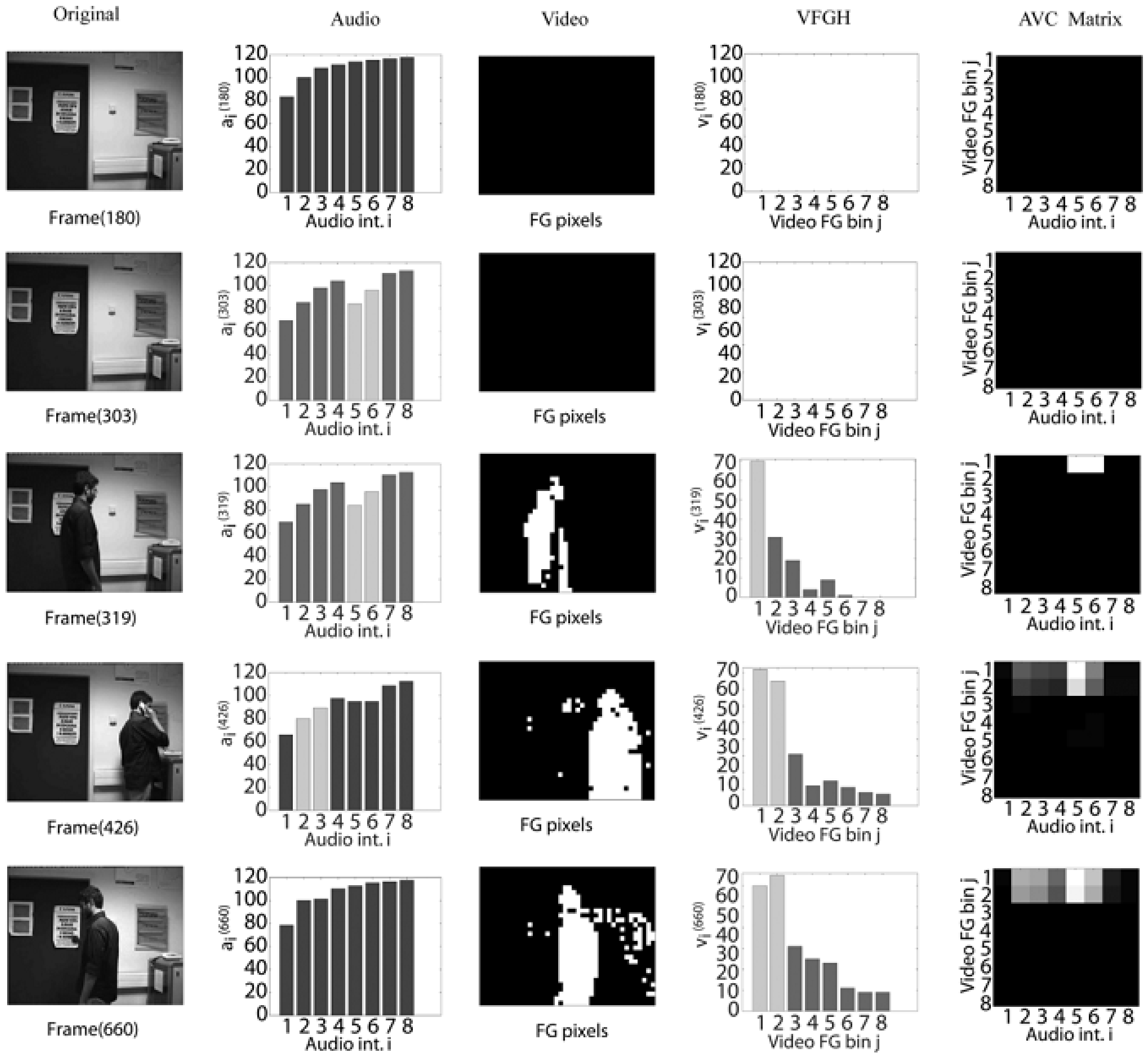

Query
Feature Extraction
MFCCs
Collection of Image Sequences
Omni-Face Detection
Matching by Cross-Modal Association
Retrieval Results

Audio and video analysis: audio/video synchrony analysis (illustration 2)

- Audio/video synchrony analysis
  - Human activities are related to the temporal relations between audio and video signal
  - Current event (the novelty) is considered as the foreground information → Foreground/Background modelling framework
  - FG/BG segmentation : based on time-adapted mixture of Gaussians (TAPPMOG)
- Video and audio histogram
  - J bins for grey level histogram of FG pixels
  - I frequency subbands for histogram of FG audio segments
  - Several Gaussians for each modal histogram
- AVC matrix to encode the degree of simultaneity of audio and video patterns
  - AVC(i,j,t) : mean of weight of activated gaussian in both audio and video TAPPMOG models at time t
- Audio /video event detection
  - AVC(i,j,t+1) - AVC(i,j,t) ≠ 0
- Audio/video event recognition
  - Model the content of each AVC matrix accumulated on a time interval T (KNN)

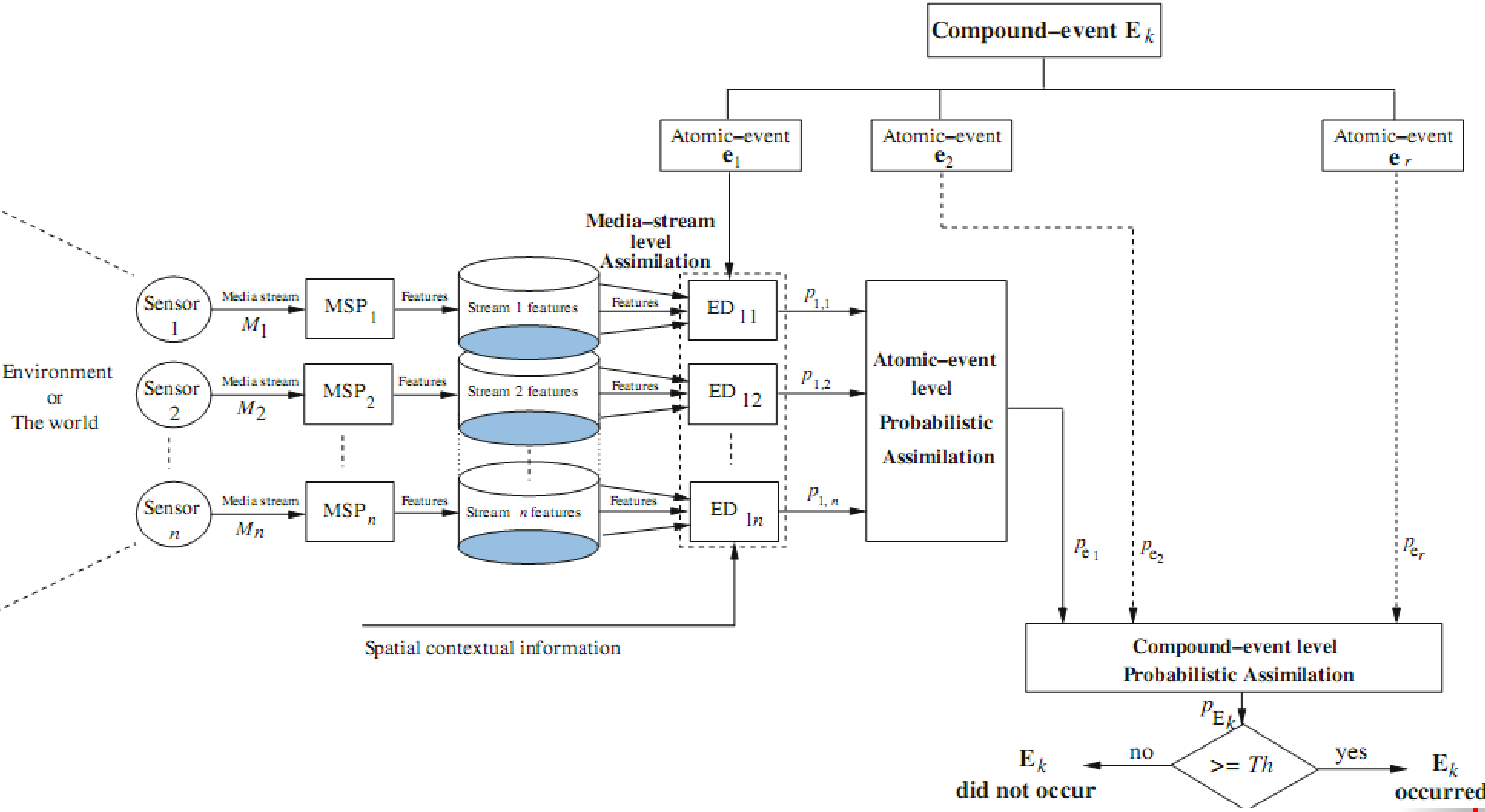


Audio and video analysis: audio/video synchrony analysis (illustration 2)



Audio and video analysis: hierarchical event detection (illustration 3)

