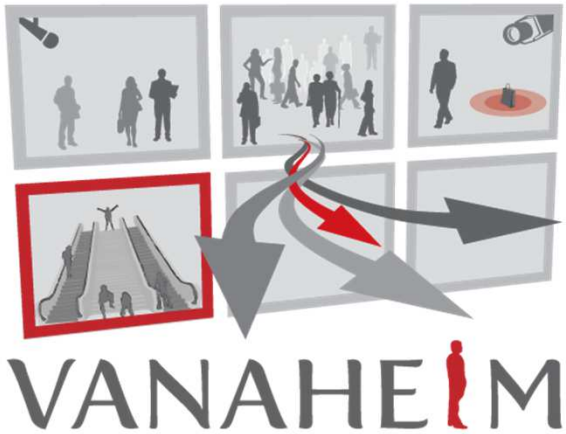


ABNORMAL AUDIO EVENT DETECTION (TCF-IFSTTAR)

François CAPMAN, Sébastien LECOMTE and Bertrand RAVERA (TCF)

Sebastien Ambellouis (IFSTTAR)

VANAHEIM - FP7-ICT-2009-4 - Grant Agreement 248907

Plan

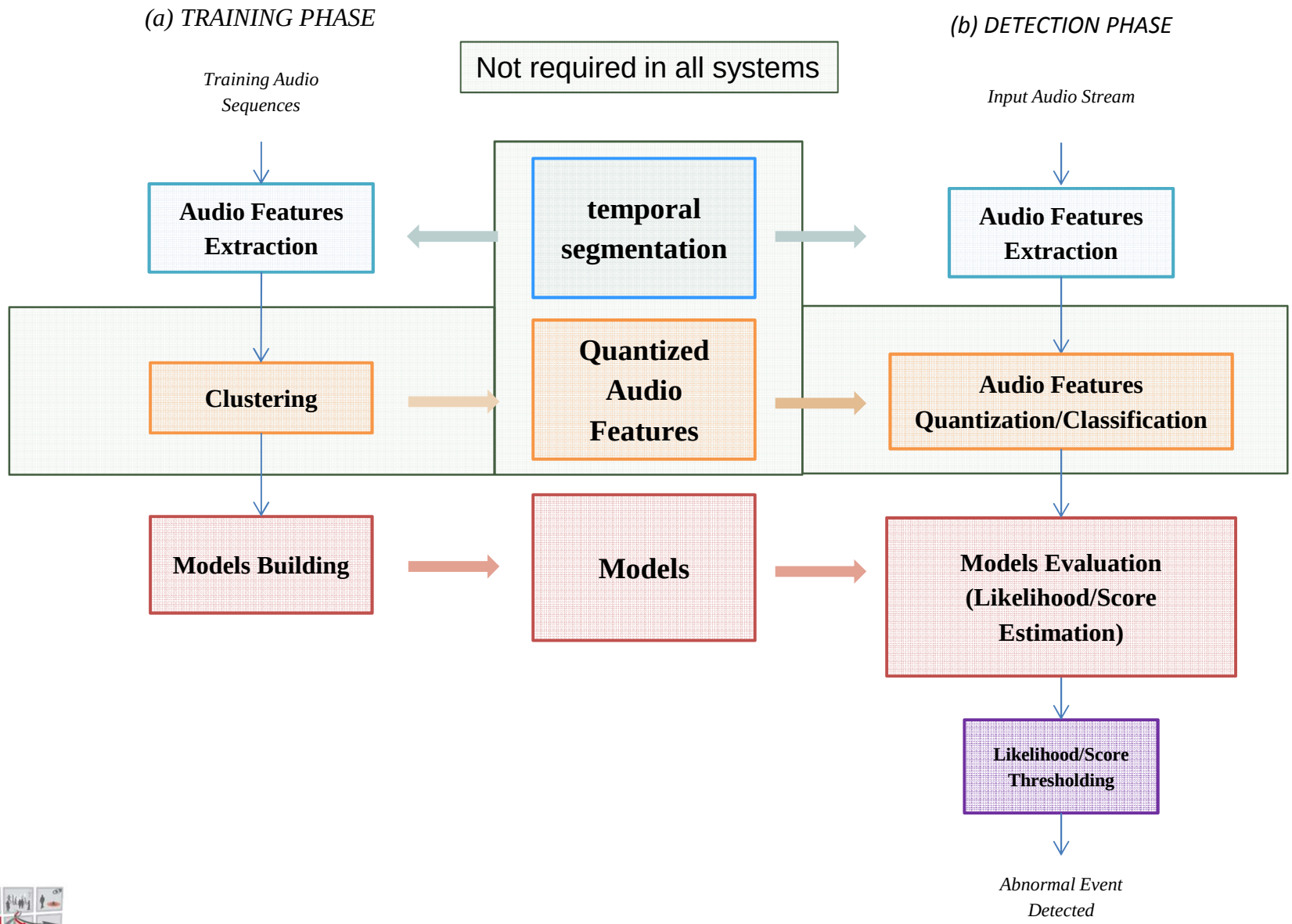
- Generic Audio Surveillance System presentation
- Methodology for Audio Surveillance System performances evaluation
- Multi-Level audio segmentation
- Unsupervised Detection of Abnormal Audio Events for Surveillance Applications
  - GMM-based Audio Ambience Modelling
  - One Class SVM-based Audio Ambience Modelling
- Supervised detection and classification of Abnormal Audio Events for Surveillance Applications
  - One Class SVM-based Audio Ambience Modelling and Abnormal Event Detection

- **Context of audio surveillances**
  - Classical framework for audio analysis
    - 1- **Detection** of abnormal situations.
    - 2- **Recognition/classification** of detected events.
  - Specificities of surveillance signals
    - Noisy environments: ambience is a **non-stationary continuum** that may include lots of events.
    - **No prior on the data distribution** in the “acoustic space”.

