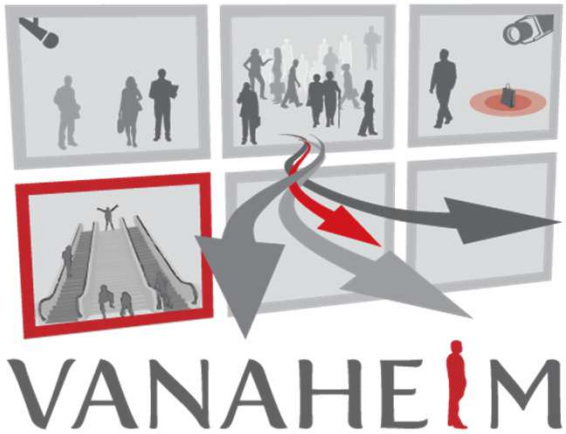


AUDIO TOPIC MODELING

François CAPMAN, Sébastien LECOMTE and Bertrand RAVERA (TCF)

Plan

Unsupervised Audio Analysis/Structuring based on PLSA and applications to

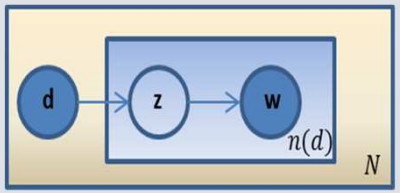
- **Surveillance Systems**
- **Speech segmentation**

PLSA model presentation

- Probabilistic Latent Semantic Analysis (PLSA – Hoffman 2000)
  - Fundamental idea of probabilistic topic models such as PLSA
    - A document is a **mixture of topics/concepts** where a topic is a **distribution of words**
  - Topic model is a **generative model for document** (definition of a statistical process to generate documents from words)

- Topic model is a **generative model for document** (definition of a statistical process to generate documents from words)

$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = P(d_j)\sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j)$$



$P(w_i|z_k)$ is the probability describing how **topic refer to word** (**model parameter : word-topic distribution**)

$P(z_k|d_j)$ is the probability describing how **document refer to topic** (**model parameter : topic-document distribution**)

$P(d_j)$ is the prior probability to pick a document

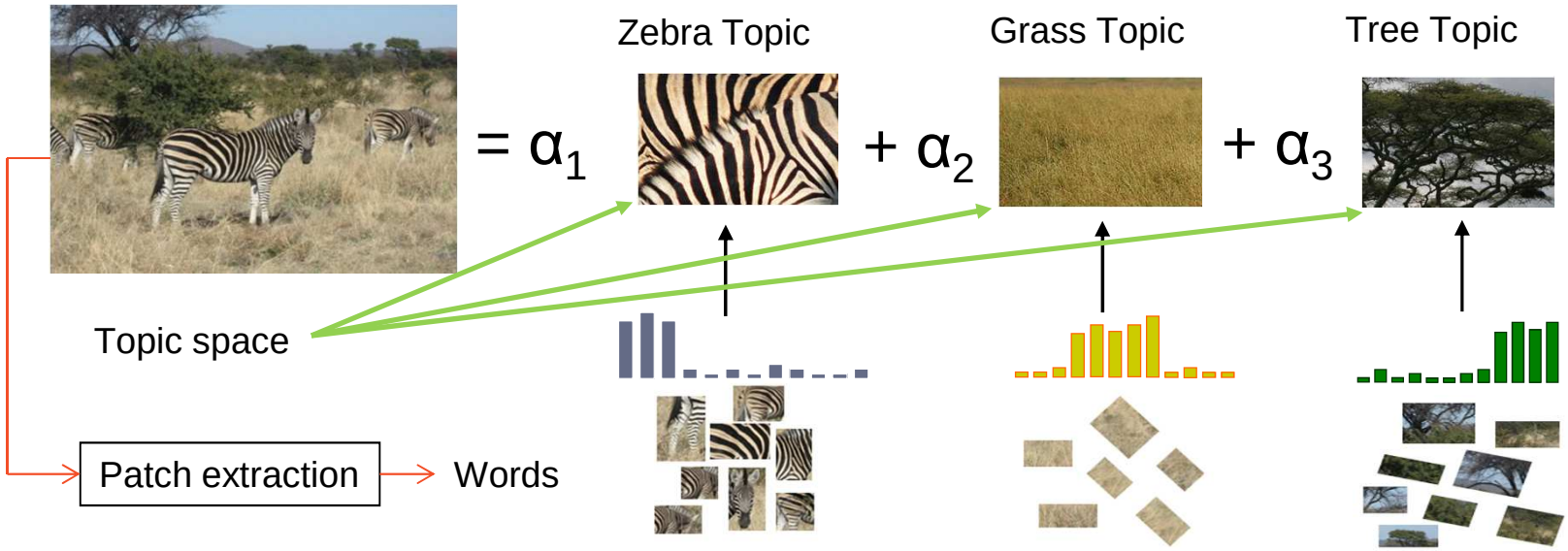
$P(w_i, d_j)$ is the joint probability to have a word in a document

- The parameters of the PLSA model are calculated using EM algorithm by maximizing the log-likelihood of the PLSA model over of training document D_{train}

$$L(P|D_{train}) = \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i, d_j)\log P(w_i, d_j)$$

$n(w_i, d_j)$ gives how often the word w_i occurs in a document d_j (**co-occurrence matrix**)

Basic PLSA model presentation



- PLSA and Text processing (Hoffman 2000)
  - Document structuration in topics, classification, retrieval, query on text DB, summarization, …
- PLSA and image/video processing
  - Content analysis, Image classification, Image retrieval, Query on image DB, …..
  - **Abnormality detection** with a trained PLSA model and documents Log Likelihood (IDIAP, QMUL studies, …..)
- PLSA and audio processing
  - Adaptation of PLSA tools for audio processing : Audio content analysis, audio classification (music classification, retrieval, indexation, …), **audio stream temporal segmentation, abnormal event detection** (audio based surveillance systems), …

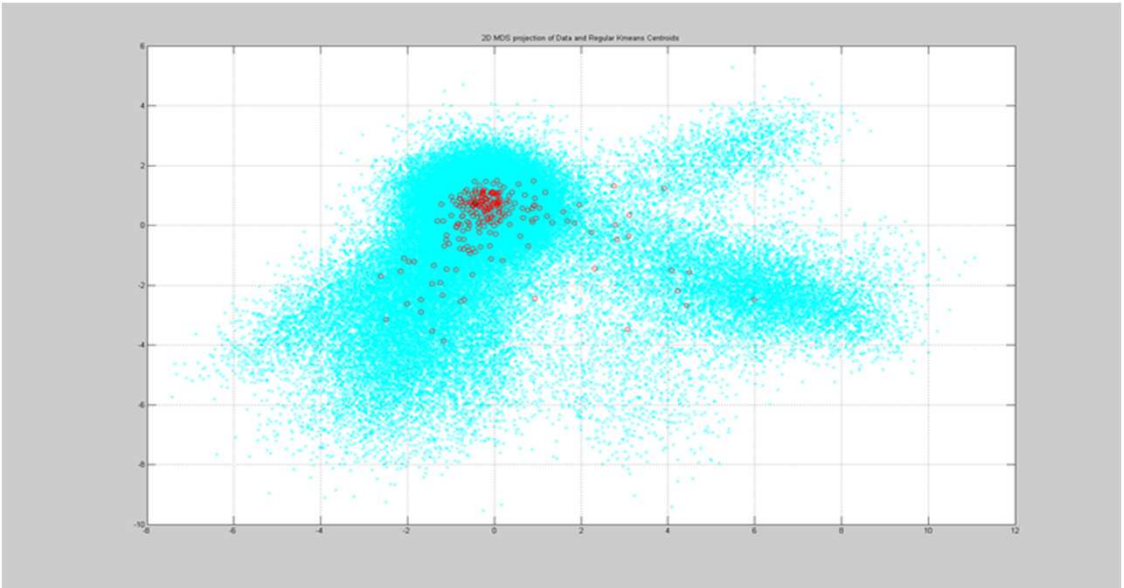
Audio words set building (unsupervised clustering)

- **Unsupervised Acoustic space clustering**
  - Acoustic features (refer to previous presentation – for this study Linear frequency Sub-band Energies LFSBE)