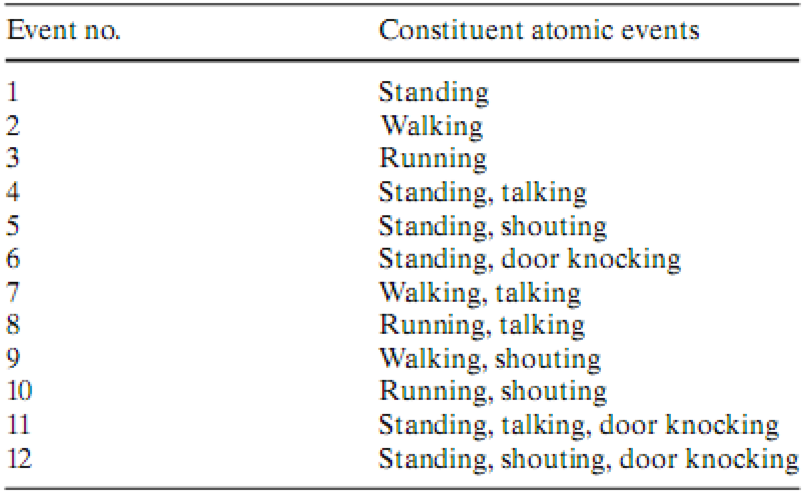
- A surveillance system using audio and video streams
- This work propose to assimilate information at low level for each media stream and at decision level (features assimilation) for multimodal streams (atomic event and compound event assimilation)



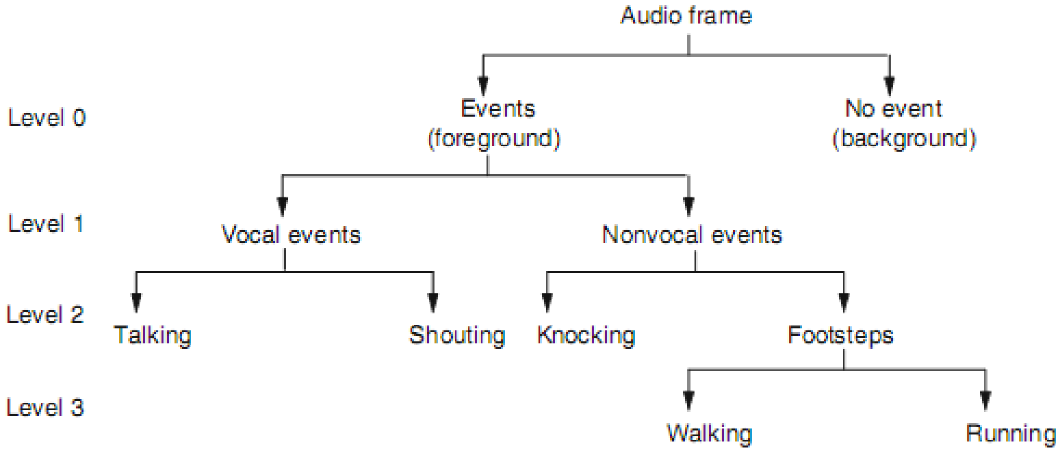
Audio and video analysis: decision level fusion (illustration 3)

- 9 atomic events
  - Standing, walking, door knocking, talking, shouting, running
- Which kind of detection
  - Standing : V
  - Walking, Running: AV
  - Door knocking, talking, shouting: A

- 12 events made of one atomic event and more
- Video based detector
  - Process BG and FG segmentation
  - Blob modelling to detect human body
  - Project blob points on the ground
  - Estimate the speed and the direction on the motion (integration on time interval T)
- Audio based detector
  - Extracted features: LFCC, LPC
  - Gaussian Mixture Model
  - Hierarchical decision

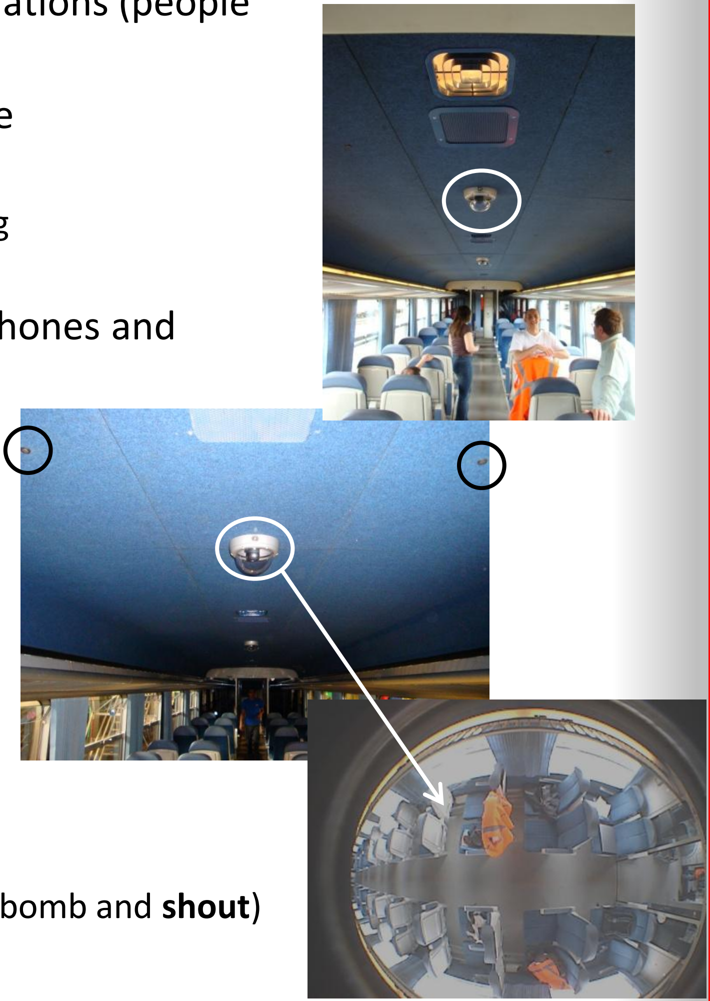


Event no.	Constituent atomic events
1	Standing
2	Walking
3	Running
4	Standing, talking
5	Standing, shouting
6	Standing, door knocking
7	Walking, talking
8	Running, talking
9	Walking, shouting
10	Running, shouting
11	Standing, talking, door knocking
12	Standing, shouting, door knocking



Case of study

- How to use the video and audio signals of a surveillance system onboard a train ?
- Functional objectives: to detect critical and dangerous situations (people fighting, violent robbery, phone snatching, tagging etc.)
- SAMSIT project : omnidirectional microphones and pinhole cameras
  - High level fusion: reasoning in a semantic space and defining an ontology (F. Bremond)
- SURTRAIN project : To use several omnidirectional microphones and fisheye cameras for a better surveillance coverage
  - Develop an audio and video **cooperative system**
    - Audio for detecting and positioning an event
    - To locate the audio event to activate the nearest camera
    - Video for identifying, positioning and tracking the person responsible for the event
    - Video study not presented here
      work done by CEA LIST
- The audio functions
  - Audio event detection: high recall and high precision (spray bomb and **shout**)
  - **Audio event localisation**



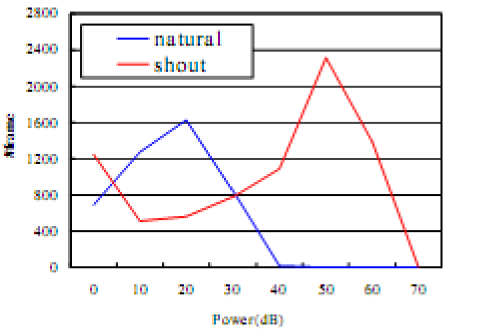
Fisheye image sample

What is a shout ?

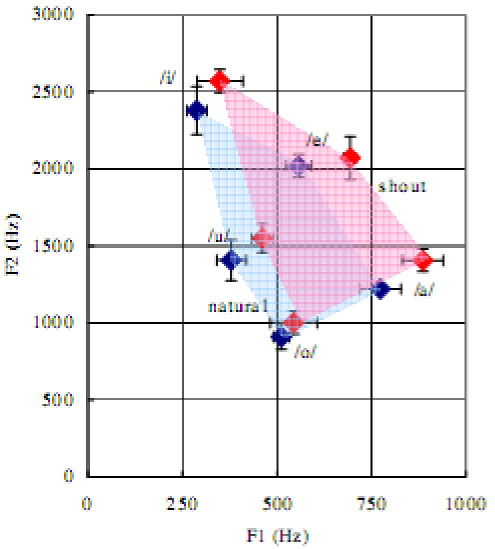
- A shout or a shouted speech are characterised by voiced segments
  - Articulatory process in which the vocal cords vibrate
  - The vocal folds are more stressed

- How acoustical properties of a shout differ from a normal speech ?
  - Fundamental Frequency (F0) is increasing
  - Formants (F1 and F2) are increasing
  - Energy is higher
  - Vowels duration is increasing

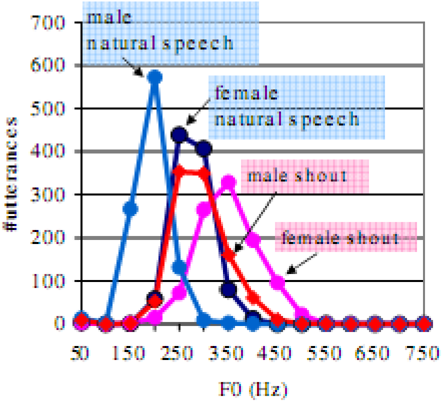
- Difficulties
  - For F0, F1 and F2 ⇨ overlapping distribution for male shout and female speech
  - F0 is correlated to intonation and phrasing
  - Energy of the source is depending on its distance to the microphone



