

AUDIO TOPIC MODELING

François CAPMAN, Sébastien LECOMTE and Bertrand RAVERA (TCF)





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

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 <u>topic refer to word</u> (model parameter : word-topic distribution) $P(z_k|d_j)$ is the probability describing how <u>document refer to topic</u> (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 <u>EM algorithm</u> 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 (<u>co-occurrence matrix</u>)





- 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), ...





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 <u>optimal representation of data set in terms of audio</u> <u>vocabulary coverage</u>



Constrained radius unsupervised clustering







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

| wasa_or_st_orunbem (confidencies wasajouu) | |
|--|--|
| 7- 6- 5- 4- | |
| | |
| | |
| fational de la cale de la contra fate an de contra fate an de contra de la contra de la contra de la contra de | and the weath of the second barrent the contraction from a second by the first second of the plant of the second of the second second by the second particular terms of the second particular terms of the second particular terms of the second s |







- > 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) \triangleright
 - Audio significations
 - \triangleright No Trains
 - No discussions \triangleright
 - No security announcement \geq
 - Very high Ambiance Music (singing voices)
 - > Topic semantic
 - Type of ambiance without train







- 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







> Topic-based interpretation of PLSA analysis



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



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

\rightarrow 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 log-likelihood analysis (doc. Size [1s; 60s], delay 0,5s) \geq



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



 \rightarrow Find optimal document sizes (longer ones with high DLL : optimal document search under constraints) 15



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

 \rightarrow 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 schema (common to training and test phases)







Unsupervised topic-based audio structuration results

VANAHE

- > 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







PROGRAMM

Delayed Multi Document Size PLSA







PROGRAMM

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

VANAHE

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

via2a_01_27_00.flt.pcm [Configuration: Waveform] 1 tafrædædrada . 0449.. 0539. 6669.. 8650.. 0320.. 02d9.. 3629.. 2649.. 4529.. 2649.. 4219.. 2010.. 2349.. 4519.. 3019.. 3019.. 3019.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35219.. 3649.. 35210.. 3 .tsa 114.192 .493.044 18.853 13.676 9.954.450 19.561 493316 19.417 .681 18.137 .363483841 17.553 20.218 19.611 19.139 18.839 19.114 19.256 13.796 18.409 19.335 19.144 19.516 20.177 19.531 19.862 19.054 .tsa 114.192 .498.044 18.853 493316 24.605 25.967 24.738 24.149 24.219 24.026 23.588 23.954 24.500 24.984 25.334 25.249 24.021 24.477 23.288 23.566 .841 17.553 tsa 114.192 .498.044 18.853 493316 24.605 .841 17.553 51.539 49.225 48.959 49.548 51.471 25.249 49.268 47.816 493316 .841 17.553 tsa 114.192 .498.044 18.853 78.334 24.219 74.305 78.306 77.128 74.285 .tsa 114.192 .498.044 18.853 493316 841 17.553 104.322 48.959 104.633 77.128 74.285 .tsa 114.192 .498.044 186 104.052 155.846 151.526 23.566 .841 17.553 .tsa 114.192 .498.044 186 210.450 209.020 .841 17.553 23.566 .tsa 114.192 .498.044.186 841 17.553 210.450 209.020 23,566 .tsa 114.192 .498.044.186 .841 17.553 104.052 317.101 23.566

8616 -8221 time 19:35 19:40 19:50 19:55 20:00 20:05 20:10 20:15 20:20 20:25 20:30 20:35 20:45 19:45 20:40 20:50 20:55 21:00 21:05

VANAHEIM

6.

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)

via2a 01 27 00.flt.pcm [Configuration: Waveform] 19. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. 0.00 |9. . 13. 262 13. 868 13. 701 13. 613 13. 732 13. 701 13. 039 13. 937 13. 448 13. 168 209. 0.498. 4098. 4098. 439. 4429. 1429. 281 13. 176 2 499711. 973 2 410. 796 2 34212. 729 8. 325 9 400400412. 665 4 404800412. 351 8. 214 5 3512. 709 7. 411, 048. 724 6 2913. 182 7. 716 2 347 48. 279. 176 13. 687 17.520 .94711.973 .27110.796 .342242 16.980 .949045 16.749 .470450057 16.434 .24252512.709 7.411.046.724 62913.182 7.716 .249746.272 9.176 13.687 .tsa .756 18.237 17.809 18.598 18.751 18.458 18.601 18.519 17.672 13.987 17.718 99711.973 .27110.796 .332262 16.980 .909005 20.678 23.587 23.791 22.684 23.617 22.875 8.775 22.584 24.036 22.580 .450057 21.322 .52512.709 7.411.048.724.62913.182 7.716.289778.272 23.223 .tsa 3.808 .95711.973 .450057 21.322 52512.709 7.411.046.724 62913.182 7.716 289776.272 48.404 47.366 47.574 47.844 22.580 41.078 .909005 20.678 23.223 .tsa 450057 21.322 8.517.99711.973 52512.709 7.411.046.724.62913.182 7.716.289778.272 .909005 20.678 74.931 72.847 72.069 41.078 23.223 .tsa 9.214 22.580 .99711.973 .909005 20.678 .450057 21.322 52512.709 7.411.046.724 62913.182 7.716 289776.272 99.494 97.777 23.223 41.078 .tsa 450057 21.322 52512.709 7.411.048.724 52913.182 7.716.289778.272 8.517.95711.973 . 909005 152.127 72.069 41.078 20.678 23.223 .tsa 52512.709 7.411.046.724 62913.182 7.716 289778.272 13.045 .99711.973 .450057 21.322 203.286 .909005 20.678 23.223 41.078 .tsa . 909005 .450057 52512.709 7.411.046.724 62913.182 7.716 289776.272 11.973 20.678 21.322 250.957 41.078 23.223 .tsa 52512.709 7.411.048.724 62913.182 7.716 289778.272 .909005 20.678 .450057 21.322 250.957 11.973 41.078 23.223 .tsa kHz 7. 6-5-3-HILIDAY. 2-8616 -8221 time 23:50 22:40 22:45 22:55 23:00 23:05 23:10 23:15 23:20 23:25 23:30 23:35 23:40 23:45 23:55 24-00 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)







- 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)