- **State of the art / Classical approaches for audio surveillance**
  - **Supervised** : we know what we look for (event model or other knowledge of event)
  - **Unsupervised :** Only ambience is known (event not belonging to ambiance model are abnormal)

- **Presented studies**
  - Improvement of the **detection stage,** focusing on **unsupervised GMM and OC-SVM systems**.
  - Improvement of the **detection and classification stages**, focusing on **supervised OC-SVM systems**

Audio surveillance system presentation

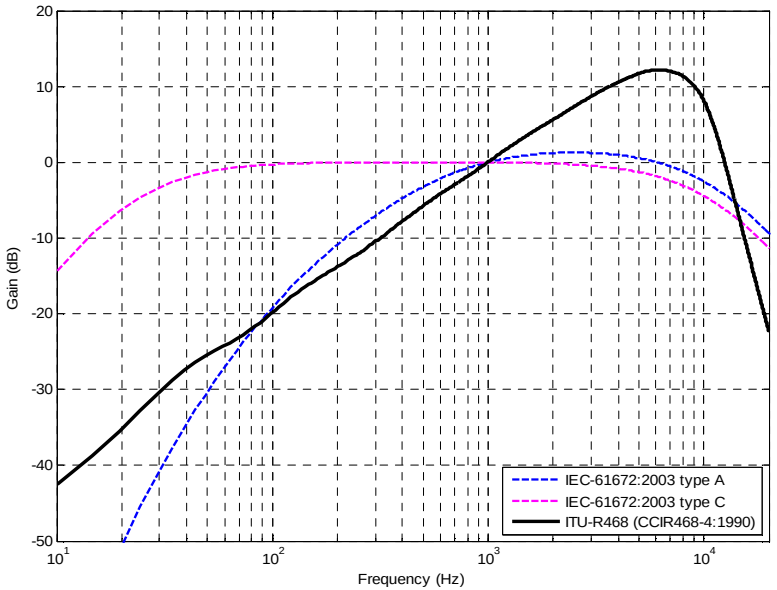


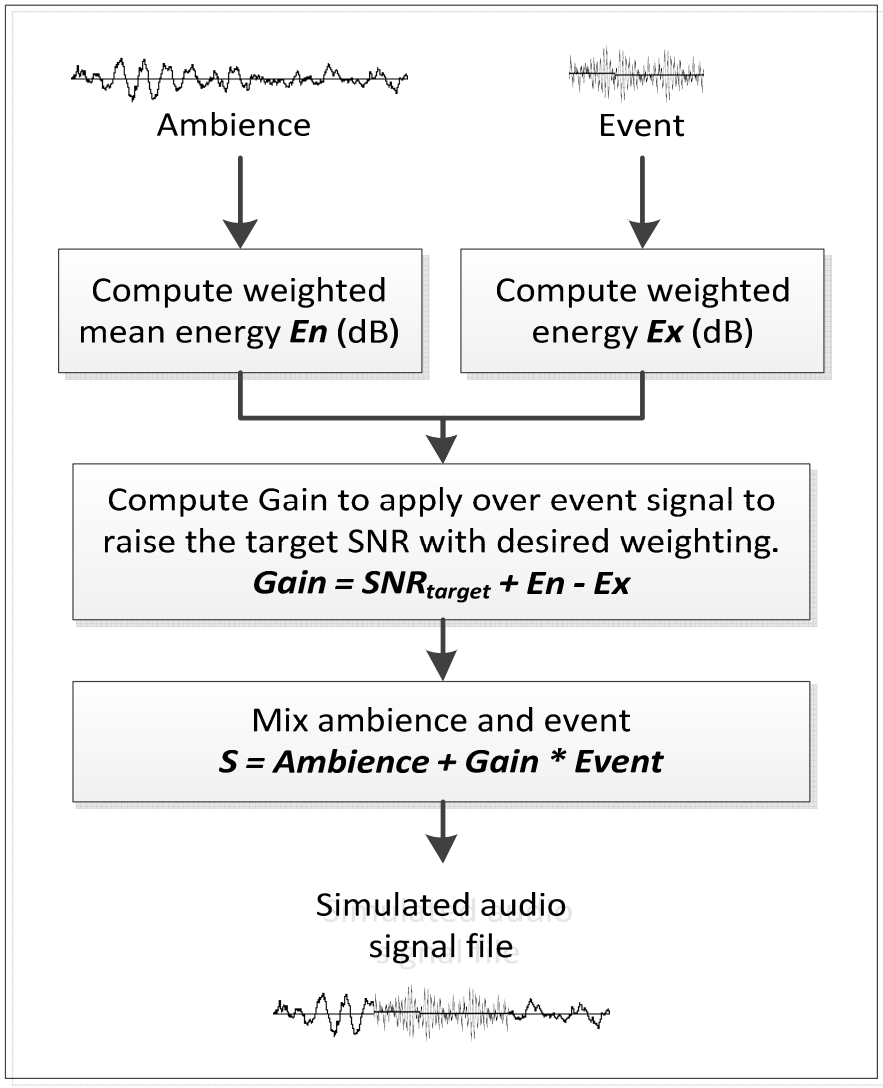
Audio surveillance system performances evaluation

- Problem description
  - Normal ambience material is easily available.
  - Abnormal events are extremely rare (Fortunately !!!!)

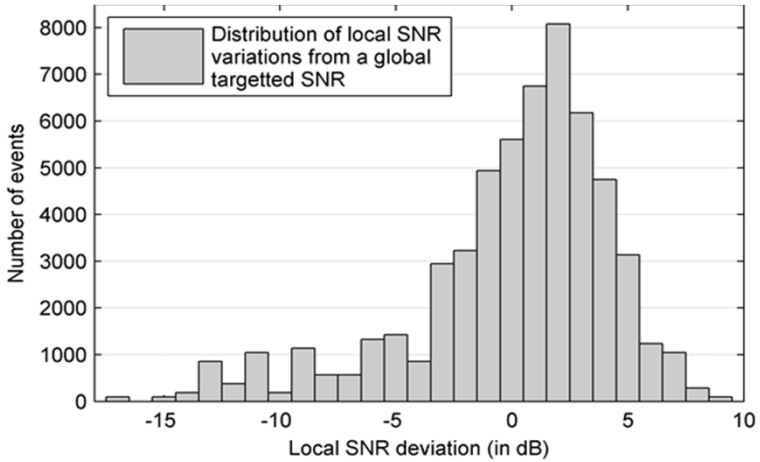
- Proposed system for evaluation purposes
  - Mixture of Normal ambiance material and “Artificial Abnormal Events” (Professional Audio Data Bases).
  - But it requires:
    - Normal ambience and Abnormal events pre-filtering (weighted measure of SNR)
    - Database building with precise and adapted Normal/Abnormal Event Ratio evaluation

- **Pre-filtering (weighted measure of SNR)**
  - Important part of ambience’s mean energy is located in low frequencies.
  - Abnormal events are energetic in full band or high-frequency
  - Use weighted spectrums in order to **reinforce the so-called “utile part of signal”**, which is where ambience and event spectrums overlap. (ITU-R468 and Low Frequencies debiasing)
  - This approach also gives a more **perceptive evaluation of SNR** related to high frequencies.