Power histogram from [Nan089]



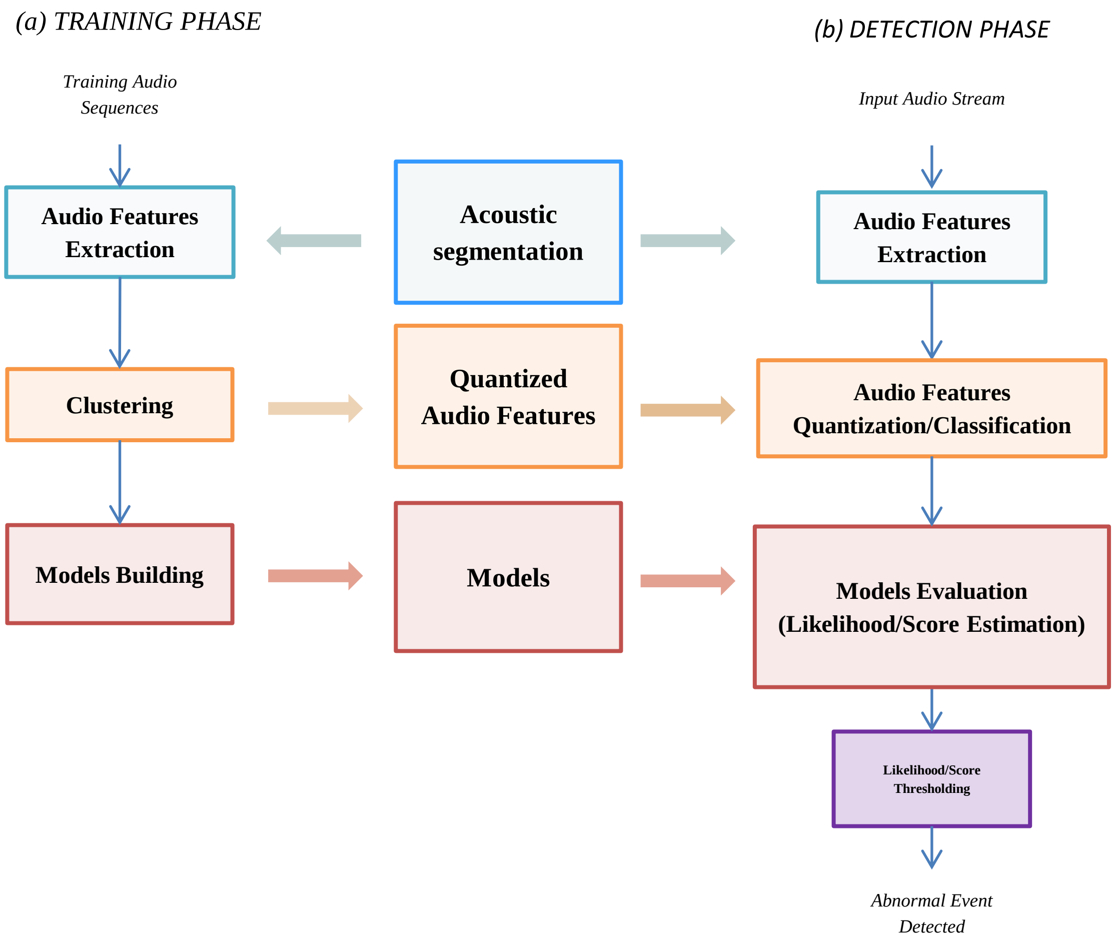
Formants distribution from [Nan089]



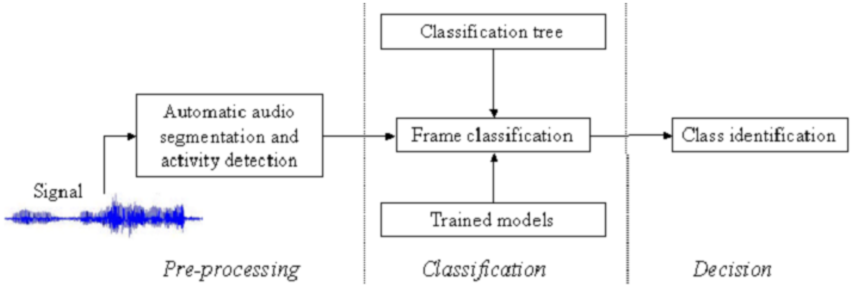
F0 distribution from [Nan089]

Supervised learning (reminder)

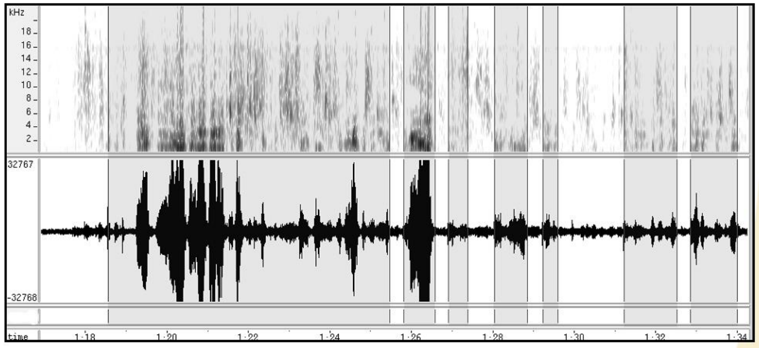
- Two solutions have been proposed
  - EVAS / SAMSIT project
  - SURTRAIN project



Shout detection - first proposition

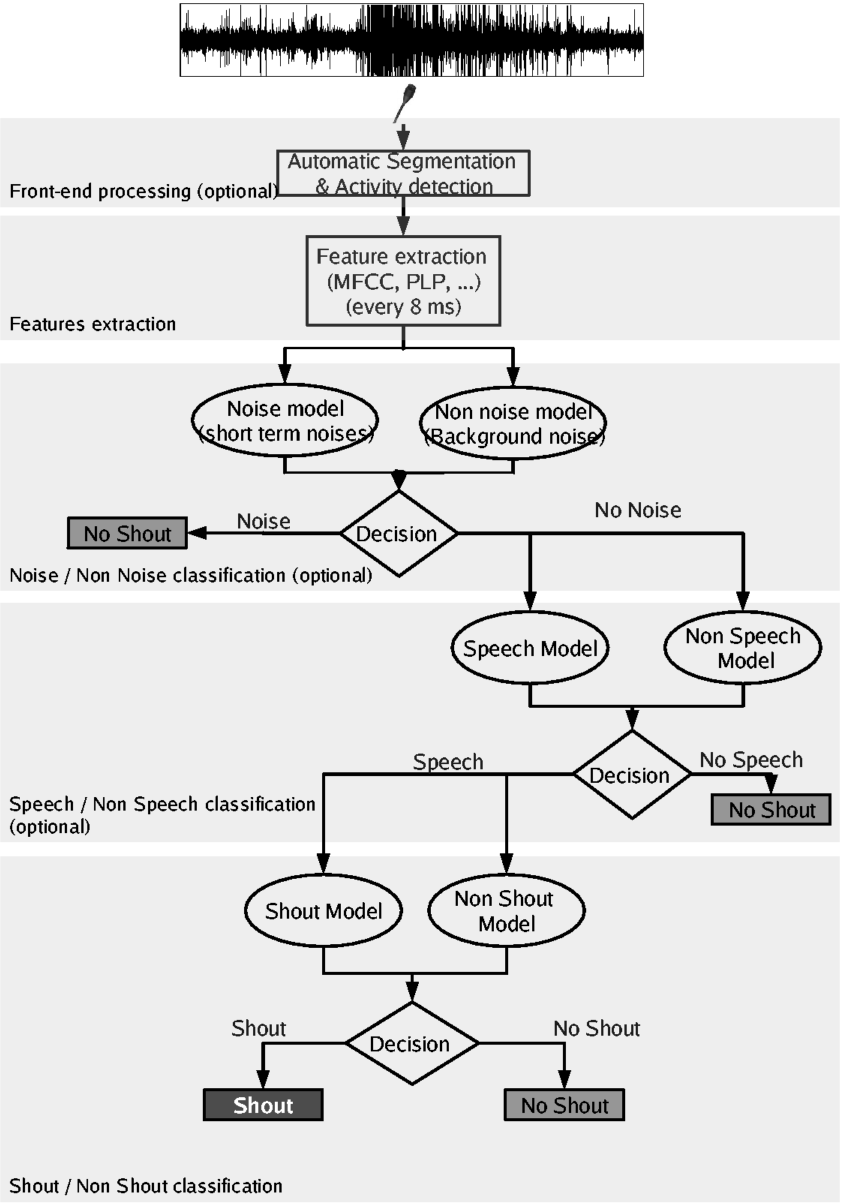


- Features based modelling: MFCC (Mel Frequency cepstral Coefficients), PLP (Perceptual Linear Prediction Coefficients), LPC (Linear Prediction Coefficients) + Energy + first and second derivative
- GMM and SVM
- To reduce complexity and increase performances
  - Automatic audio segmentation and activity detection (in gray)



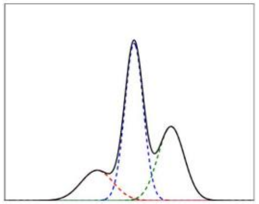
  - Use of decision tree

Shout detection - first proposition

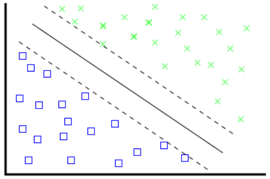

Automatic Segmentation & Activity detection
Front-end processing (optional)
Feature extraction (MFCC, PLP, ...) (every 8 ms)
Features extraction
Noise model (short term noises)
Non noise model (Background noise)
No Shout
Noise
Decision
No Noise
Noise / Non Noise classification (optional)
Speech Model
Non Speech Model
Speech
Decision
No Speech
No Shout
Speech / Non Speech classification (optional)
Shout Model
Non Shout Model
Shout
Decision
No Shout
Shout
No Shout
Shout / Non Shout classification

Shout detection - first proposition

- The data
  - Recorded by ourselves in a regional train
  - Several scenarii with actors (each played several times and once for ”normal condition scene”) – SAMSIT and EVAS project
    - Fight scene involving two people or more
    - Fight scene involving two men and a woman
    - Violent robbery scene (two guys attack one person)
    - Bag and mobile phone snatching (a lady)
  - Total duration: 2402s
  - Shout duration: 138s
- Better results for PLP and SVM modelling



False Alarm rate: 0.12% (3s./2400s.)
% Correct: 56 % (76.8s./140s.)



