A Generic System for audio indexing: application to speech/music segmentation and music genre recognition

G. Peeters peeters@ircam.fr

IRCAM (Sound Analysis/Synthesis Team) - CNRS (STMS)
Projet RIAM Ecoute http://projet-ecoute.ircam.fr/

Introduction

- Automatic audio indexing
  - Increasing amount of applications based on audio indexing
    - sound event (musical instrument, sound FX) recognition,
    - music genre/mood,
    - singer genre, speaker recognition
    - speech/music segmentation
    - ...
  - Safe time -> develop a unique generic and modular system for indexing
  - Existing generic systems:
    - WEKA (Waikato University)
    - Extractor Discovery System EDS (Sony CSL Paris)
    - jAudio + ACE (McGill Univeristy)
    - Marsyas (Tzanetakis)
    - M2K (IMIRSEL's)
    - ...

Generic audio indexing system
Requirements

- Two main actions the system must perform:
  - training: a classification model is learned from hand-labeled data
  - indexing: the classification model is used to label (or segment) unknown data
  - <u>-> the two actions must be clearly separated since they are not used by the same people</u>

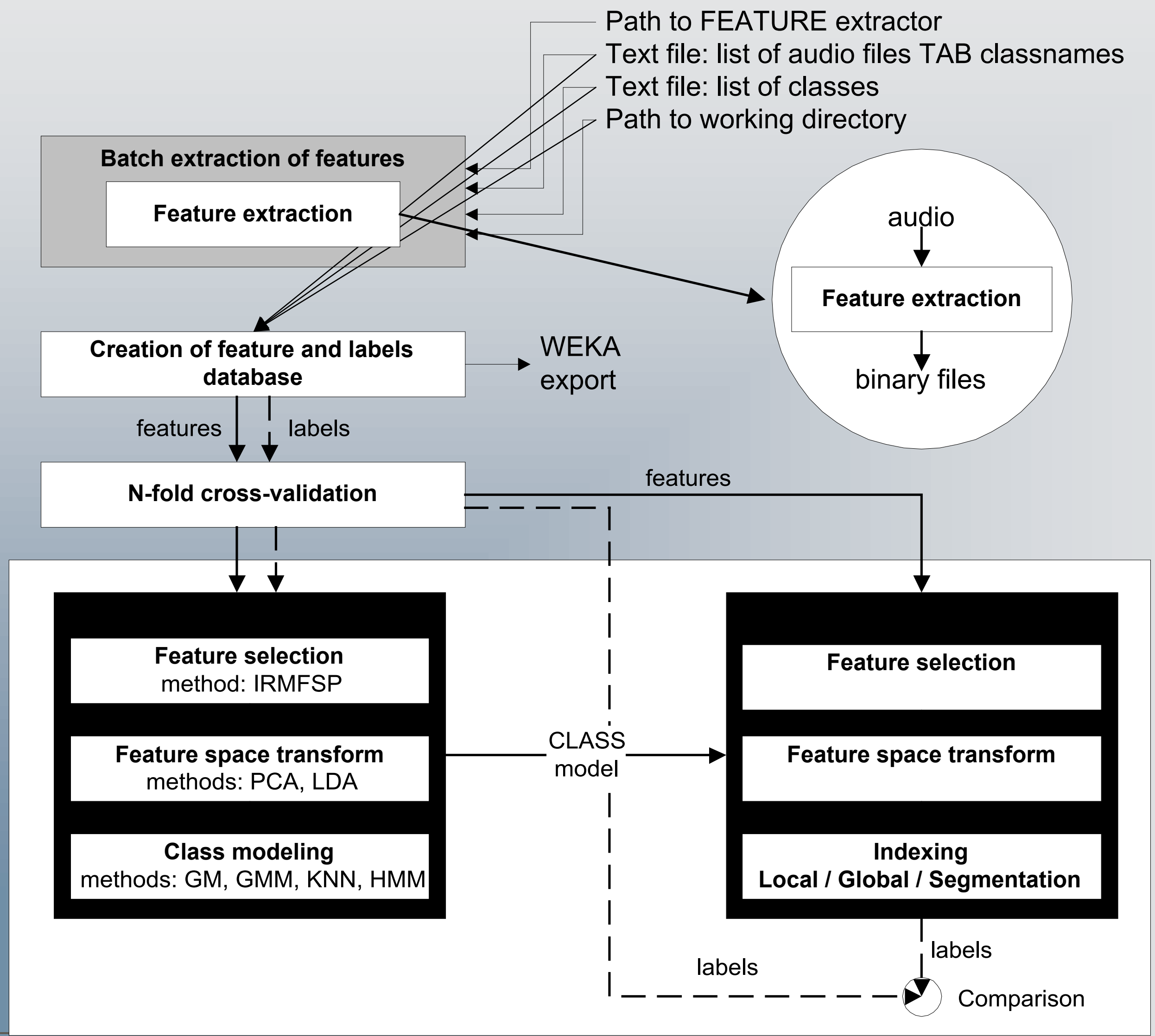
- <u>Training</u>:
  - extract features from a set of audio files
  - find a mapping between
    - the characteristics of the features and
    - hand-annotated labels of the audio files
  - labels ? they define the problem to be solved.
    An audio file can have a unique or a succession over time of labels
  - <u>-> the set of files and the corresp. labels must be easy to define and modify by the user</u>