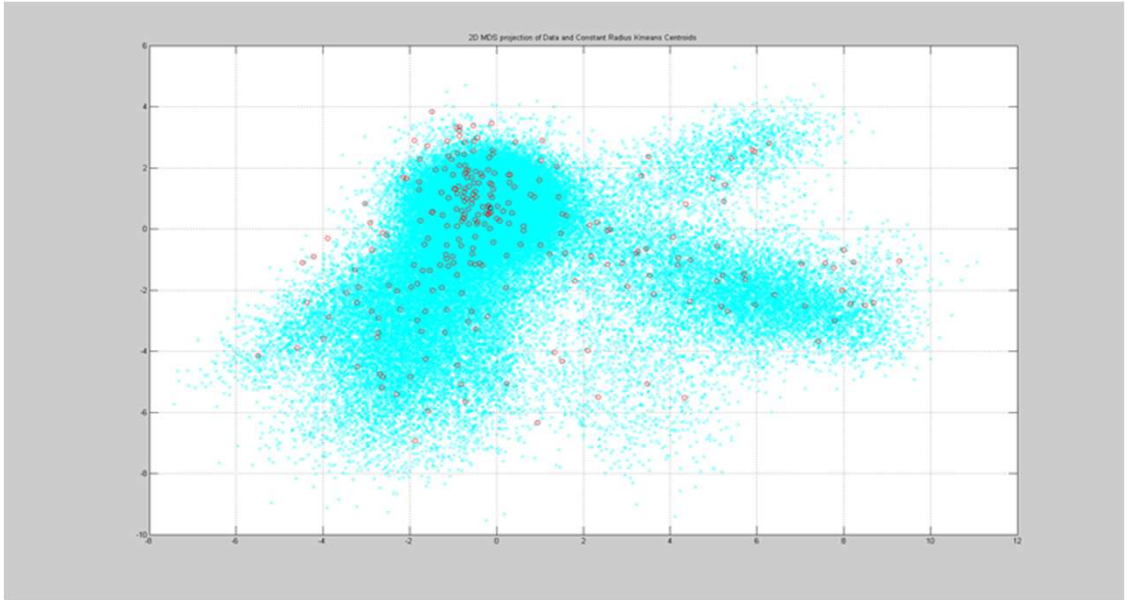
- **Possible approach**
  - Well known K-means algorithms (widely used in speech/audio compression)
  - K-means suffers from well-known drawbacks
    - K-means solution depends on **initialisation**
    - **Acoustic space topology (or data topology) not well maintained after clustering**

- **Proposed approach**
  - K-means with constrained clusters volume and centroid trajectories monitoring (algorithm not presented here)
  - **The goal is to obtain not an optimal clustering driven by distortion minimization but an optimal representation of data set in terms of audio vocabulary coverage**

Constrained radius unsupervised clustering

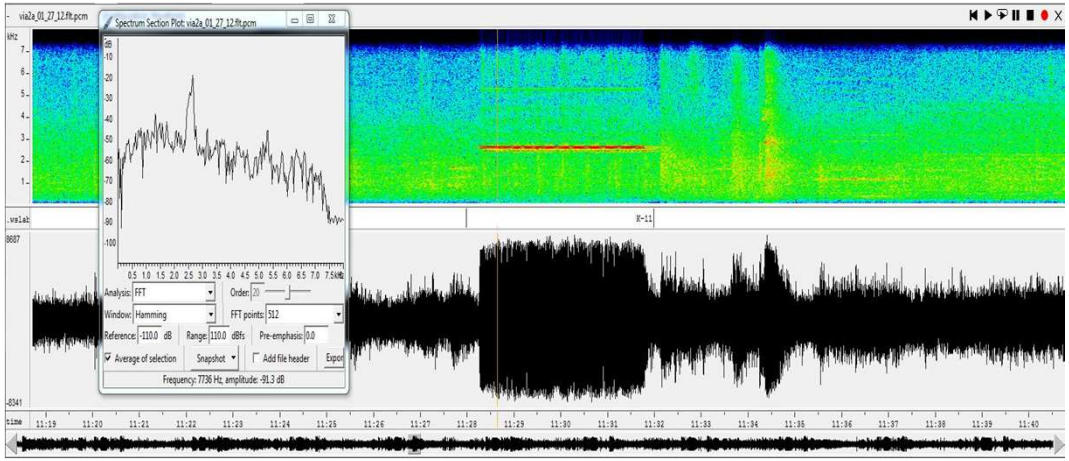


Regular K-means

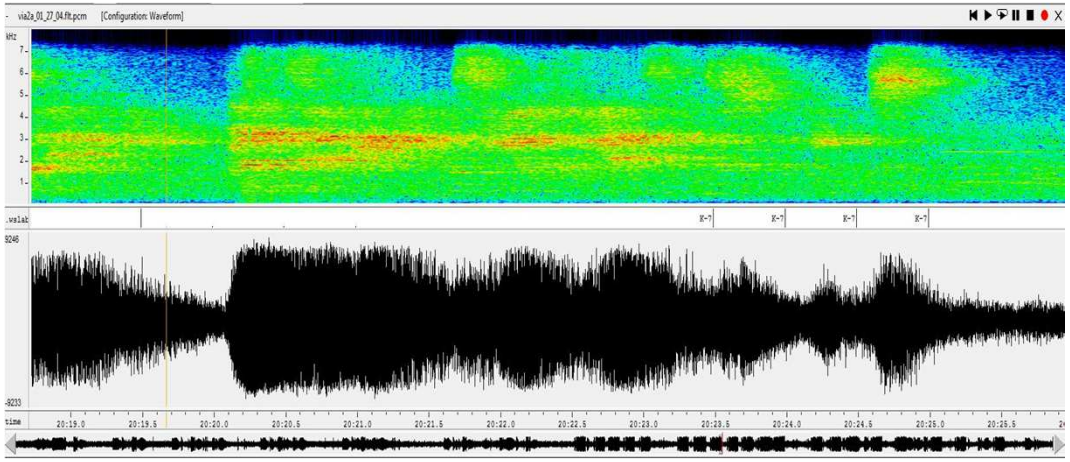


Constrained radius unsupervised clustering

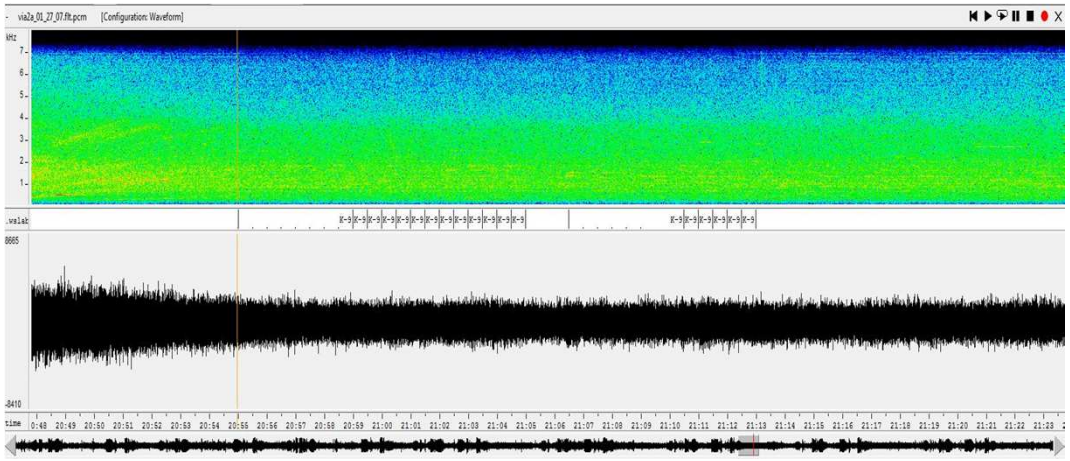
- **Analysis of topics signification with audio event semantic (Topic 11/15)**
  - Audio significations
    - One frequency tone (with HF components) corresponding to train's doors opening and closing announcement
    - Very soft discussions
    - No security announcement
    - No Ambiance Music
  - Topic semantic
    - Type of train's doors opening and closing announcement