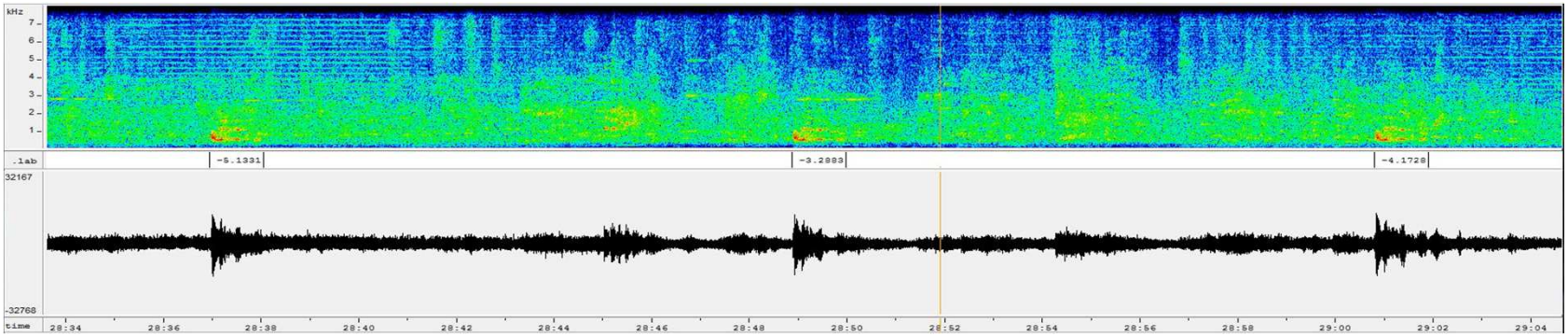
- **Database Building**
  - **<u>No specific database exists</u>**
  - Design of a framework in order to simulate adequate surveillance signals for evaluation purposes
  - Global **Normal/Abnormal Event Ratio (SNR)** targeted.
  - Local **Variable SNR** (variability related to real operational conditions – energy variation of real ambience leads to SNR variation – realistic use cases)



Audio surveillance system performances evaluation

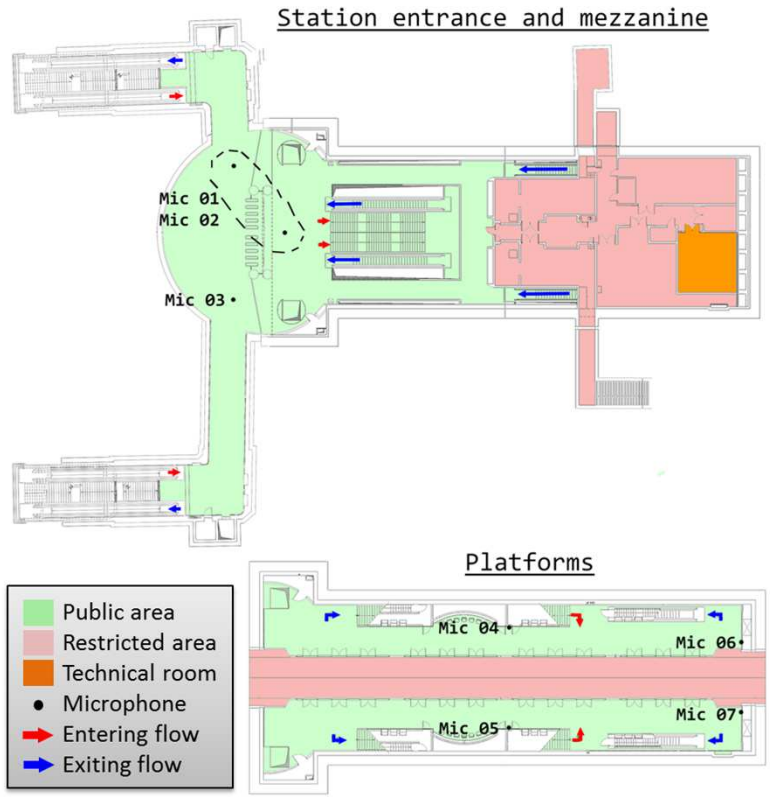
- Simulation of a complete database
  - 96 events (27 different types)
    - Telephones, sirens, various kind of screams, various kind of crowd noises (fight, cheer, bravo, applause, …), various kind of explosions, …
    - Some more exotic ones as dog noises, …
  - Audio ambience files (from real site acquisition)
  - SNR from 0dB to 30dB.
  - Several hours easily available for models training/testing.
  - **<u>Flexible and powerful tools for audio sequence generation</u>**

  - Example : Dog noise – targeted SNR 0.0 dB – Torino metro ambience



- **Signal collection system (Torino Metro Station) and protocol**

- Ambience is collected during regular metro operation (during the day)
- Abnormal event are played in station and then collected (during the night)
- **<u>"real" abnormal events used for algorithm evaluation</u>** (mixed with real ambience)



Audio Features extraction

- **Definition**
  - This procedure, the so-called **parameterization** of the signal, consists in transforming the waveform into a series of vectors of parameters. The parameters are also called **acoustic features**.

- **Audio features Types**
  - Loudness features (relatives to energy considerations)
  - Time-Domain features (ex. Zero crossing rate)
  - Frequency-Domain features,
    - Linear frequency sub-band energies (LFSBE),
    - Mel frequency sub-band energies (MFSBE),
    - Linear Frequency Cepstral Coefficients (LFCC),
    - Mel Frequency Cepstral Coefficients (MFCC).
  - Statistical features (ex. Power Spectrum Density (PSD) mean an variance),
  - Regression features (ex. PSD linear regression),
  - Parametric features (ex linear predictive coding coefficients extraction LPCC).