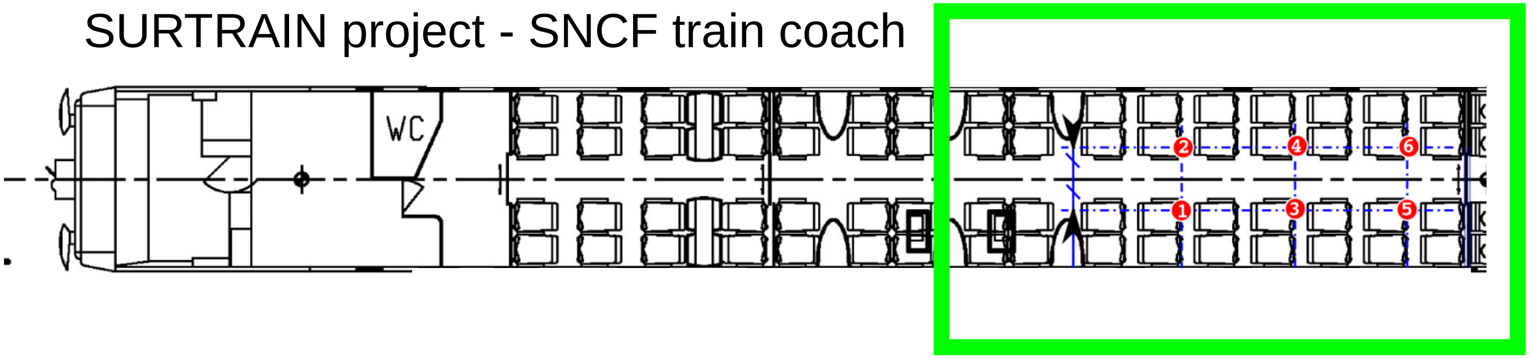
False Alarm rate: 0.05% (1.3s./2400s.)
% Correct: 62.8% (87.9s./140s.)

Shout detection - second proposition

- Which properties ?
  - A shout is composed of voiced segments
  - The duration of voiced segments (vowels) is long
  - Energy is higher when a shout appears ...
  - ... But be careful to the distance between the source and the microphone

- The choices
  - To characterize and to model formants stationarity during a "abnormal" period T
  - To use the four first formants (f0 ... f4) and the energy
  - To model with Gaussian Mixture
  - To use a microphone array (6 microphones) to reduce the position/energy uncertainty

SURTRAIN project - SNCF train coach

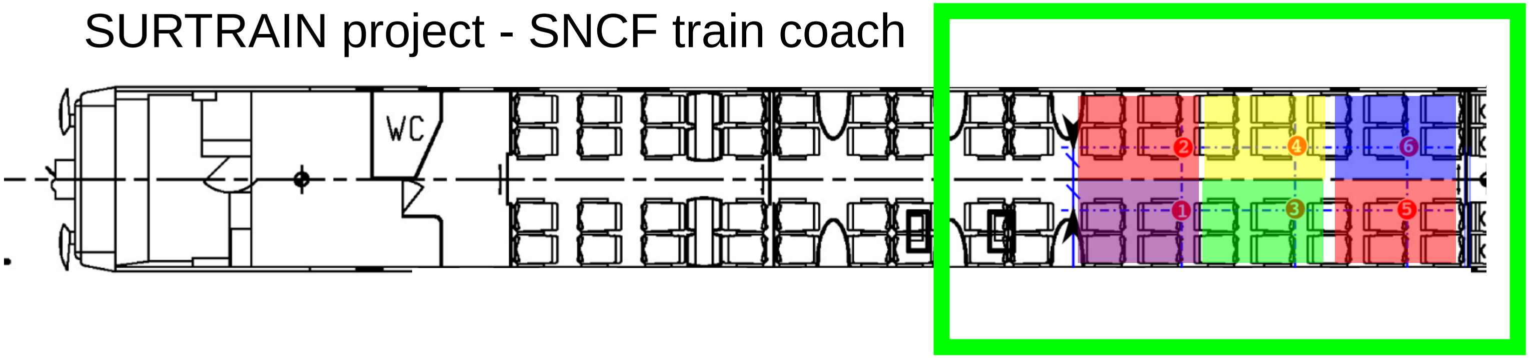


Shout detection - second proposition

- Which properties ?
  - A shout is composed of voiced segments
  - The duration of voiced segments (vowels) is long
  - Energy is higher when a shout appears ...
  - ... But be careful to the distance between the source and the microphone

- The choices
  - To characterize and to model formants stationarity during a period "abnormal" T
  - To use the four first formants (f0 ... f4) and the energy
  - To model with Gaussian Mixture
  - To use a microphone array (6 microphones) to reduce the position/energy uncertainty

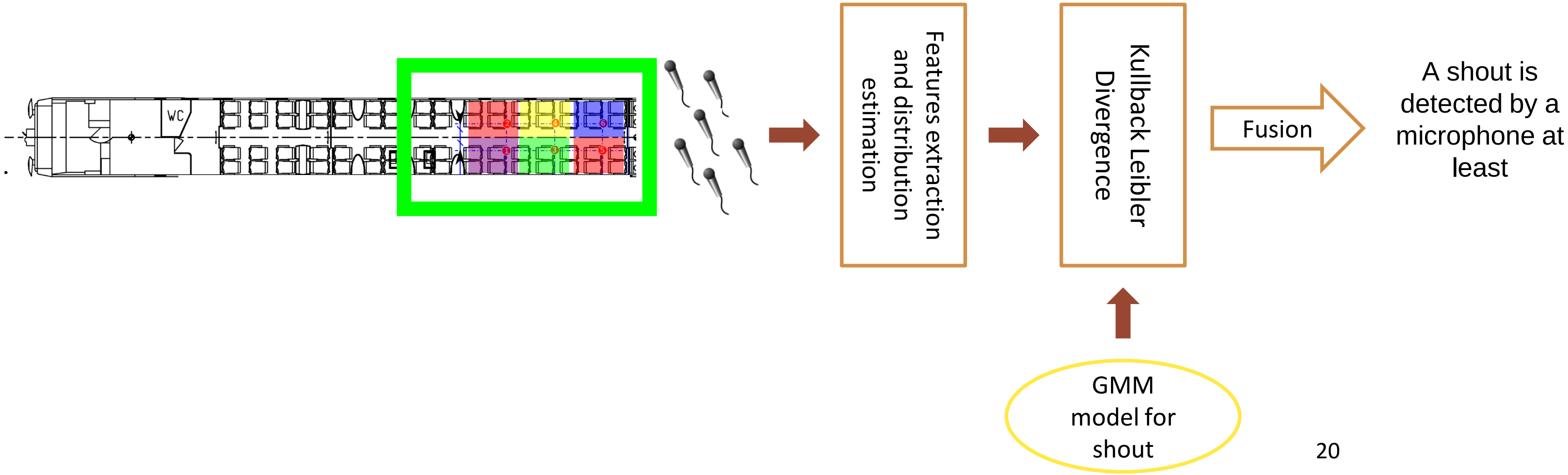
SURTRAIN project - SNCF train coach



Shout detection - second proposition

- Which properties ?
  - A shout is composed of voiced segments
  - The duration of voiced segments (vowels) is long
  - Energy is higher when a shout appears ...
  - ... But be careful to the distance between the source and the microphone

- The choices
  - To characterize and to model formants stationarity during a period "abnormal" T
  - To use the four first formants (f0 ... f4) and the energy
  - To model with Gaussian Mixture
  - To use a microphone array (6 microphones) to reduce the position/energy uncertainty