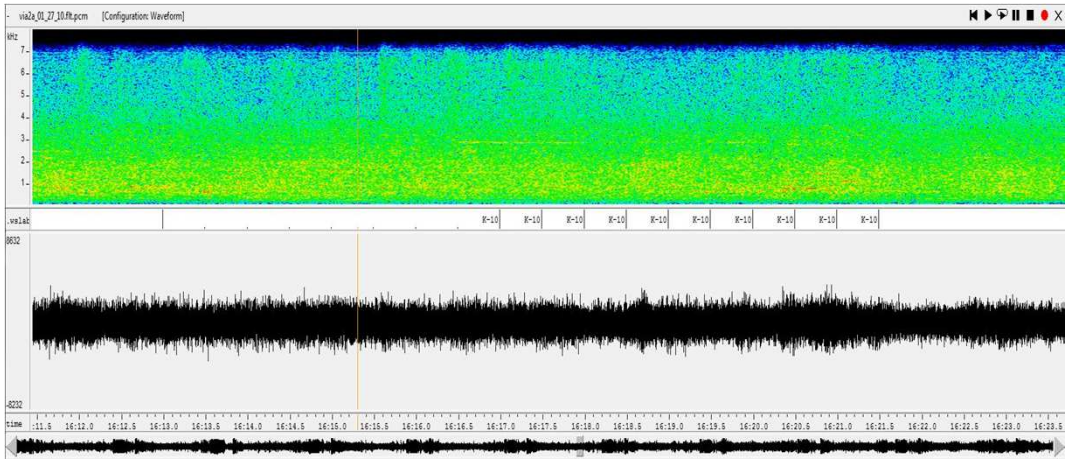
- **Analysis of topics signification with audio event semantic (Topic 7/15)**
  - Audio significations
    - No Trains
    - Very soft discussions
    - High and saturated security announcement
    - No Ambiance Music
  - Topic semantic
    - Type of security announcement



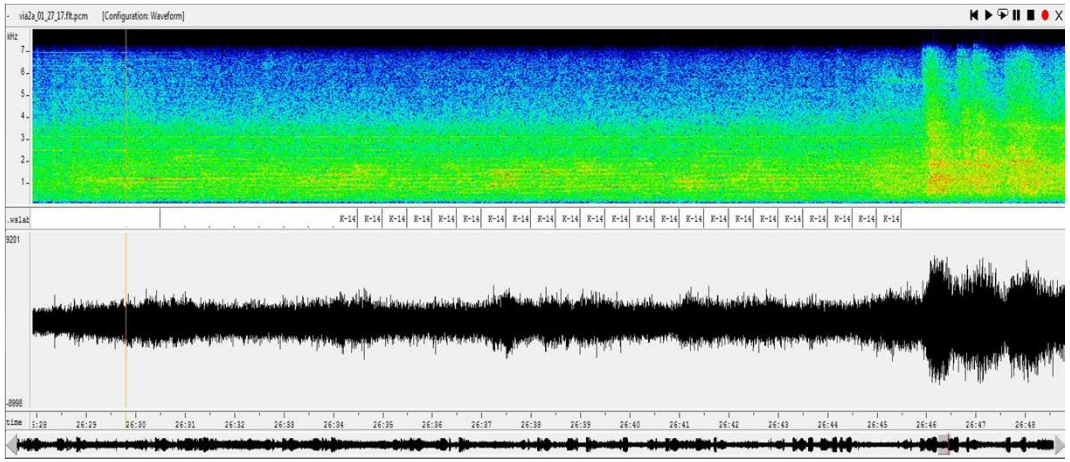
- **Analysis of topics signification with audio event semantic (Topic 9/15)**
  - Audio significations
    - No Trains,
    - No discussions,
    - No security announcement,
    - Very few Ambiance Music
  - Topic semantic
    - Type of very quiet ambiance without train



- **Analysis of topics signification with audio event semantic (Topic 10/15)**
  - Audio significations
    - No Trains
    - Discussions (several groups on the platform)
    - No security announcement
    - No Ambiance Music
  - Topic semantic
    - Type of quiet ambiance with persons on the platform but without train

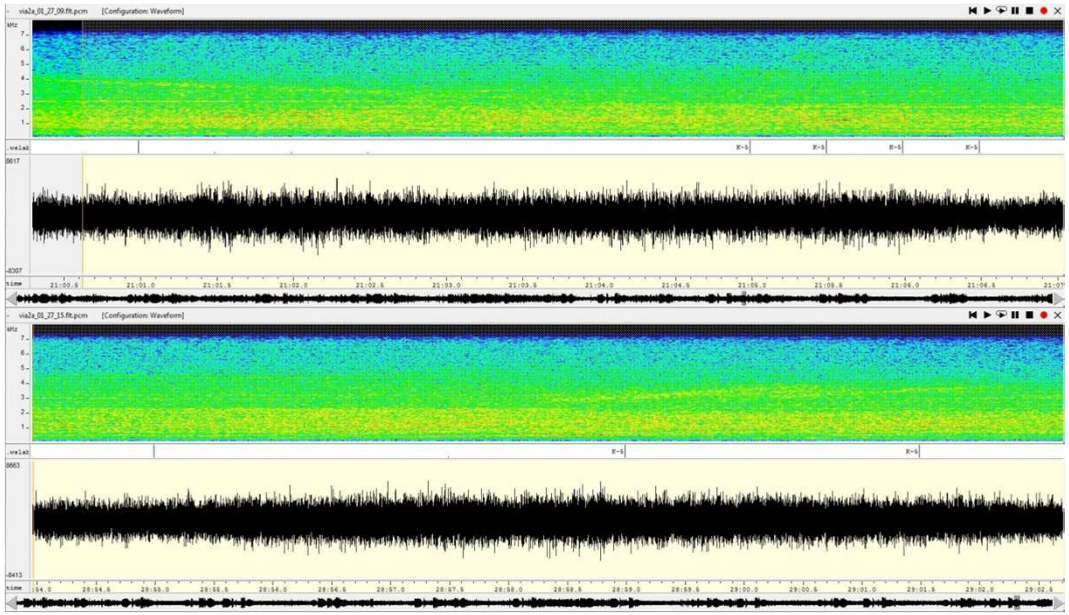


- **Analysis of topics signification with audio event semantic (Topic 14/15)**
  - Audio significations
    - No Trains
    - No discussions
    - No security announcement
    - Very high Ambiance Music (singing voices)
  - Topic semantic
    - Type of ambiance without train



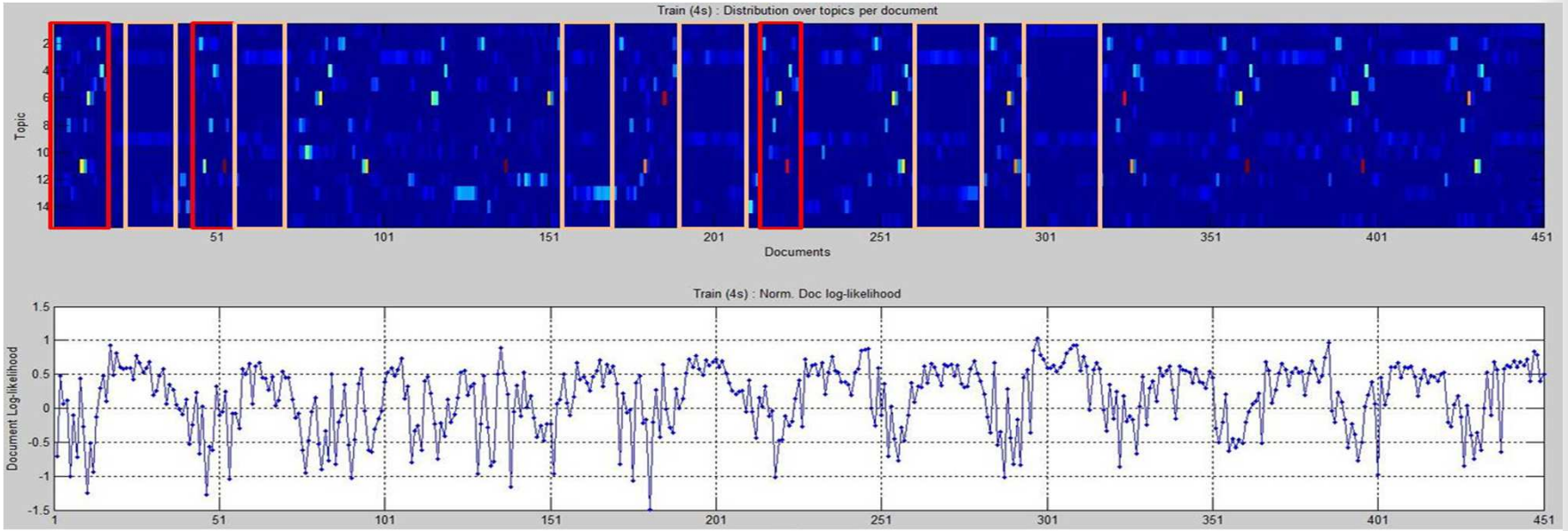
PLSA intermediary results (150 words, k=15, Ds=4s)