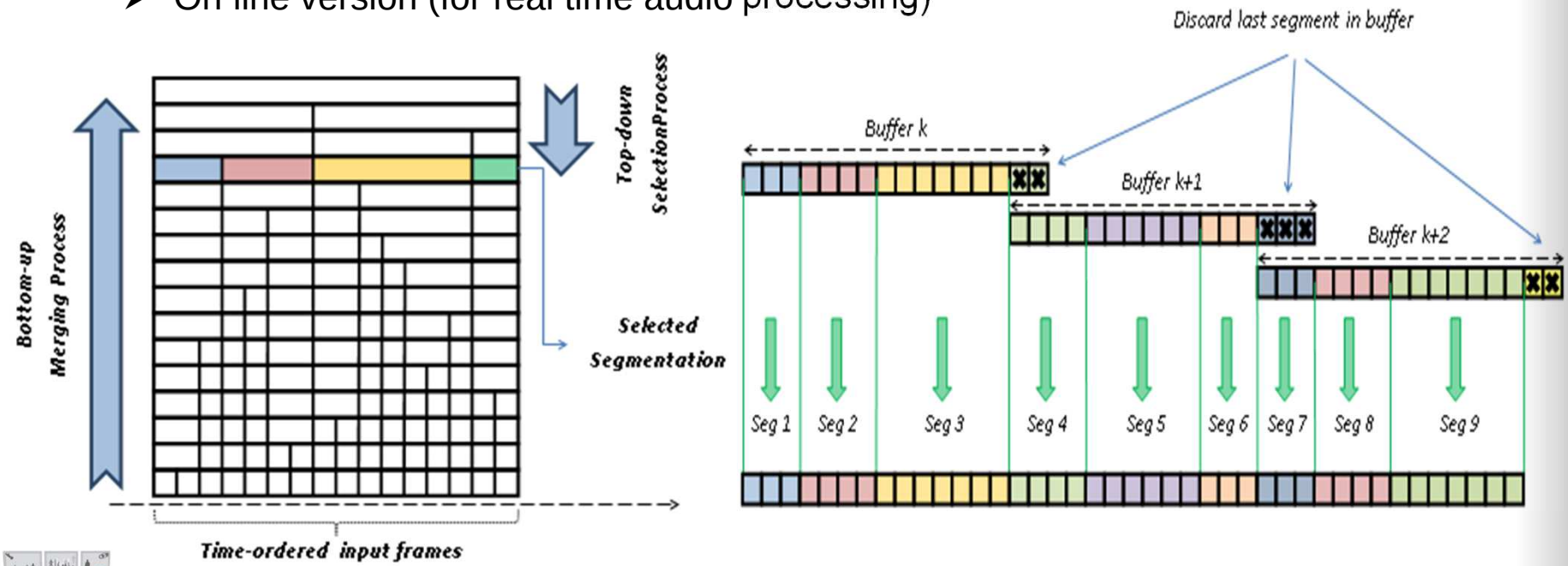
- **Audio features used**
  - **LFSBE** (from 8 to 24 bands)
  - **MFCC** (from 10 to 20)

Frame by frame extraction

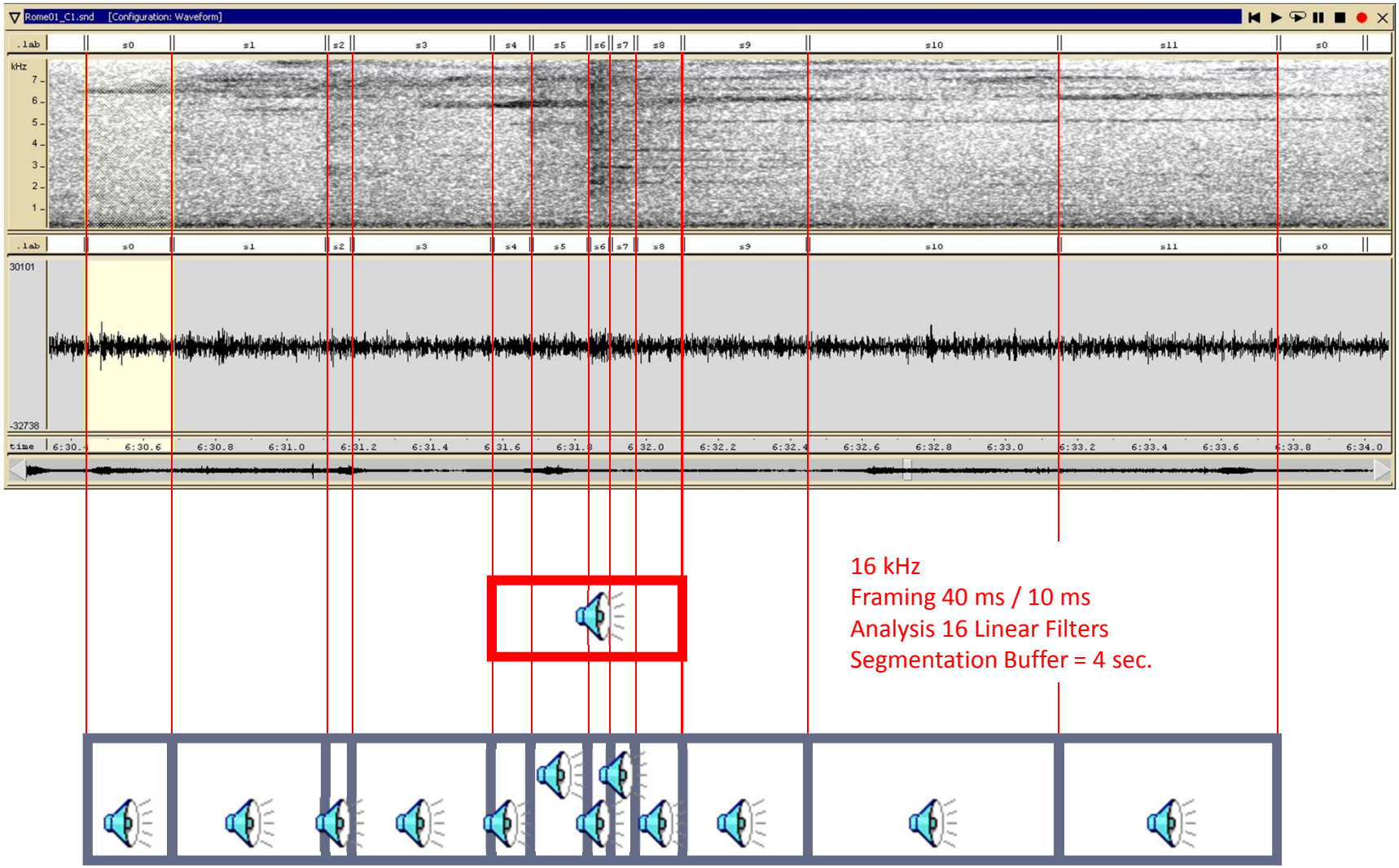
- One acoustic features set (vector) for each frame