Features extraction and distribution estimation

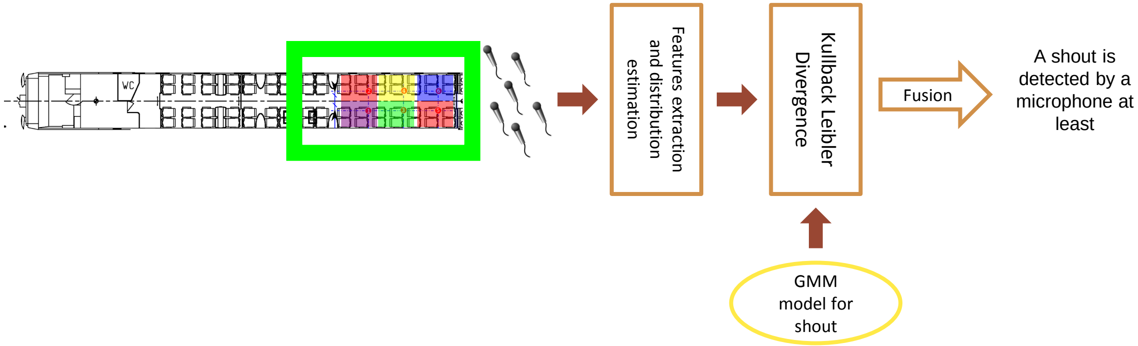
Kullback Leibler Divergence

Fusion

A shout is detected by a microphone at least

GMM model for shout

Shout detection - second proposition



Features extraction and distribution estimation

Kullback Leibler Divergence

Fusion

A shout is detected by a microphone at least

GMM model for shout

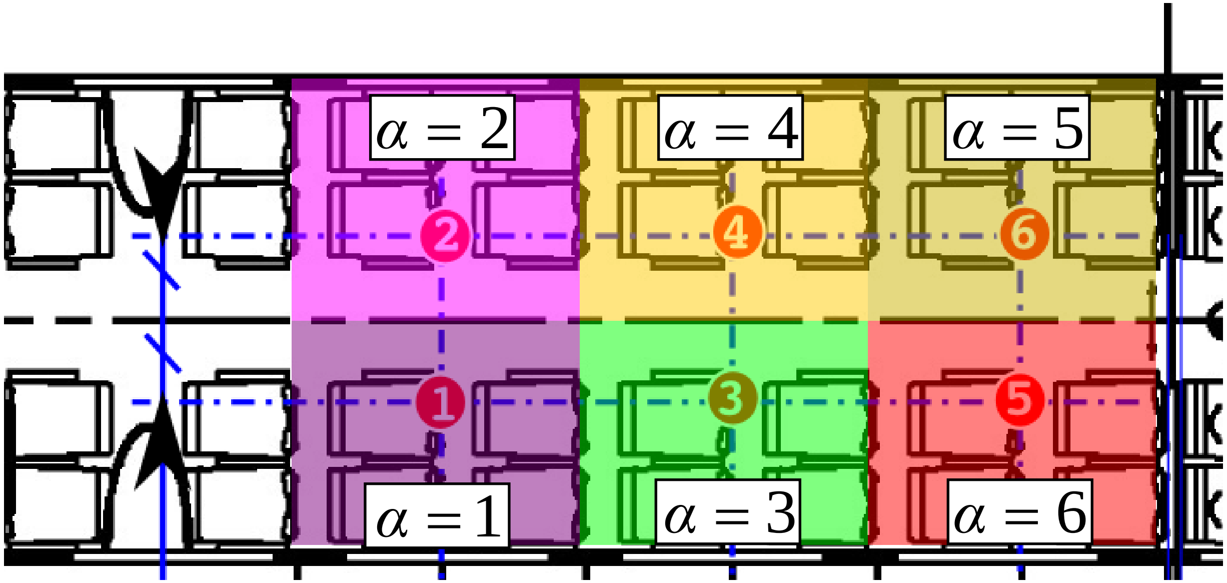
- Evaluation : offline and online
  - off-line case : SAMSIT and EVAS databases and SURTRAIN database
  - On-line case : with the system embedded on-board a train
  - Recall : 0.85 – quite good detection rate
  - Precision : 0.9 – few false detections

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

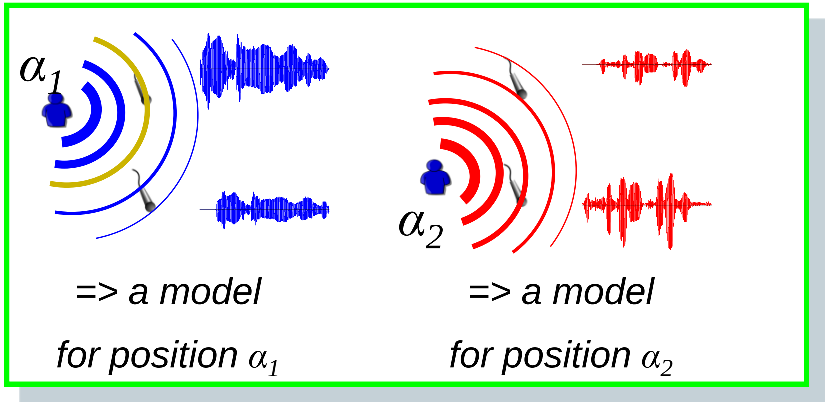
Audio source localisation

- We aim at locating sub-mixtures of audio sources in a set of areas of the train coach
- To use an array of 6 microphones
- 6 areas « centered » in each microphone



- Difficulties
  - Echoic environment: many reflections
  - Audio sources: complex mixture, very different kind of audio sources, difficult to predict a priori the frequency content of the sub-mixtures
  - We focus on the case for which the number of sub-mixtures is equal to the number of sensors

Audio source localisation



α_1

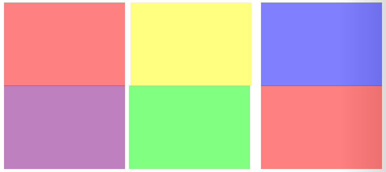
α_2

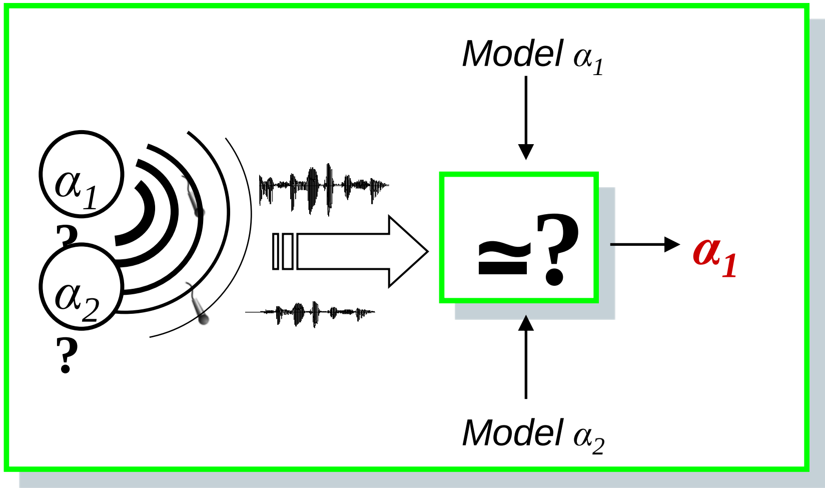
=> a model for position α_1

=> a model for position α_2

- Step 1: **Learning**

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





Model α_1

α_1 ?

α_2 ?

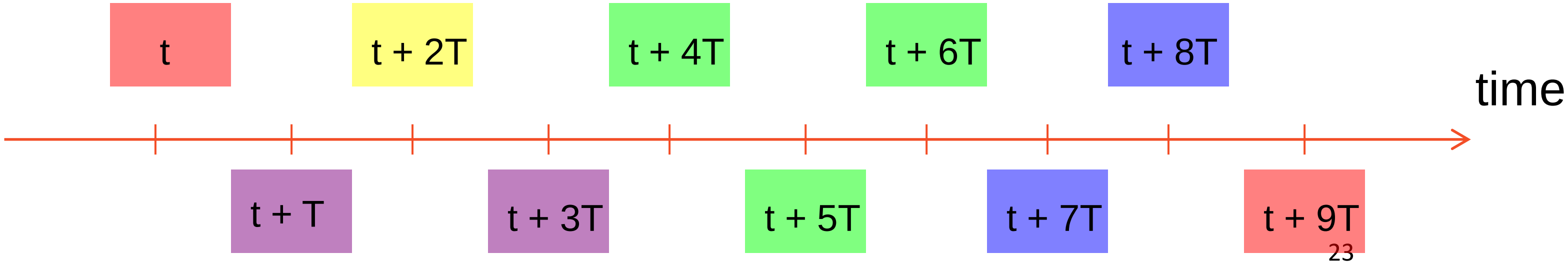
≃?