- **Analysis of topics signification with audio event semantic (Topic 5/15)**
  - Audio significations
    - Trains arrival (top) and train departure (bottom)
  - Topic semantic
    - Type of train arrival/departure
  - This topic is very interesting because , there are **no differences in terms of spectral content between arrival and departure patterns**, as clearly shown by the spectrogram. The **only difference is the time organization of these patterns**. Because **PLSA analysis is based on bag of words methods, it doesn't take into account time parameters**. That's the reason why this topic fit well on both train arrival/departure



PLSA intermediary results (150 words, k=15, Ds=4s)

- **Topic-based interpretation of PLSA analysis**



- Topic semantics (topic distributions over documents) is **a good tool to structure** audio signals
- Document Log-Likelihood temporal (DLL) analysis
  - Best documents are not related to mono-modal topic distribution over documents (mixture of topics as expected)
  - Temporal location of topic: we do not have better time location that document temporal parameters (beginning, end and duration) → **Lack of temporal precision**

PLSA adaptation to audio analysis : Delayed Multi Document Size PLSA

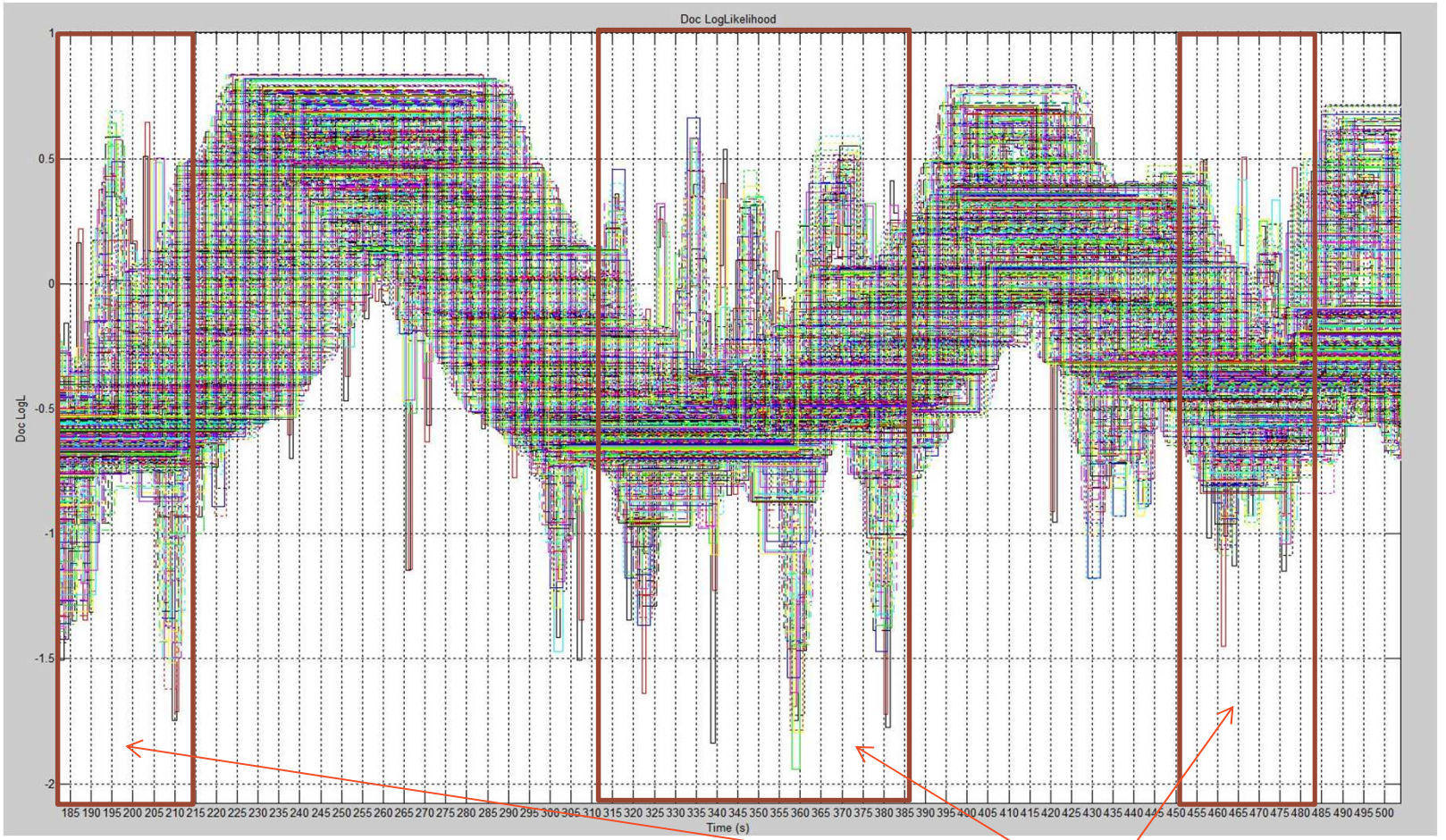
- Motivation (from analysis of PLSA results)
  - **No topic temporal topic localization available** (only document temporal localization)
    - Not enough precise
  - **Document Log-Likelihood** (DLL) values differs when
    - Temporal support of events differs and are not the same as PLSA temporal resolution

    → **Best DLL are obtained when PLSA temporal resolution is close to event duration**

- Adaptation of PLSA (2 key issues)
  - PLSA with **variable document size**
    - We expect to obtain with PLSA a collection of models well fitting a large audio event duration
      - from short event as impulsive event (door opening, …..)
      - To long event as audio ambiance between train arrival/departure
  - PLSA with **variable document size according delayed analysis schema**
    - We expect to obtain with PLSA **a collection of models well fitting a large audio event duration and with fine temporal localization**

Delayed Multi Document Size PLSA

- Delayed Multi document size log-likelihood analysis (doc. Size [1s ; 60s], delay 0,5s)



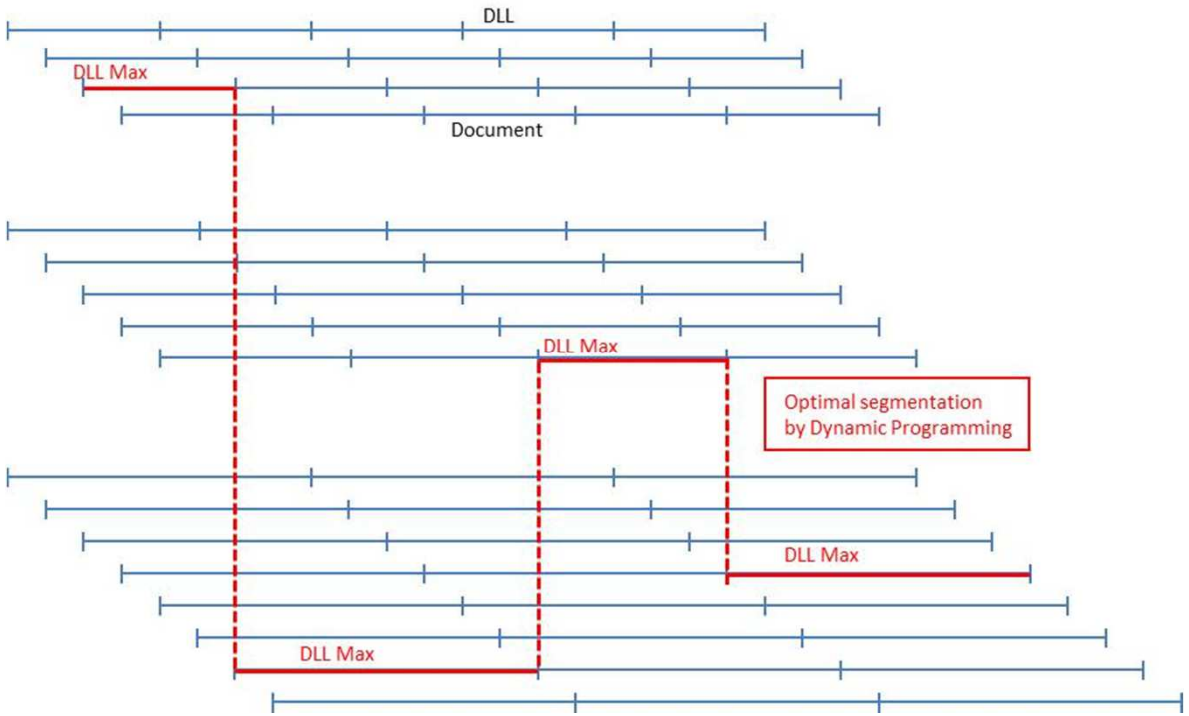
- When document size is adapted to audio event duration
  - High DDL occurs with well fitted PLSA models (Short event appears well inside long event)