- Performances of the system depends strongly on
  - the choice of the features
    <u>-> must be easy to be changed + include an automatic feature selection alg.</u>
  - choice of the model to represent the mapping between features and classes
    (SVM, KNN, GMM, ANN)
    <u>-> must be easy to be changed</u>

- Testing the performances of the system:
  - Cross-database, N-fold cross validation, Leave-One-Out

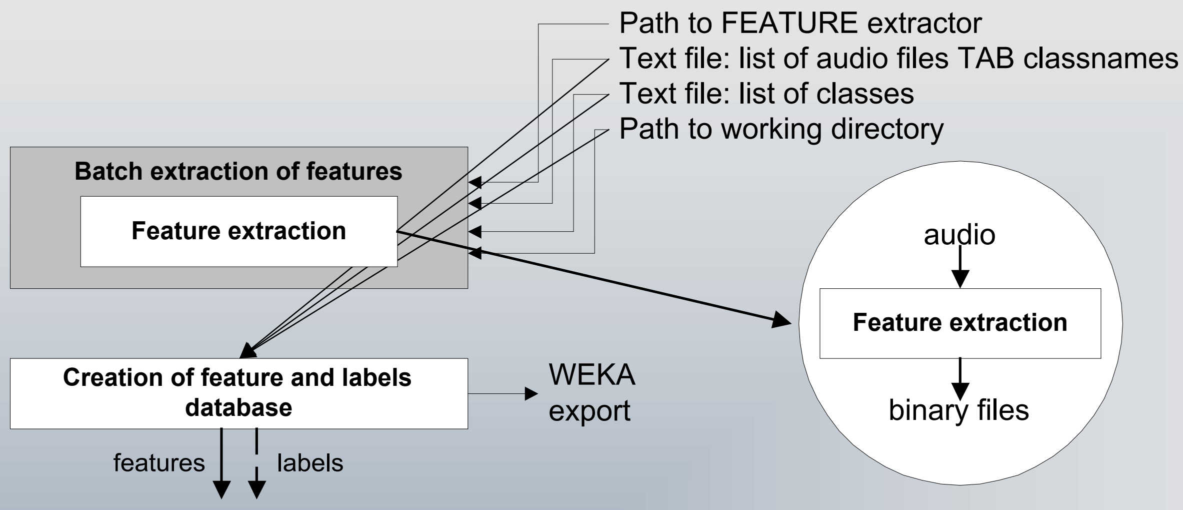
Generic audio indexing system

Description of the system



Generic audio indexing system

Description of the system



Generic audio indexing system
Describing a new indexing problem

- Defining a new problem using text files
  - 1) Text file defining the list of audio files and the corresponding annotations
    - path_audio_file \t class

```
J:\sound\wav_training\16-air_suite_from_les_fetes_d_he.wav   class:classical
J:\sound\wav_training\1-sartinal_i.wav   class:electronic
J:\sound\wav_training\6-three.wav   class:metal_punk
```