α_1

Model α_2

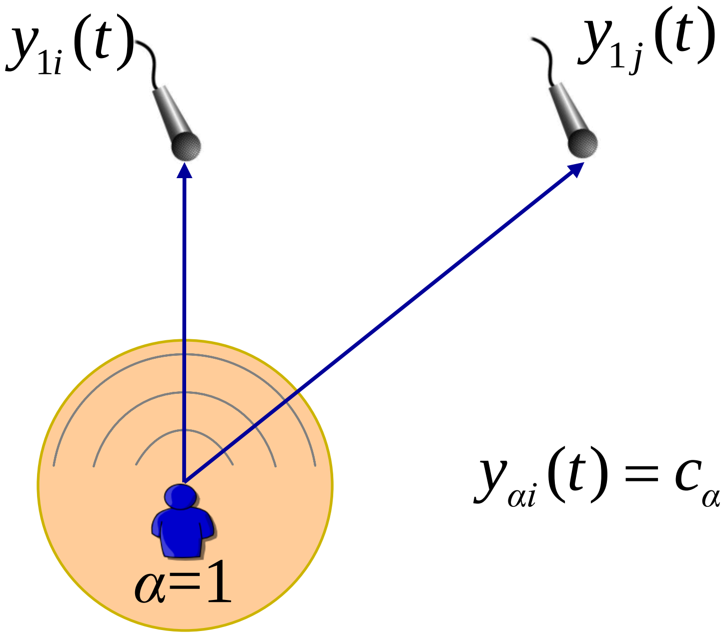
- Step 2: **Localisation**

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



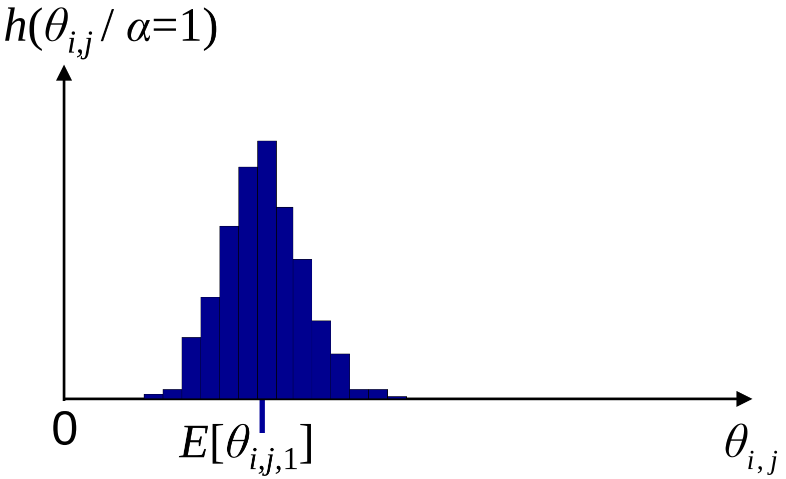
Audio source localisation

- Simple case



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

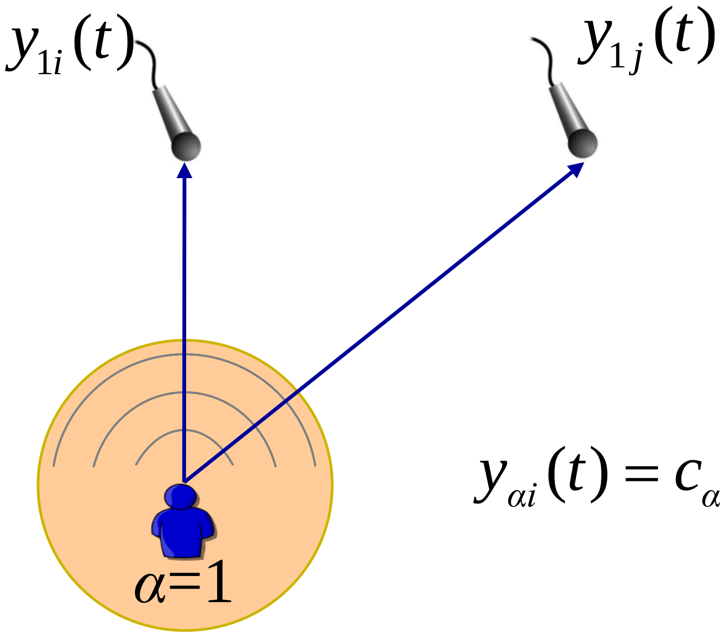
$$\theta_{i,j,\alpha}(t) = \frac{c_{\alpha i}}{c_{\alpha j}} \square \frac{|y_{\alpha i}(t)|}{|y_{\alpha j}(t)|}$$



Position **1**

Audio source localisation

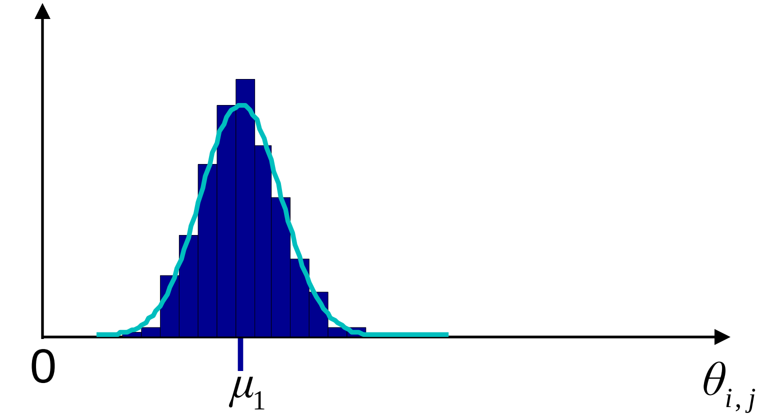
- Simple case



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

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

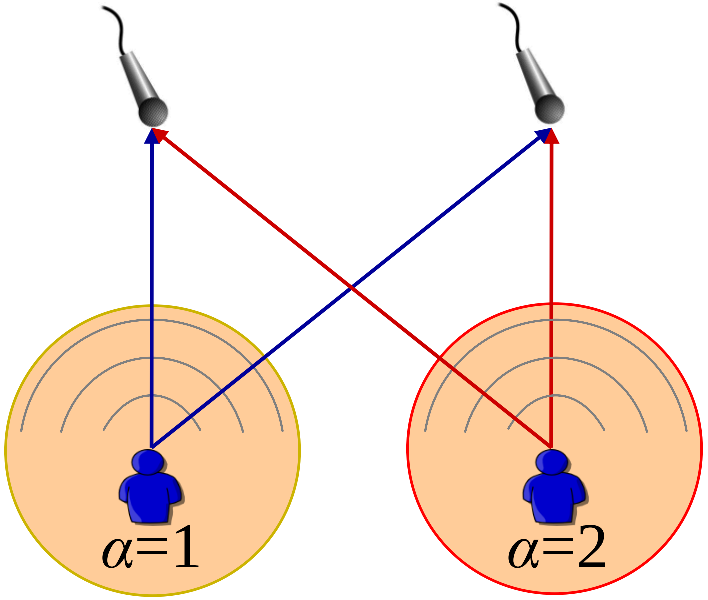
$p(\theta_{i,j} / \alpha=1) = \mathcal{N}(\theta_{i,j};\, \mu_1 \sigma_1)$



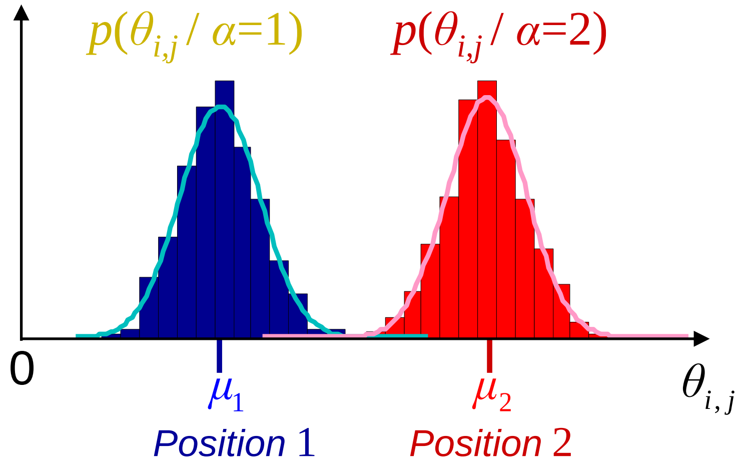
Position **1**

Audio source localisation

- Simple case and multi position

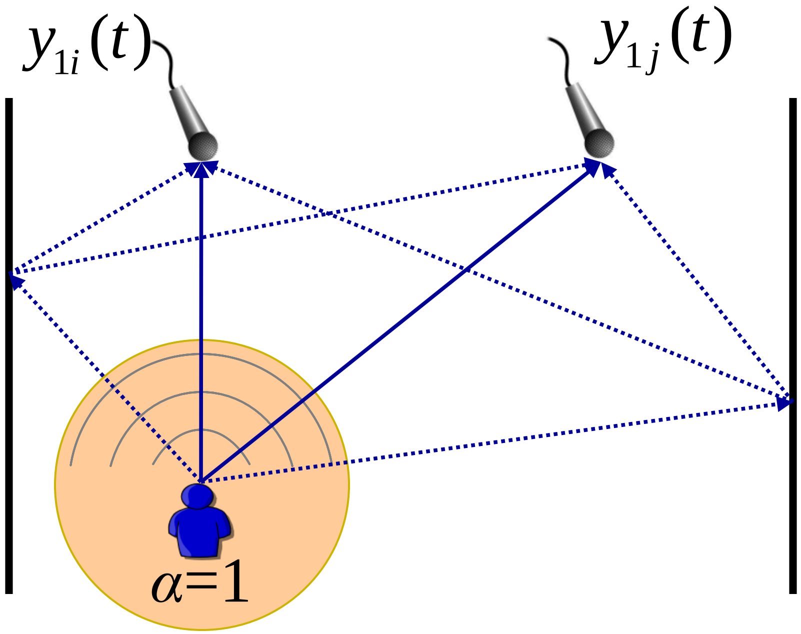


$$\theta_{i,j,\alpha}(t) = \frac{c_{\alpha i}}{c_{\alpha j}} \square \frac{|y_{\alpha i}(t)|}{|y_{\alpha j}(t)|}$$