→ **Find optimal document sizes (longer ones with high DLL : optimal document search under constraints)**

Delayed Multi Document Size PLSA

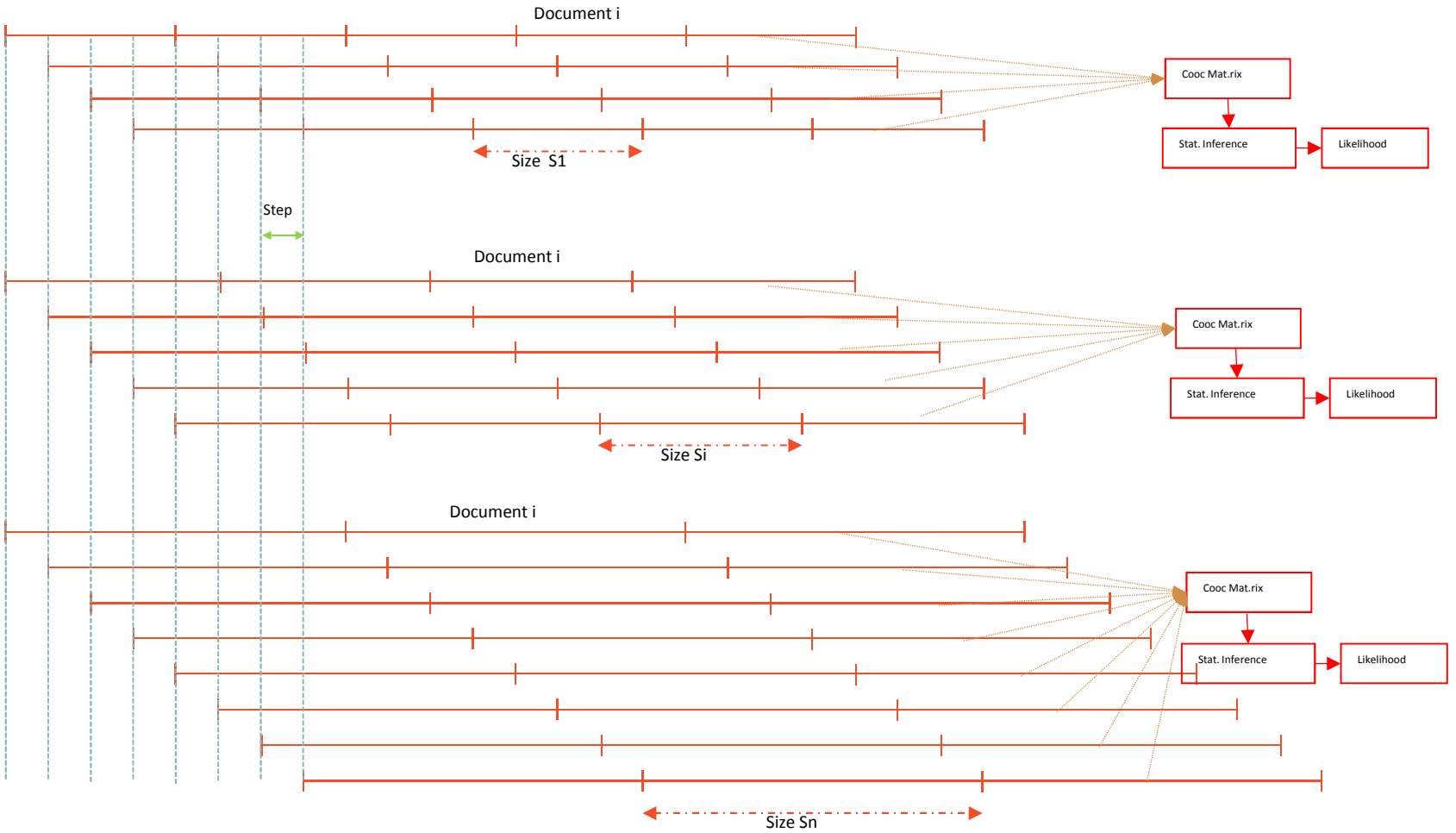
- Unsupervised topic-based audio structuration
  - Optimal search, based on the Document log-likelihood (DDL maximum), of the non-overlapping documents
  - Optimal segmentation or structuration obtained with **Dynamic Programming (DP) tool - <u>Weighted Interval Scheduling (WIS)</u>**

→ This segmentation has been developed to be robust against local statistical variations and by the way can be more easily understood by surveillance operators

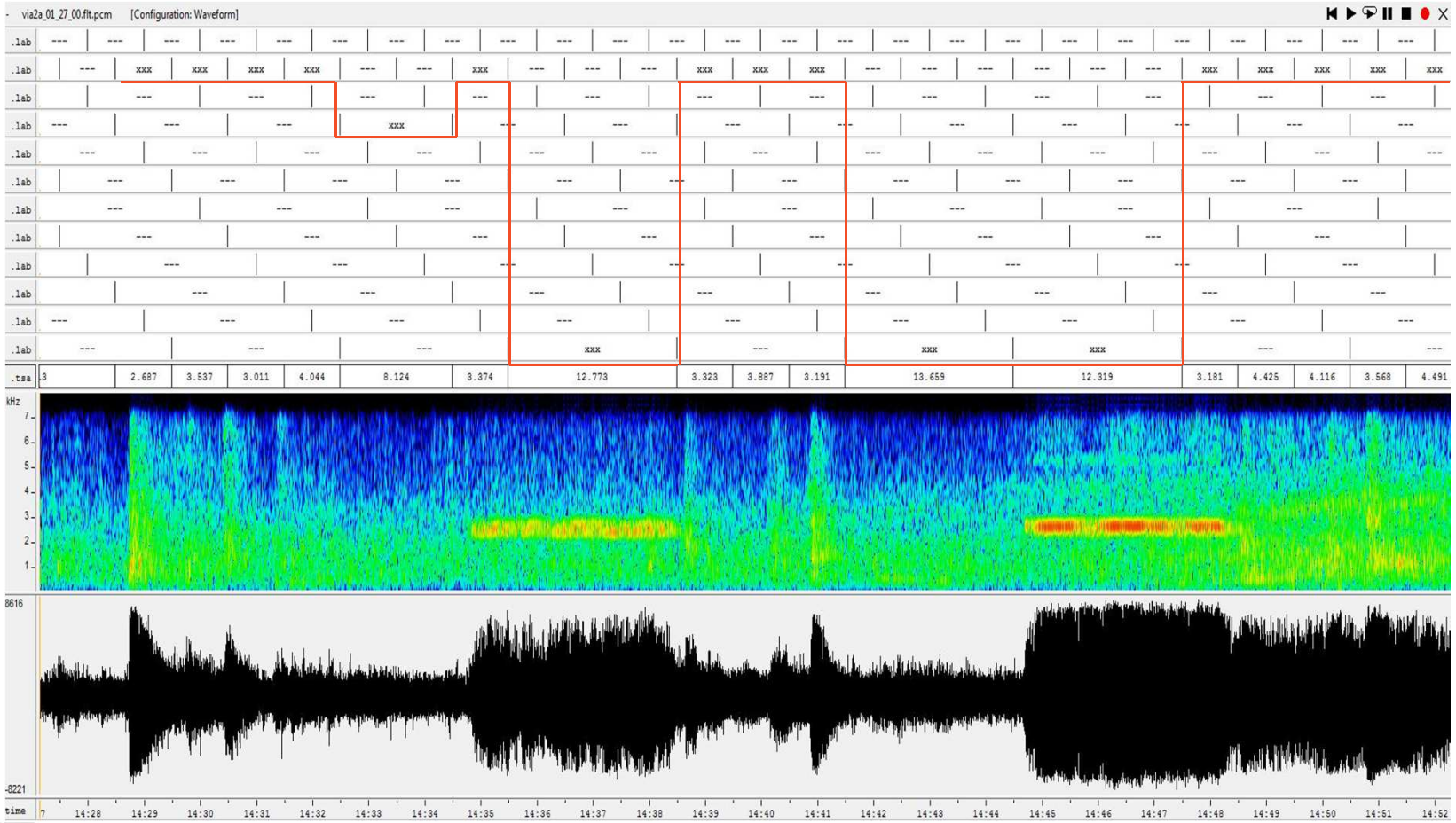


Delayed Multi Document Size PLSA

- Delayed Multi document size PLSA schema (common to training and test phases)