Multi Level segmentation

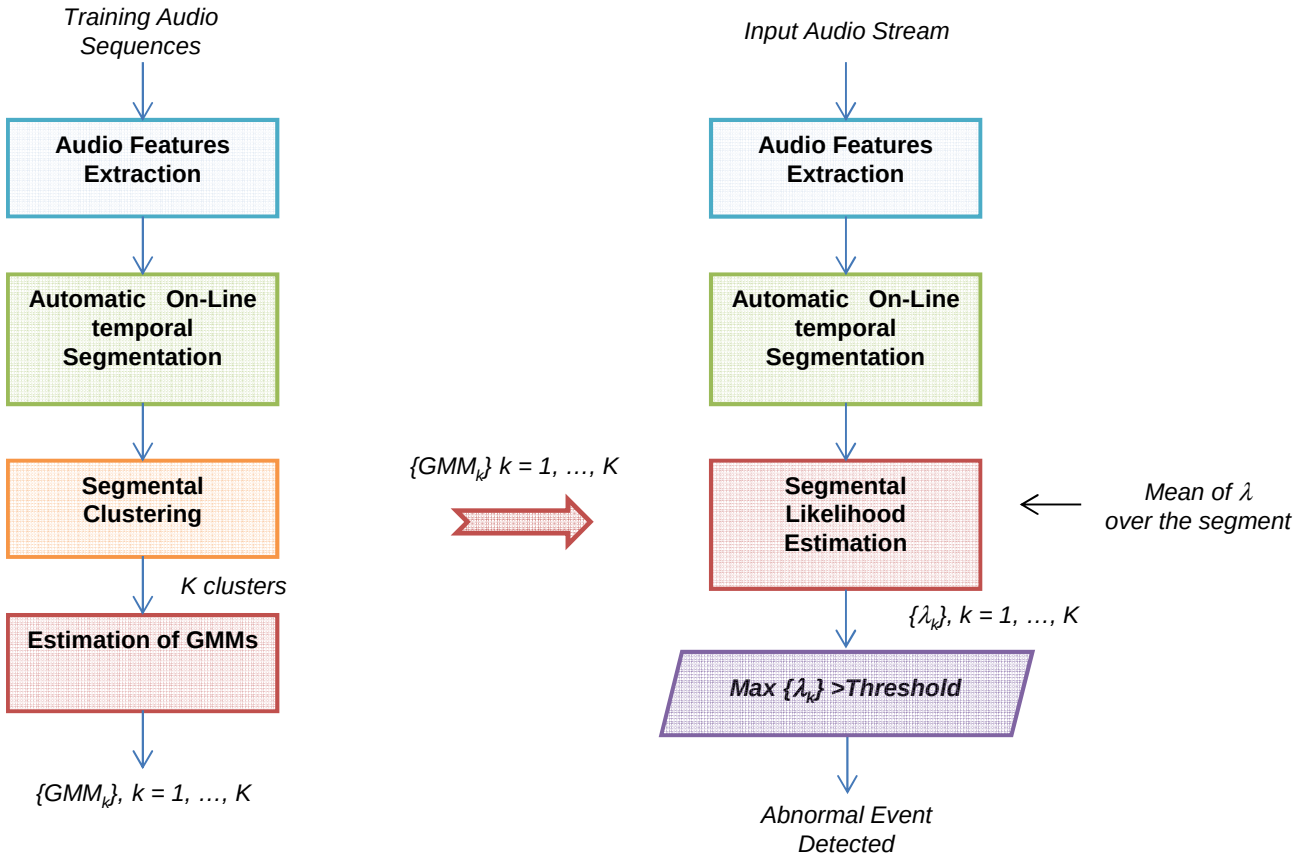
- **Dendrogram** based bottom-up acoustic description which varies from fine to coarse (based **on acoustic parameters correlation** measurement)
- **BIC based** segmentation
- One acoustic parameters vector for several regrouped frames (mean over the segment)
- On line version (for real time audio processing)