    - path_audio_file \t path_to_wavesurfer_file

```
L:\speechmusic\00_0003BB8A.wav      class:file:L:\speechmusic\00_0003BB8A.lab
L:\speechmusic\01_0003BB8D.wav      class:file:L:\speechmusic\0003BB8D.lab
```

- wavesurfer_file : time_begin \t time_end \t class\n
  time_begin \t time_end \t class\n ...

```
0.0000000 14.3621269 jingle-jingle001
14.3621269 220.5964666 music-music
220.5964666 237.0186157 talk-talk
237.0186157 244.4753160 jingle-jingle008
244.4753160 293.8578211 talk-talk
293.8578211 490.3433231 music-music
490.3433231 496.4073322 jingle-jingle028
496.4073322 671.0000000 music-music
671.0000000 677.0000000 jingle-jingle022
677.0000000 874.6100077 music-music
874.6100077 880.4419017 jingle-jingle004
880.4419017 897.5603964 talk-talk
```

Generic audio indexing system
Describing a new indexing problem

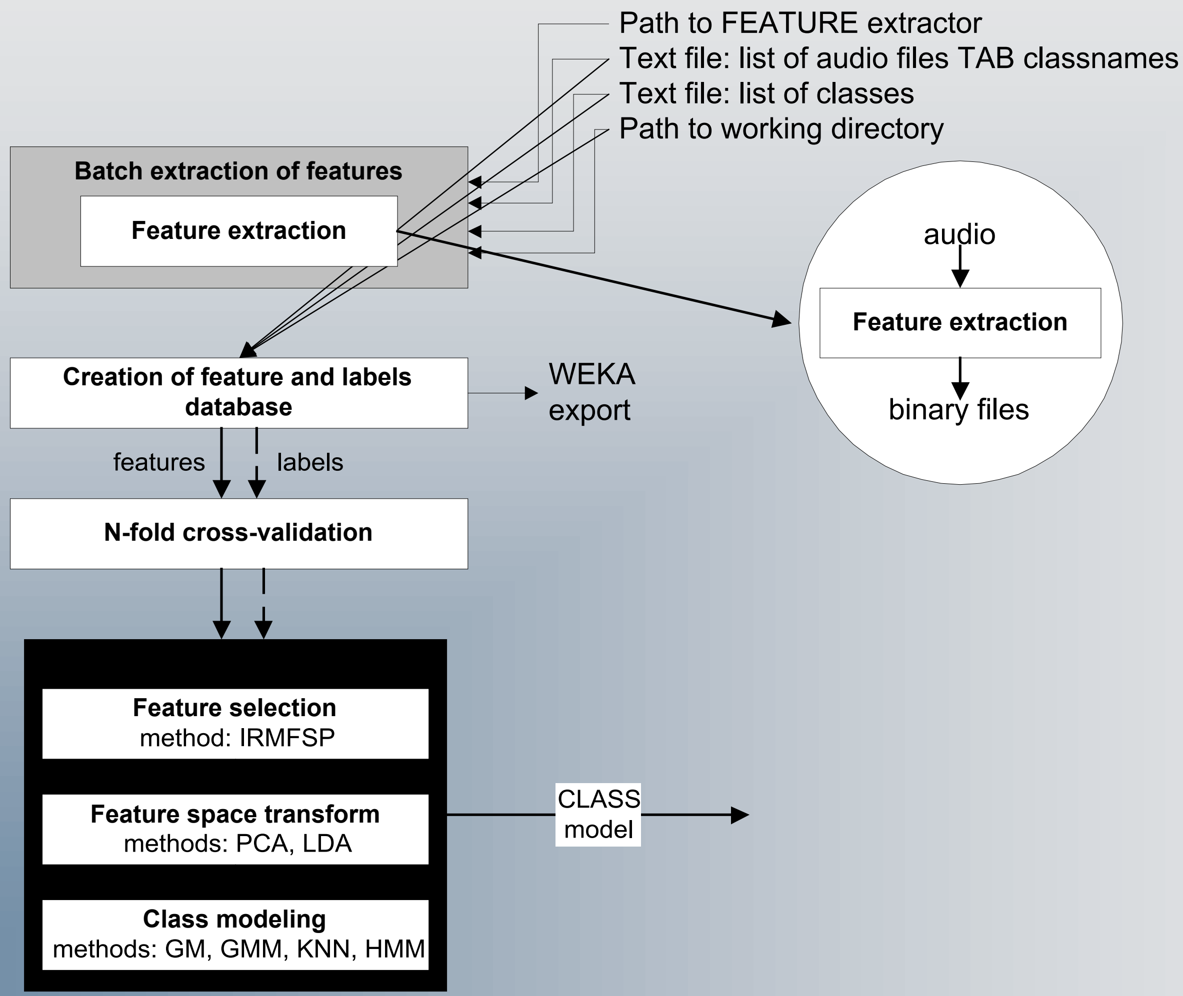
- 2) Text file defining the list of classes
  - the list can be a subset of the annotated classes
  - the list can perform a mapping between annotated class and new class names
    - allows mapping between different labels form different data-sets
    - allows to create a hierarchy among classes (“talk-talk” + “ads-talk” = speech)

```
music-music music
talk-talk   speech
ads-talk    speech
```