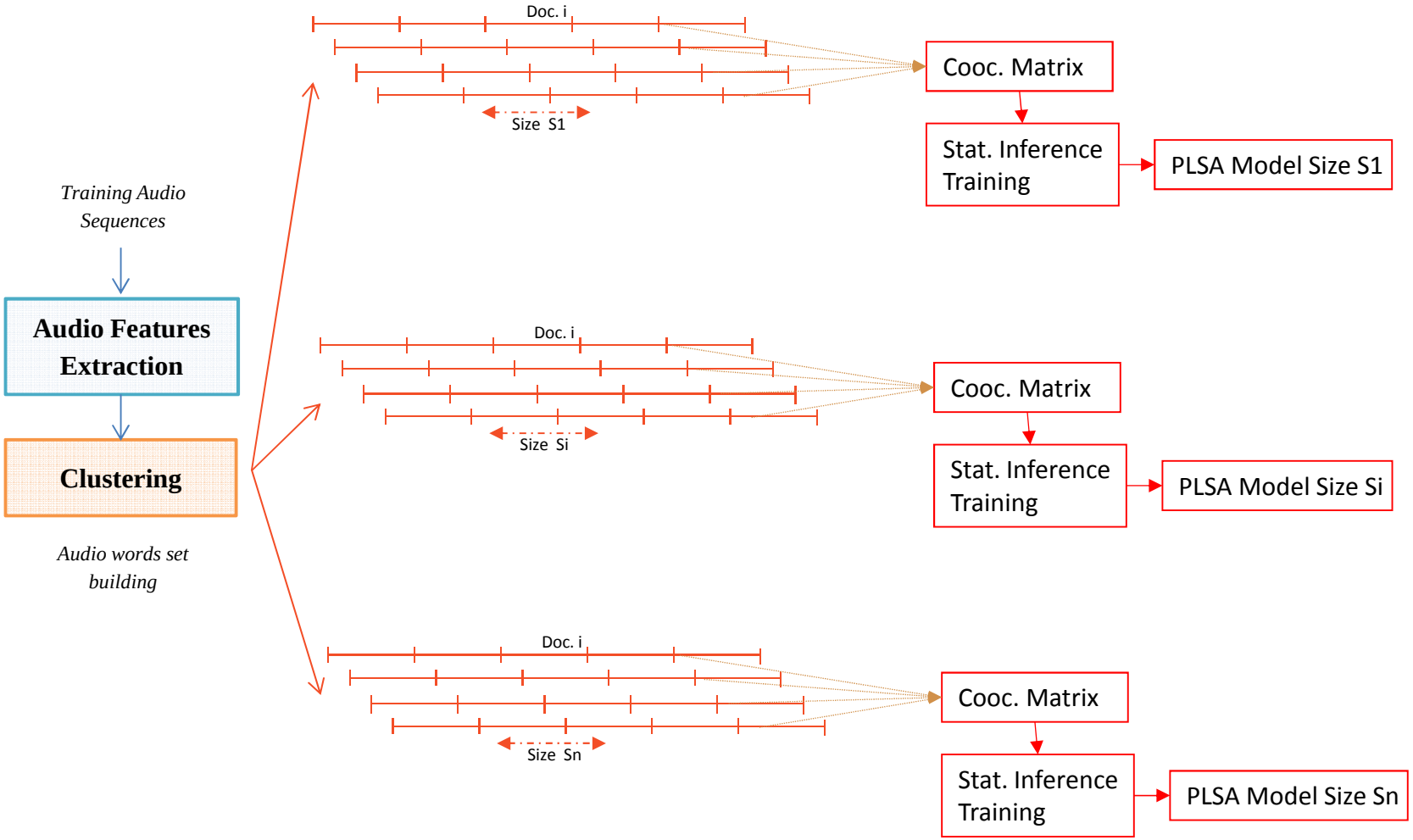


- Unsupervised topic-based audio structuration results
  - Audio data : signal collected on 27th of January , Document size 1s,2s and 3s delayed by 0,5 s
    - Label with “xxx” shows part of the optimum WIS path and label with “---“shows documents which don’t belong to optimum scheduling or optimum segmentation. Short events are well segmented, such as Frequency tones and Impulsive events (door opening/closing).