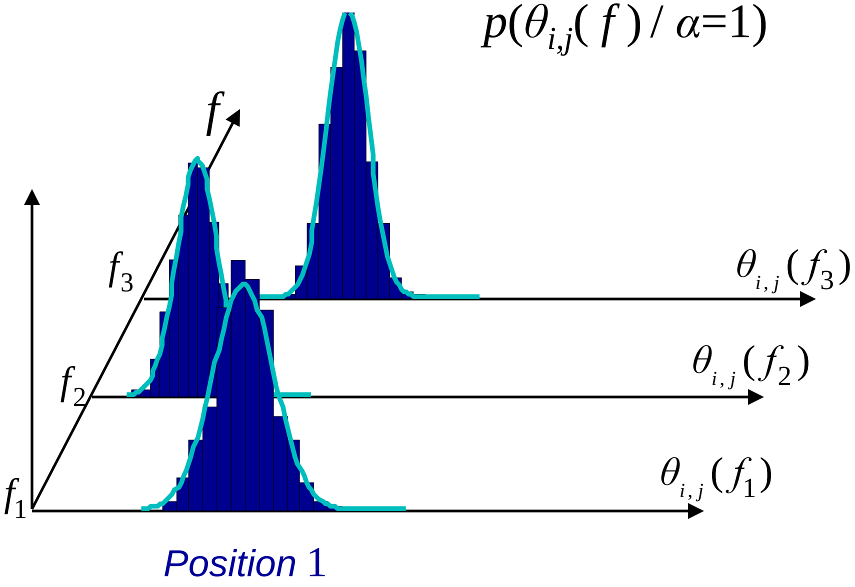
Audio source localisation

- Reverberant case



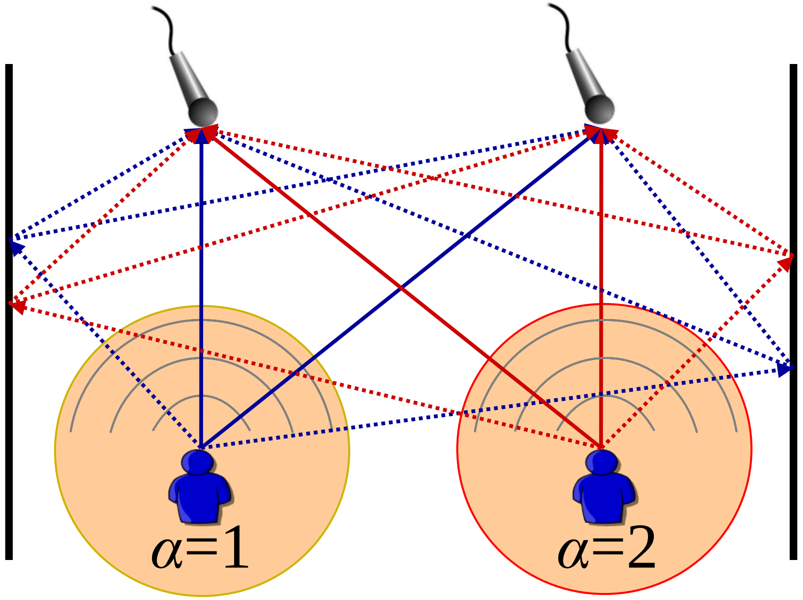
$$y_{\alpha i}(t,f)=c_{\alpha i}(f)s_{\alpha}(t,f)$$

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

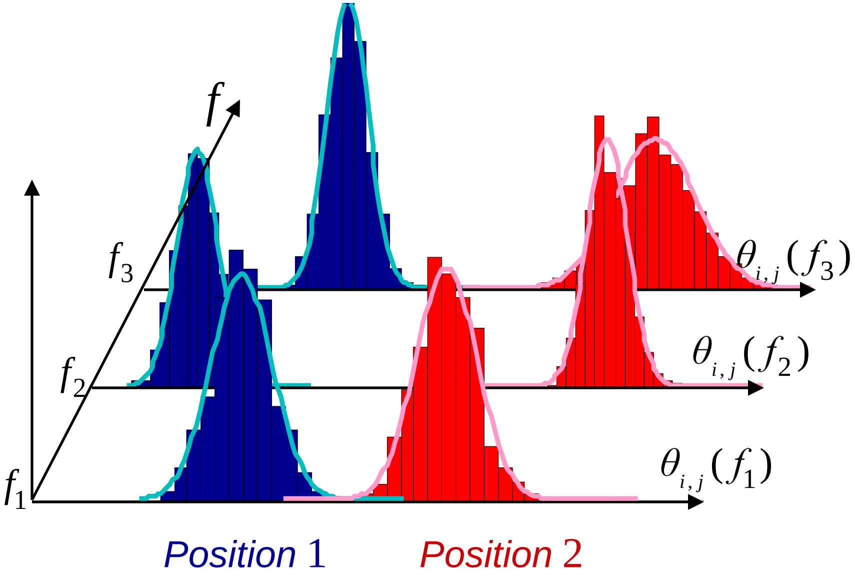


Audio source localisation

- Reverberant case and multiposition

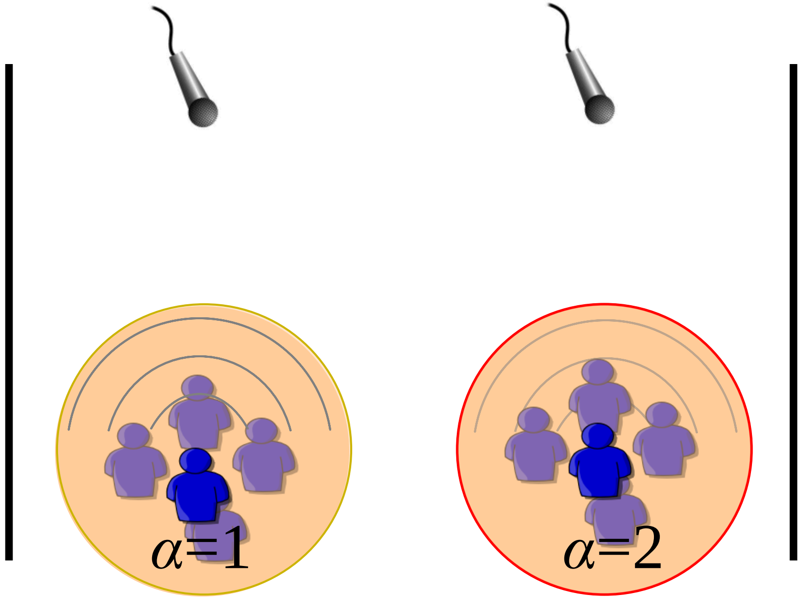


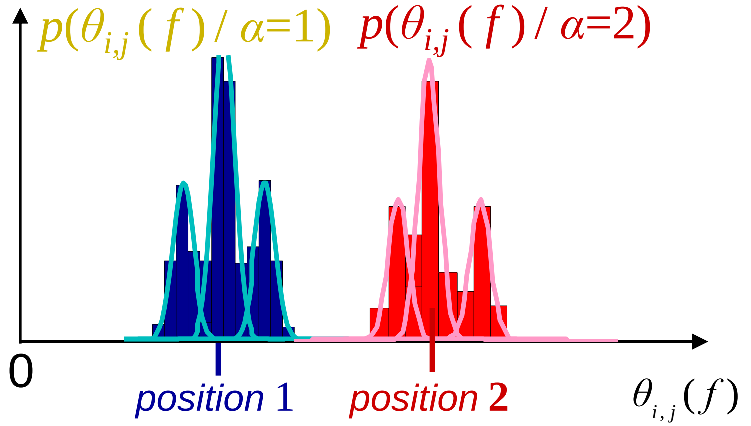
$$\theta_{i,j,\alpha}(f) = \frac{c_{\alpha i}(f)}{c_{\alpha j}(f)} \square \frac{y_{\alpha i}(t,f)}{y_{\alpha j}(t,f)}$$



Audio source localisation

- Variability of the position : multi-gaussian solution



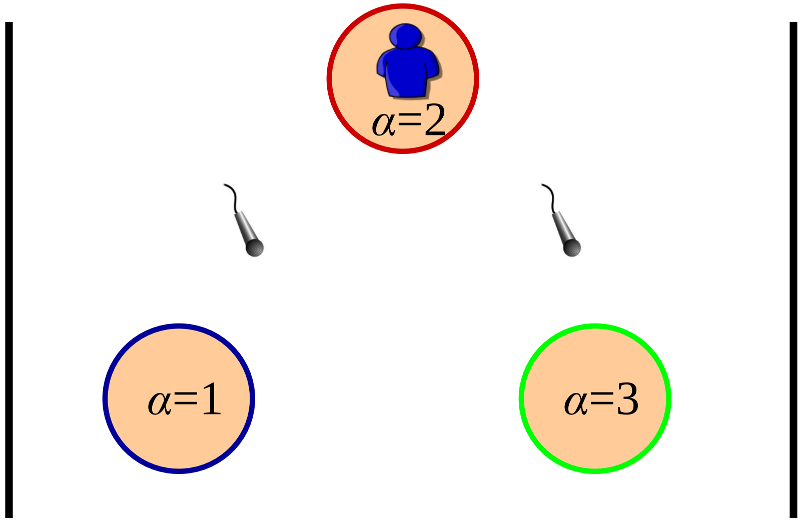


$$p(\theta_{i,j}(f)/\alpha) = \sum_{g=1}^{G} p(g)\mathcal{N}(\theta_{i,j}(f);\mu_{\alpha_g},\sigma_{\alpha_g})$$