- 3) Path to the feature extractor to be used
  - feature extractor = independent executable
  - input = audio filename
  - output = feature filename
    - file format of the feature filename
      -> must guarantee compatibility between features and class definitions
      - features values
      - feature names
      - identifier to the used feature extractor
      - version of the feature extractor
      - parameters of the feature extractor

Generic audio indexing system

Description of the system



Generic audio indexing system

Description of the system

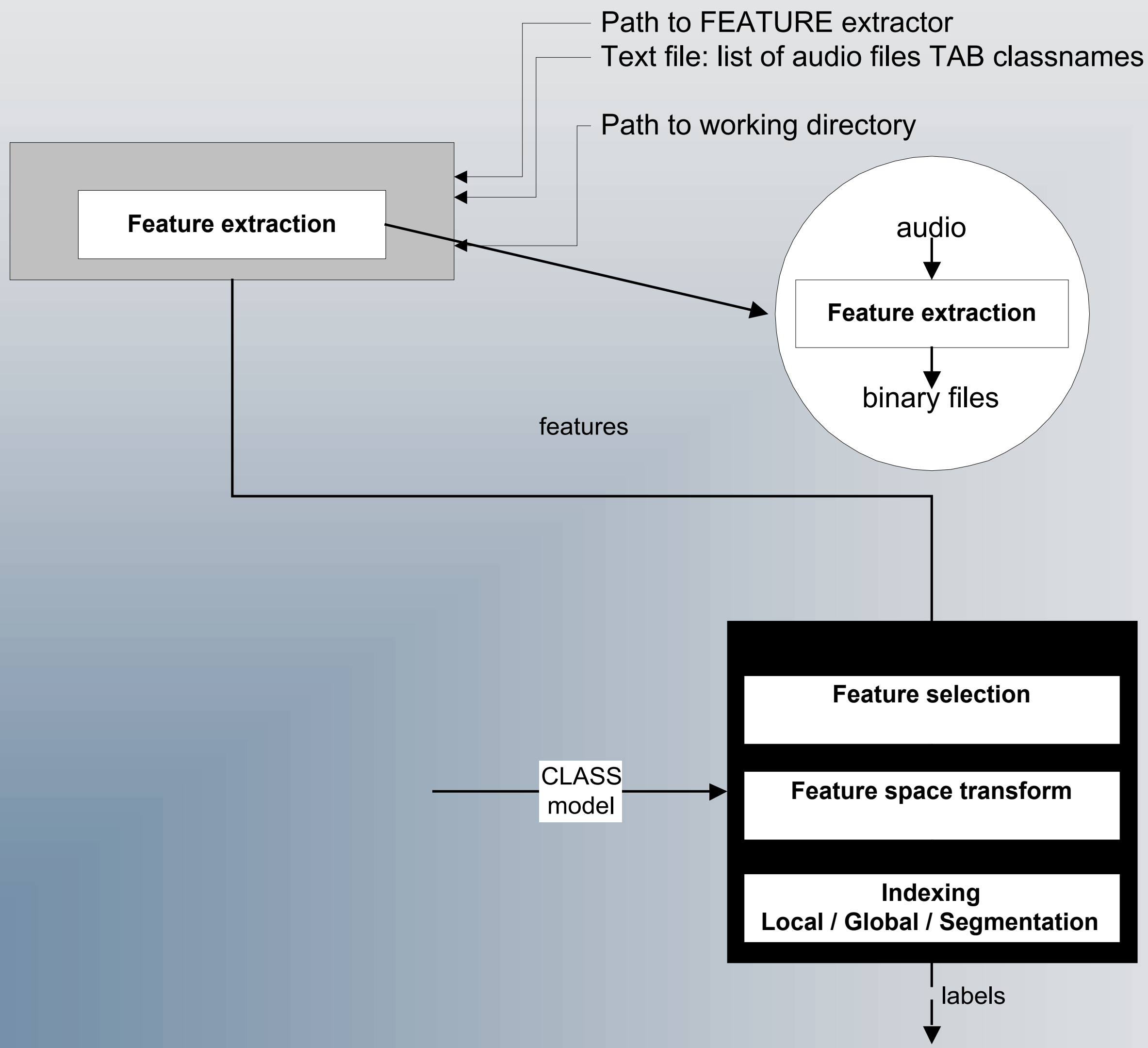
- Training
  - Batch feature extraction

  - Creation of a database containing feature values + corresponding classes
    - Weka export
    - Systat export

  - Training
    - Feature selection
      - IRMFSP algorithm (Peeters 2003)
    - Feature space transform
      - PCA: reduces the dim. of feature space while preserving most of the variances of the data
      - LDA: reduces the dim. of feature space while maximizing the class separation of the data
    - Class modeling
      - Multi-dimensional Gaussian modeling,
      - GMM,
      - HMM
      - K-Nearest Neighbors