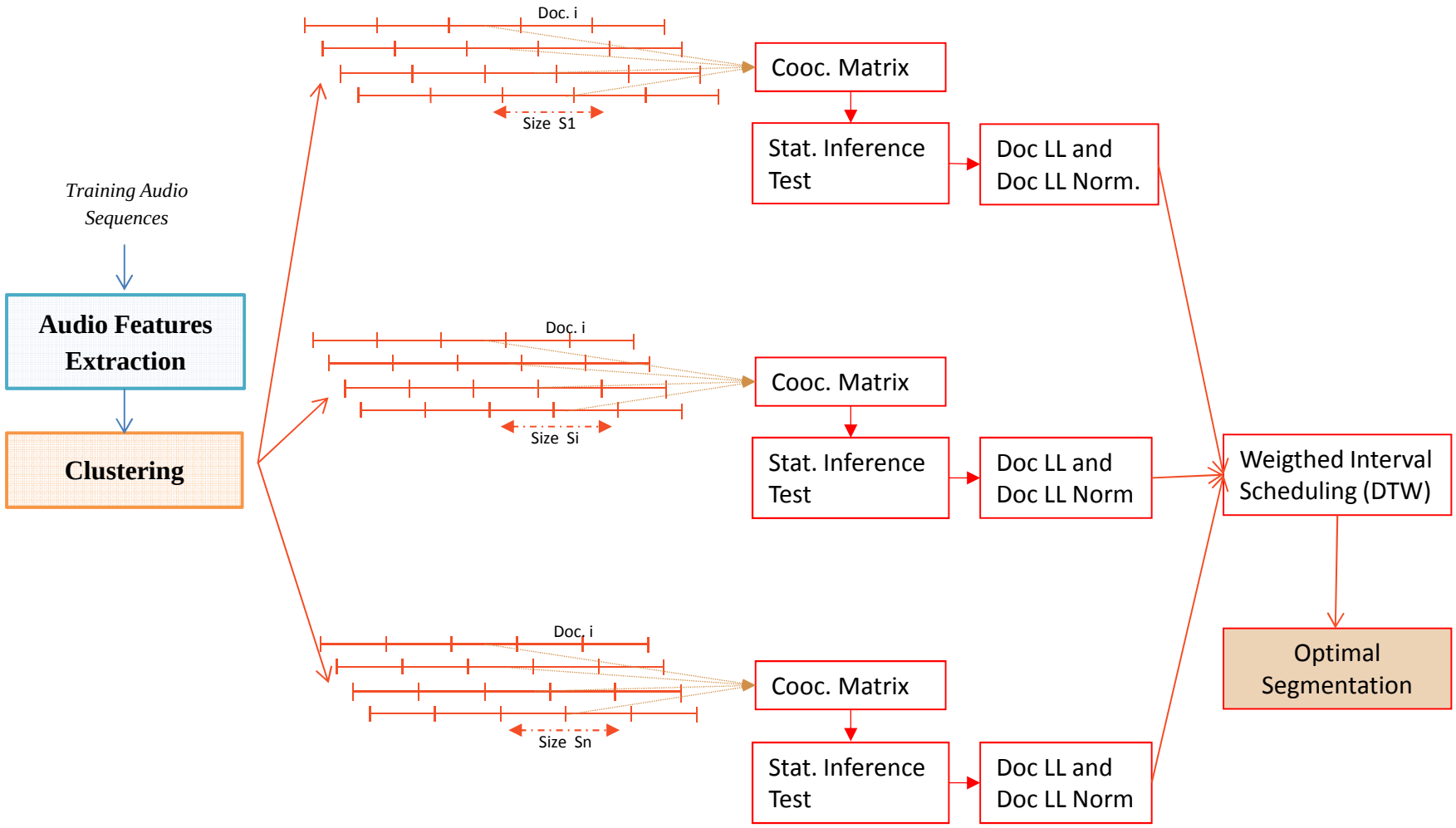
Delayed Multi Document Size PLSA

(a) TRAINING PHASE

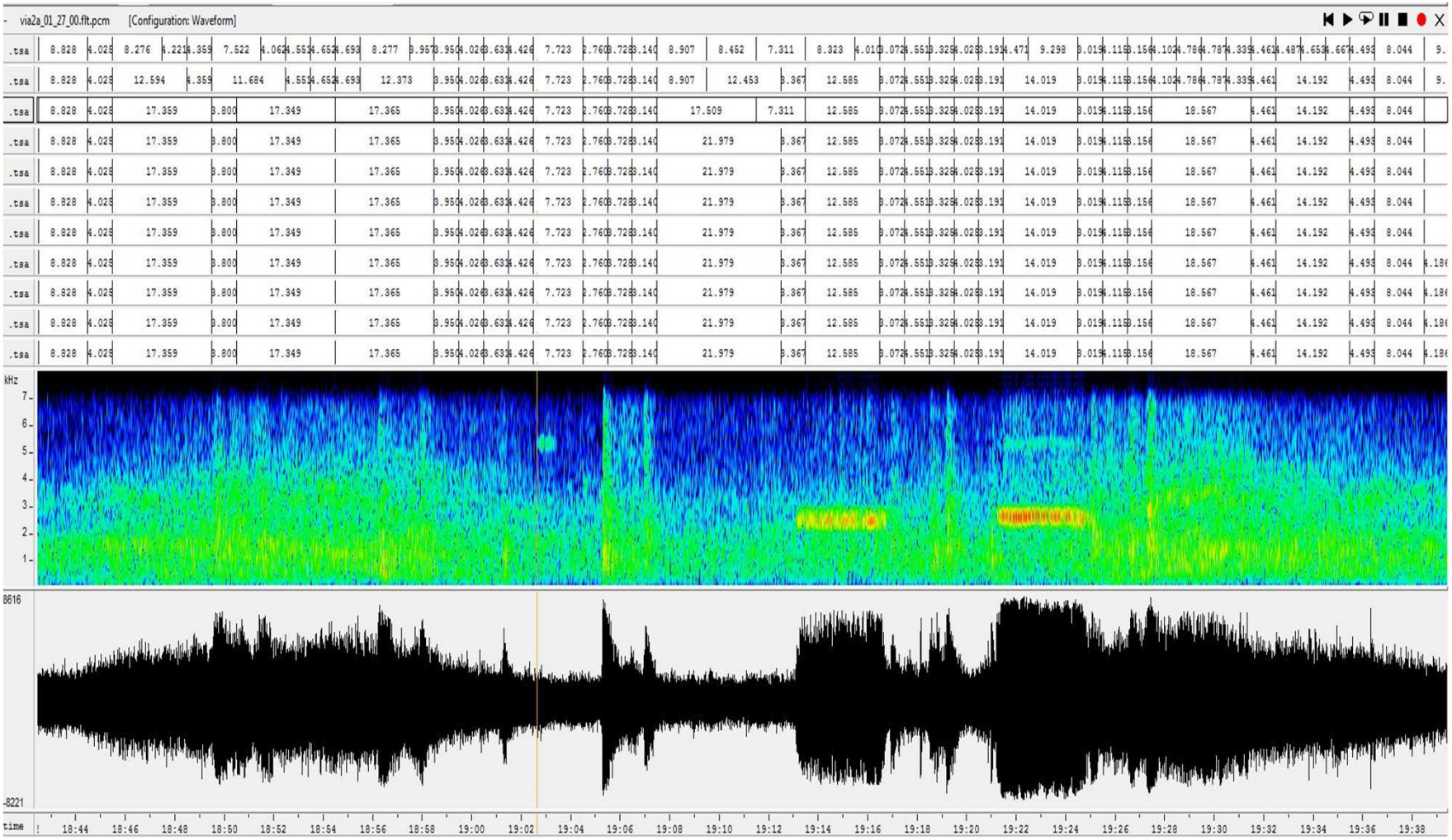


Delayed Multi Document Size PLSA

(b) TEST PHASE (SEGMENTATION)



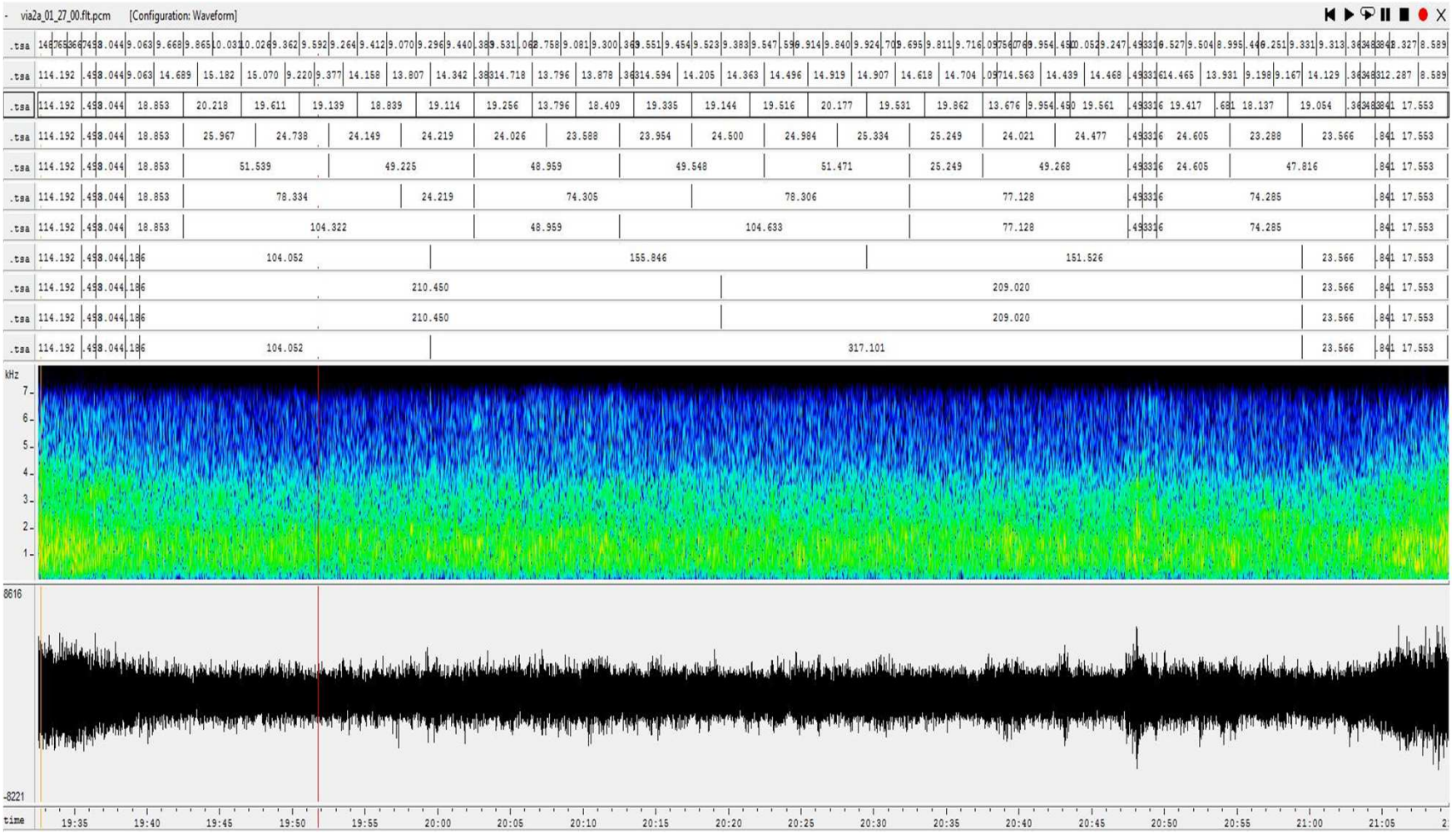
- Unsupervised topic-based audio structuration results (full analysis)



Topic-based audio segmentation results (train arrival and departure)

Document size 1s, 2s, 3s, 4s, 5s, 10s, 15s, 20s, 30s, 40s, 50s and 60s delayed by 0,5s