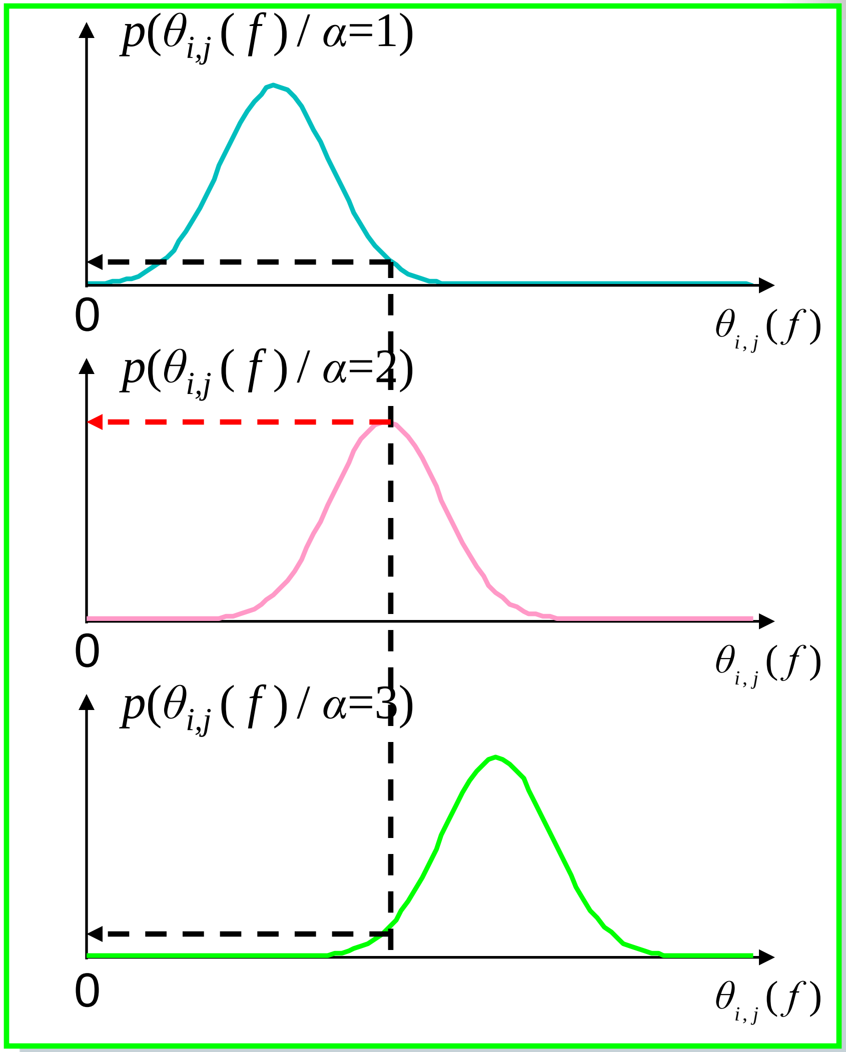
Audio source localisation

- Localisation : M.A.P

Maximum a posteriori



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



Audio source localisation

- Audio sources 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
  - 3 Gaussians per position model
  - Max frequency used *F* = 16kHz
  - Frequency sampling *Fs* = 48k Hz
  - Estimation at every t = 10 ms

Audio source localisation

- Decision made with several pairs of microphones:

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

- Decision made on all frequencies

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

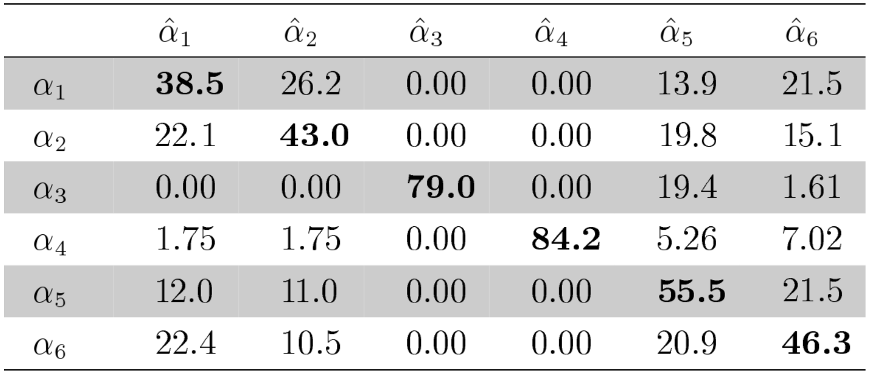
- Decision made on T consecutive time frames

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

Audio source localisation

1 frame - 1 pair of micros

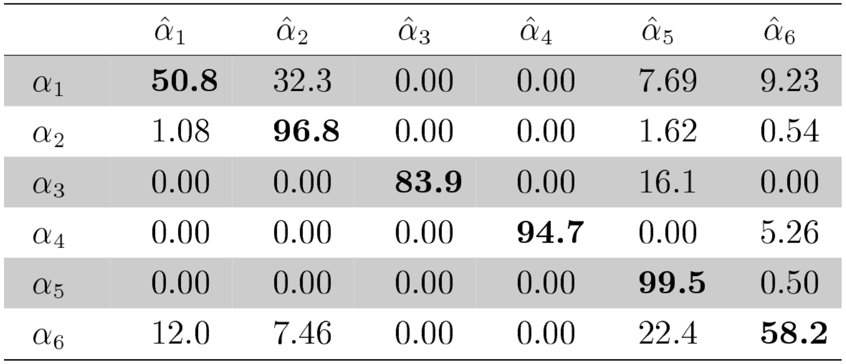
$$\theta_{3,4}(t,f)$$



	$\hat\alpha_1$	$\hat\alpha_2$	$\hat\alpha_3$	$\hat\alpha_4$	$\hat\alpha_5$	$\hat\alpha_6$
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 pair 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**

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

Audio source localisation

1 frame - 1 pair 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 pair of micros

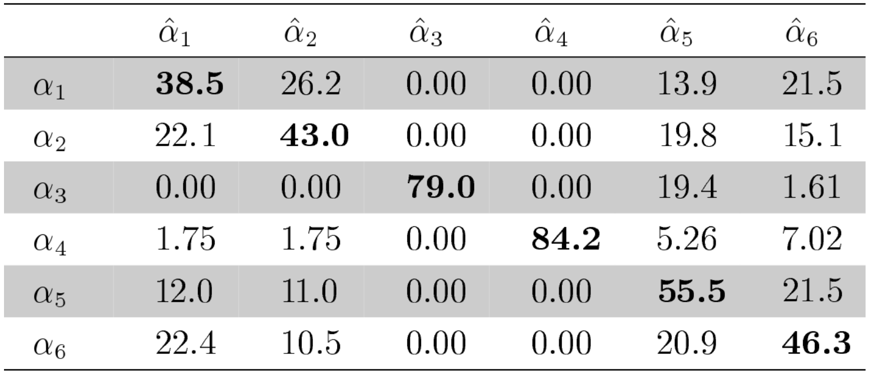
$$\theta_{3,4}(t,f)$$

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
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**

Audio source localisation

1 frame - 1 pair of micros

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



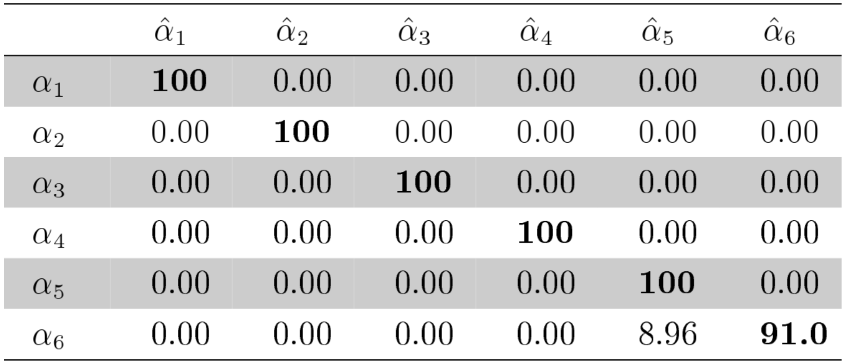
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
α_1	**38.5**	26.2	0.00	0.00	13.9	21.5
α_2	22.1	**43.0**	0.00	0.00	19.8	15.1
α_3	0.00	0.00	**79.0**	0.00	19.4	1.61
α_4	1.75	1.75	0.00	**84.2**	5.26	7.02
α_5	12.0	11.0	0.00	0.00	**55.5**	21.5
α_6	22.4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

5 frames - 6 pairs 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**

Conclusions

- SURTRAIN system
  - A system that jointly uses audio and video signal processing for security application
  - A system embedded and tested in real condition
  - Audio processing for the detection and the localisation of audio source mixture
  - Audio processing for identification of « major source » in the mixture
  - Video processing is initialised thanks to audio outputs