Multi level segmentation



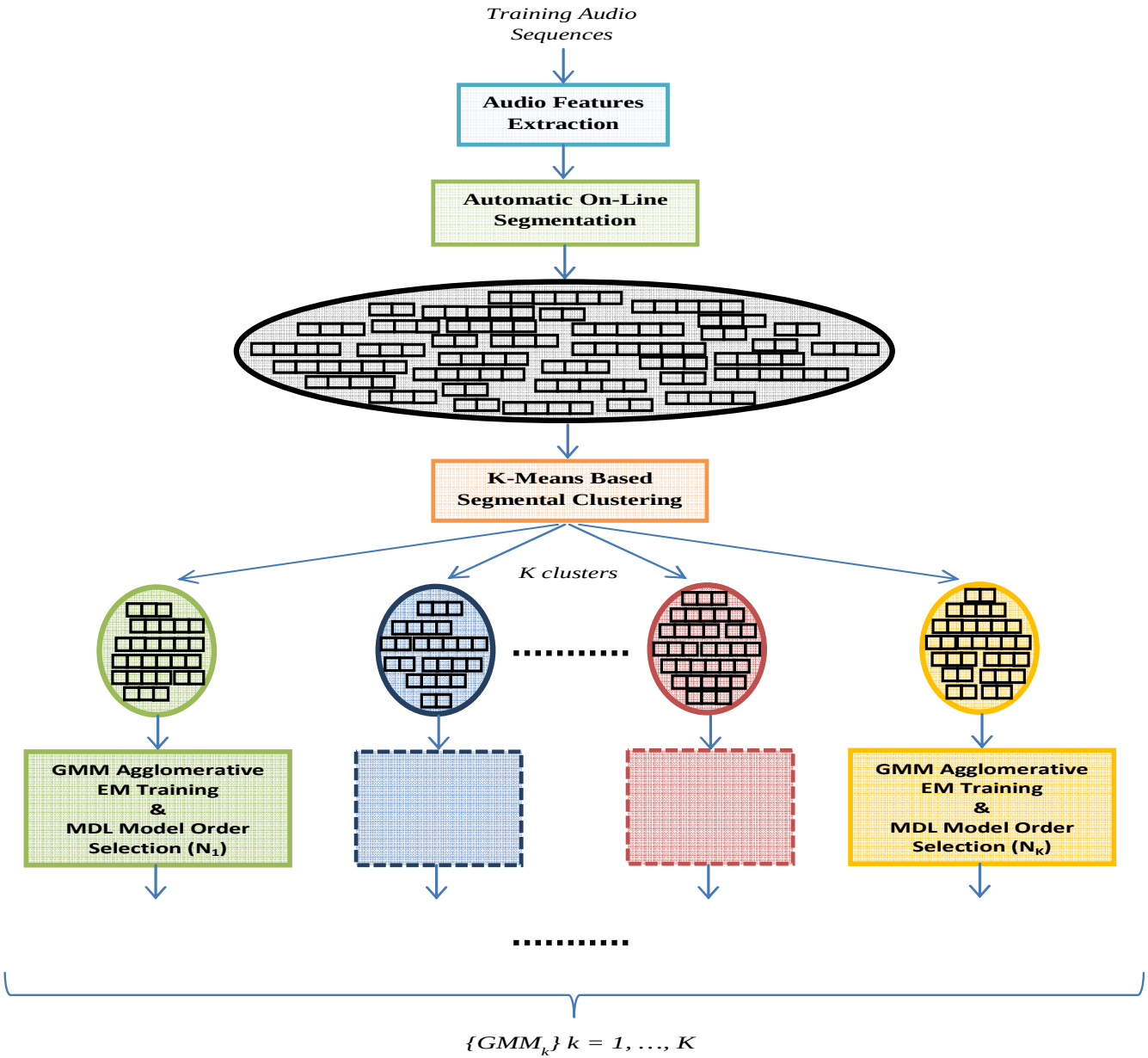
GMM for Audio Abnormal Events Detection



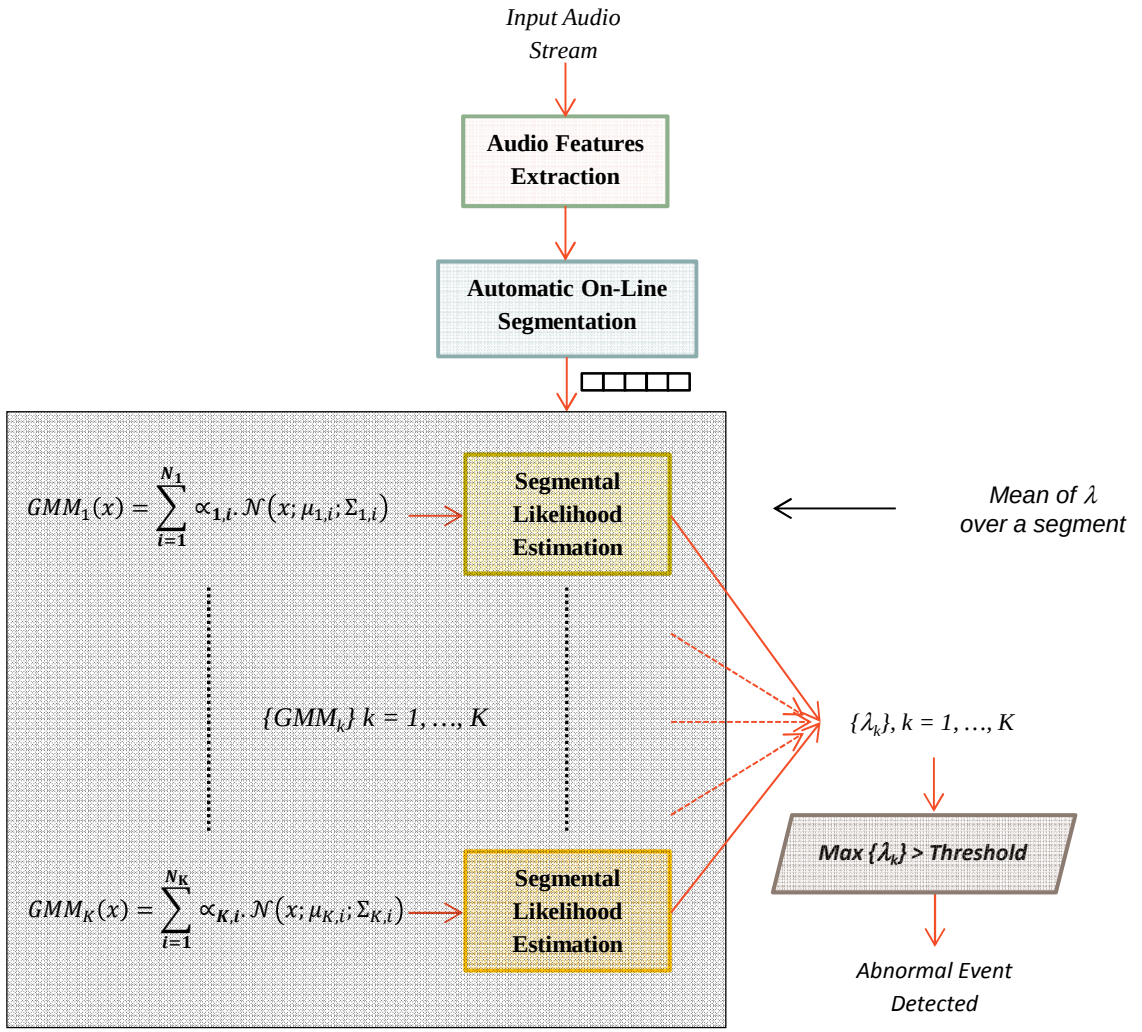
(a) TRAINING PHASE

(b) DETECTION PHASE

GMM for Audio Abnormal Events Detection



GMM for Audio Abnormal Events Detection

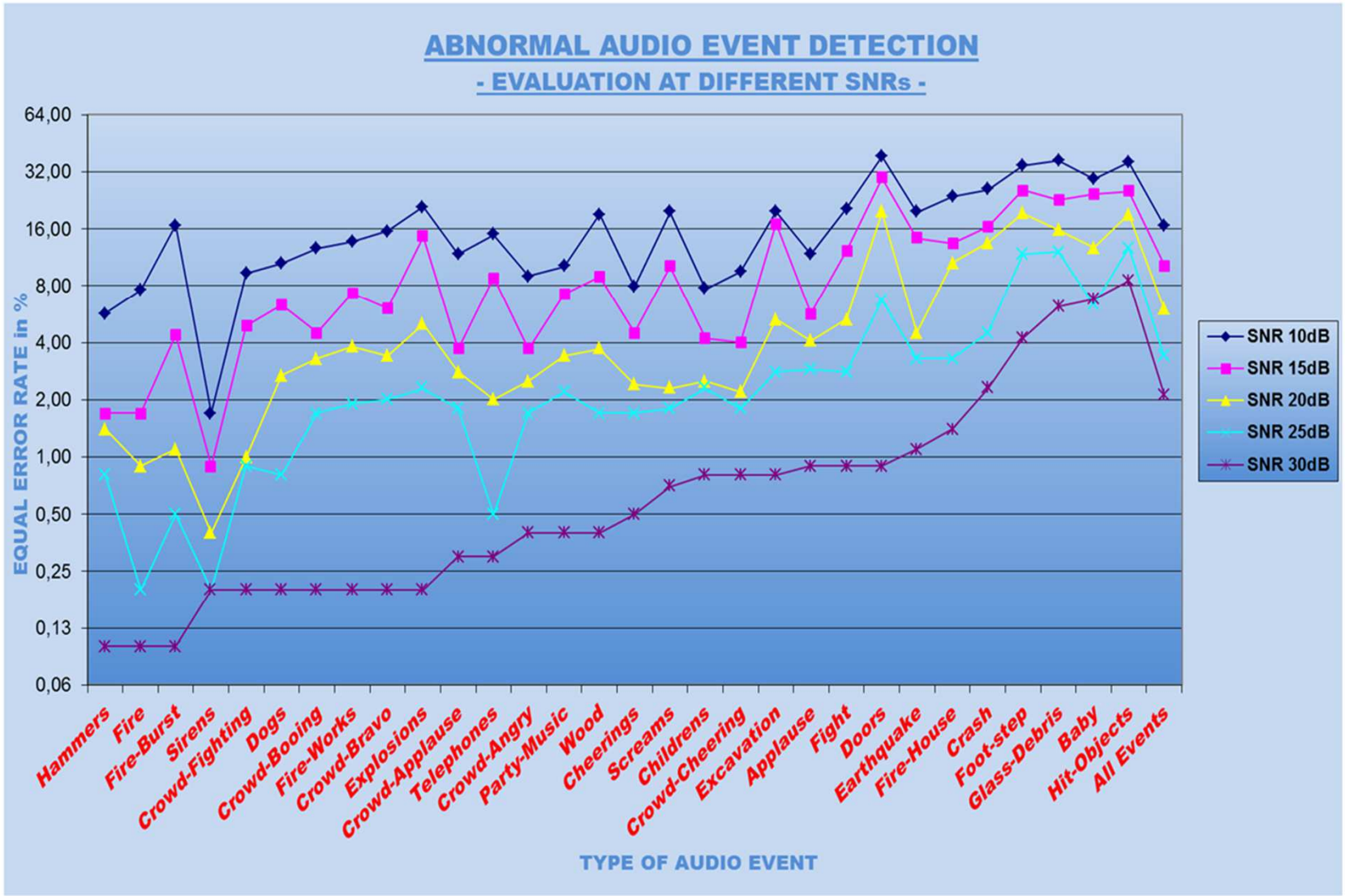


GMM for Audio Abnormal Events Detection (Evaluation)

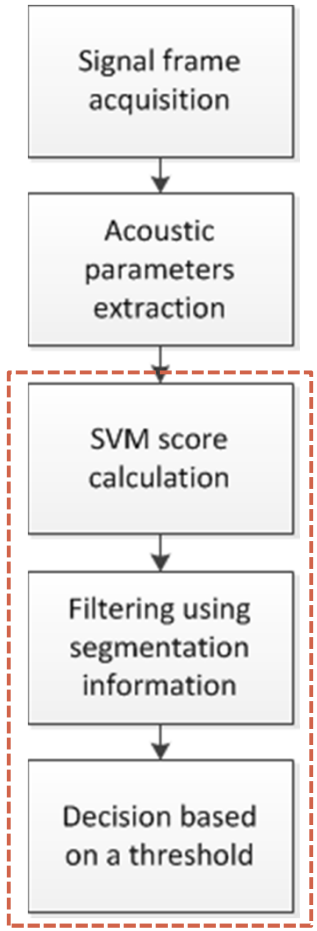
	Number of tests
Hammers	4
Fire	3
Fire-Burst	3
Sirens	3
Crowd-Fighting	3
Dogs	4
Crowd-Booing	3
Fire-Works	3
Crowd-Bravo	3
Explosions	4
Crowd-Applause	3
Telephones	4
Crowd-Angry	3
Party-Music	4
Wood	9
Cheering	4
Screams	4
Children	4
Crowd-Cheering	3
Excavation	4
Applause	4
Fight	4
Doors	5
Earthquake	6
Fire-House	4
Crash	2
Foot-step	3
Glass-Debris	4
Baby	4
Hit-Objects	4
All Events	115