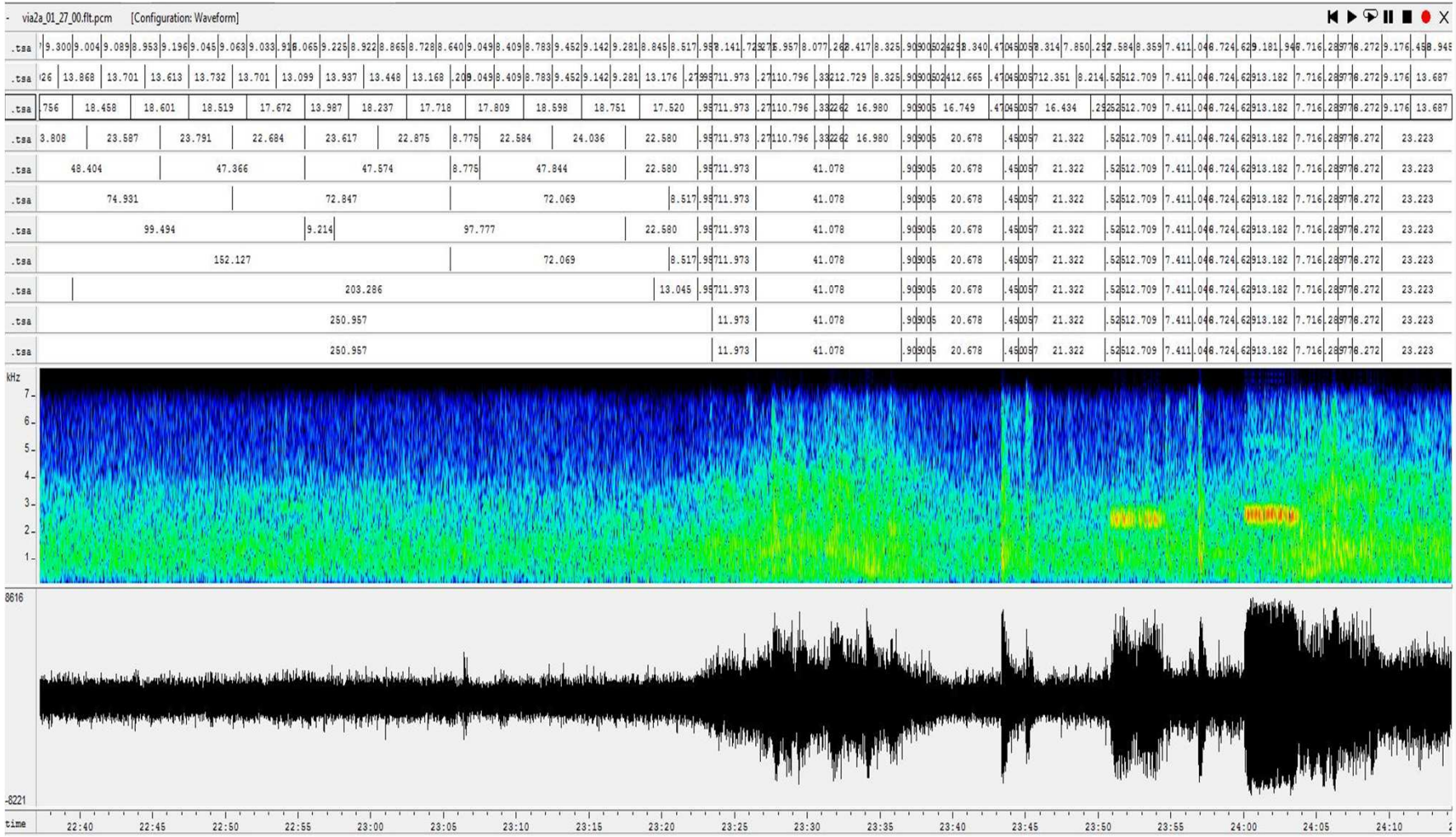
- Unsupervised topic-based audio structuration results (full analysis)



Topic-based audio segmentation results (station ambiance between 2 trains)

Document size 1s, 2s, 3s, 4s, 5s, 10s, 15s, 20s, 30s, 40s, 50s and 60s delayed by 0,5s

- Unsupervised topic-based audio structuration results (full analysis)



Topic-based audio segmentation results (station ambiance and then train arrival) Document size 1s, 2s, 3s, 4s, 5s, 10s, 15s, 20s, 30s, 40s, 50s and 60s delayed by 0,5s

Motivations

- The main problem of our studies related to unsupervised audio stream segmentation is the **performance evaluation**
  - **No databases easily available** (as least from my own point of view).
  - **No ground truth available**.
  - Database annotation is time consuming and requires several human runs to converge to a final usable manual annotation.

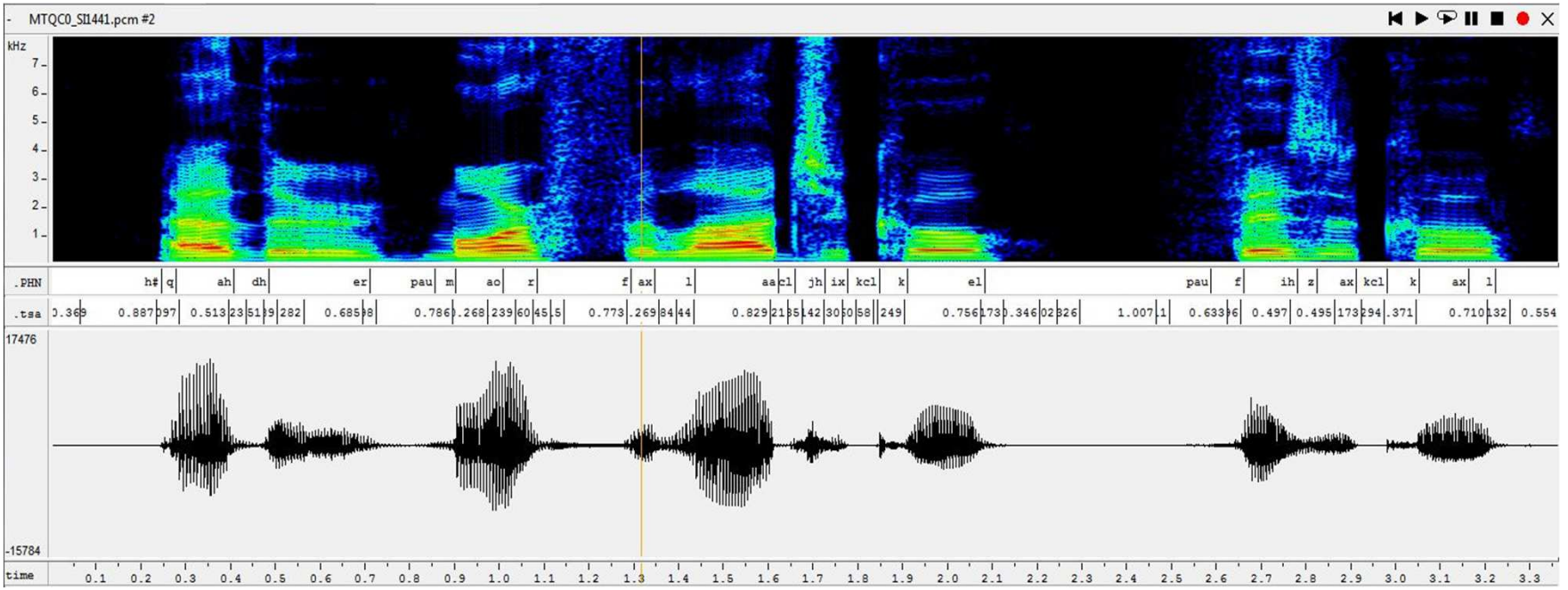
- We need an evaluation on **well recognized database**
  - **TIMIT database** (4620 speakers for training and 1680 speakers for test including male and female)

- **New addressed task : "Full unsupervised Phoneme Boundaries Identification"** with the following parameters
  - 3 specific audio words dedicated to silence (adapted audio clustering)
  - Sampling Freq. 16Khz, Temporal windows size 20 ms , temporal shift 10ms, 24 LFSBE, 64 audio words (with 3 words for silence)
  - 75 topics related to 60 phonemes and silences
  - Document size : 30 ms to 170 ms (delay 10ms)

Delayed Multi Document Size PLSA (Speech analysis)

- **Results**
  - Measure interval : +- 20 ms

	P_det	P_over_seg (including pause and silence)
Train	82%	16%
Test	81%	18%



- **Results Analysis**
  - Good results for an unsupervised method
  - High over-segmentation rate
    - Topic based analysis is mainly driven by spectral content
  - **Topic semantic is strongly related to phoneme classes (added value of this approach)**