Number of ambience files for training	6 - 1h
Number of ambience files for testing	6
Duration of each ambience file (in min.)	10
Number of SNR conditions (10,15,20,25,30 dB)	5
Duration of single audio event (in sec.)	1
Number of audio events per ambience file	50
Total duration of tested audio events (in sec.)	28750
Total duration of tested audio events (in hours)	8h

GMM for Audio Abnormal Events Detection (Evaluation)



- One-Class SVM choice justification:
  - One-Class aims to define boundaries of a class
    - adapted to unsupervised ambience modeling



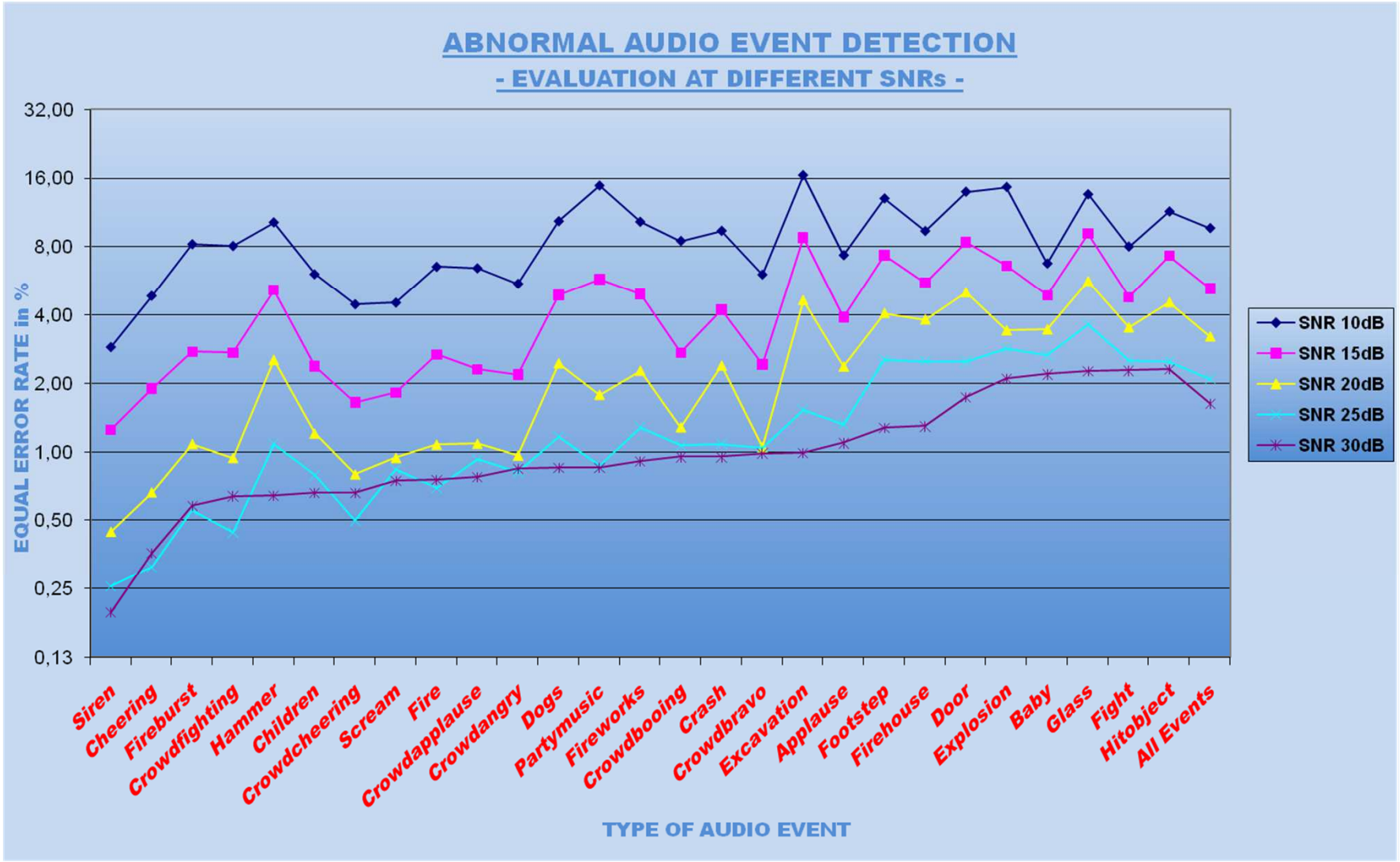
- Detection score:
  - A raw score is computed for each frame
  - Then scores are integrated (averaged) over segments as a smooth filter to get the score (as in GMM based system)
  - We apply a threshold on the score for final decision

OC-SVM for Audio Abnormal Events Detection (Evaluation)

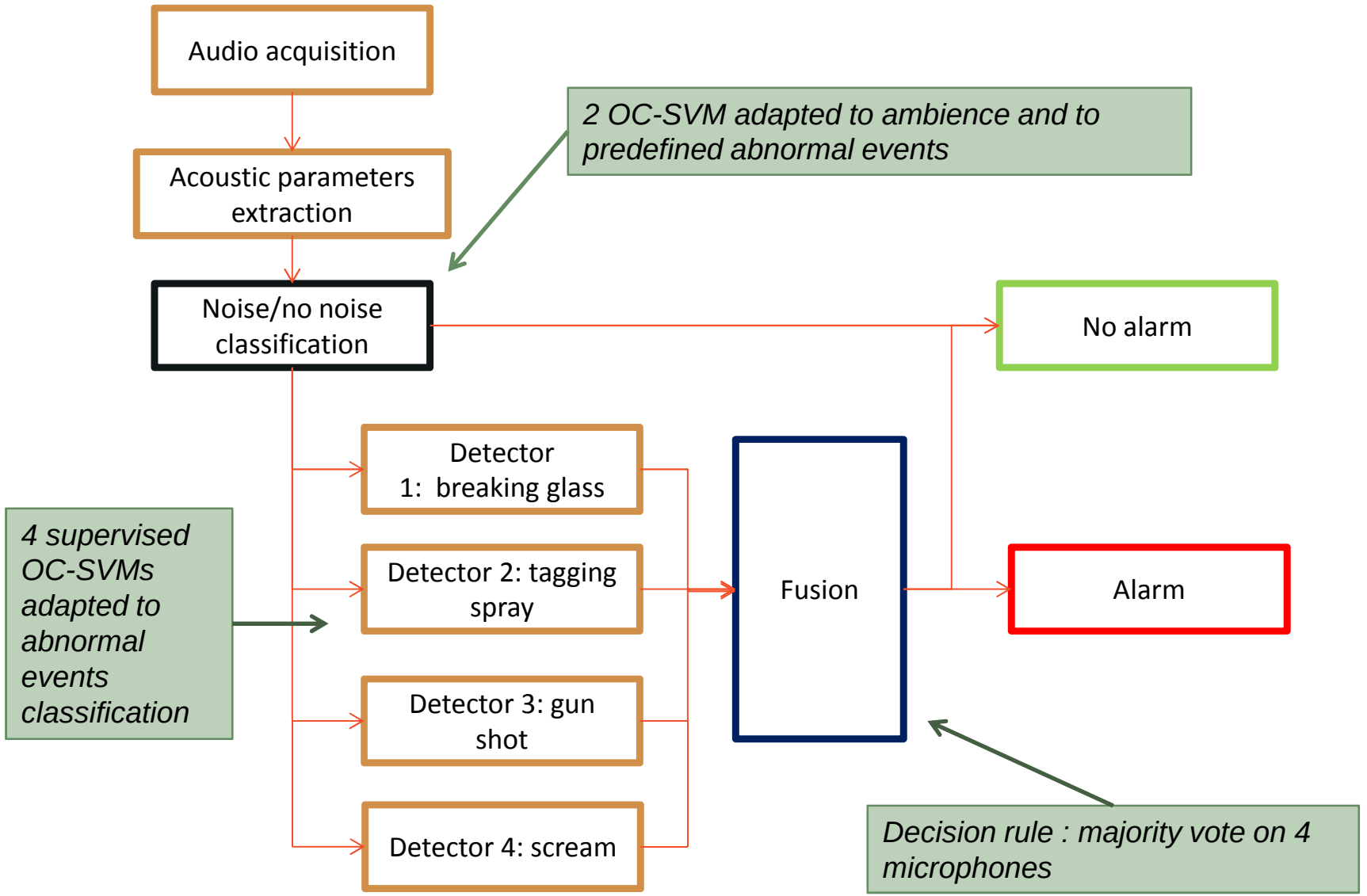
Siren	3
Cheering	4
Fireburst	3
Crowdfighting	3
Hammer	4
Children	4
Crowdcheering	3
Scream	4
Fire	3
Crowdapplause	3
Crowdangry	3
Dogs	4
Partymusic	4
Fireworks	3
Crowdbooing	3
Crash	2
Crowdbravo	3
Excavation	4
Applause	4
Footstep	3
Firehouse	4
Door	5
Explosion	4
Baby	4
Glass	4
Fight	4
Hitobject	4
All Events	96

Number of ambience files for training	6 – 1h
Number of ambience files for testing	12
Duration of ambience files (in min.)	10
Number of SNR conditions (10,15,20,25,30 dB)	5
Duration of audio event (in sec.)	1
Number of audio events per ambience file	50
Total duration of tests (in sec.)	24000
Total duration of tests (in hours)	7h

OC-SVM for Audio Abnormal Events Detection (Evaluation)



OC-SVM based Supervised detection/classification of Audio Abnormal Events



OC-SVM based Supervised detection/classification of Audio Abnormal Events (Evaluation)

- Evaluation description
  - Acted scenes for Tagging(spray) and Screaming
  - Mixed audio signals for breaking glass and gunshot (refer to evaluation protocol presentation)
- Audio material (TESS and EVAS French Funded Project – IFFSTAR Studies)
  - Ambience : 15h
  - Scream (116 - 2s for each)
  - Breaking glass (91 - 1,5 sec)
  - Gunshot (45 - < 0,5 sec)
  - SNR : from 10dB to 20dB (realistic SNR in operational cases)
  - Training 40% of DB – Test/Evaluation 60% of DB

	Pfa	Pdet
Broken Glass	<1%	98%
Gunshot	<1%	97%
Scream/Shout	3%	92%
Tagging (spray)	2%	98%