      - Clustering algorithm, Histogram learning
  - Output:
    - a CLASS model that can be used for indexing unknown data

Generic audio indexing system

Description of the system

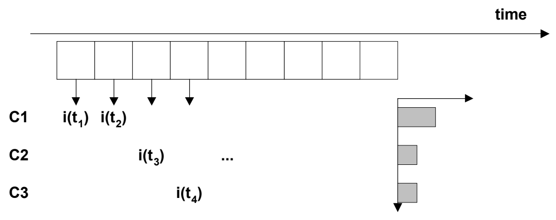


Generic audio indexing system
Description of the system

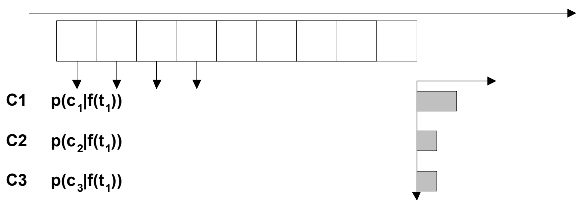
- Indexing
  - Local indexing
    - assign various labels over the file duration
      - smoothing of the labels -> median filtering, local histogram
      - can be used for segmentation -> use class changes over time
  - Global indexing
    - assign a single label to the whole file (or segment) duration
      - case 1) due to the fact that the extracted features are timeless
      - case 2) need to make a global decision from a succession of instantaneous (local) decision

Generic audio indexing system
Description of the system

- Global indexing methods: case 2)
  - Motivation
    - 1) we know that all the frames of a given segment/ file should belong to the same class
    - 2) the definition of the class does not come from the frame-class but from the distribution (or succession) of frame-class over time
  - Methods:
    - Cumulated histogram
      - $i(t) = \mathrm{argmax}(\, p(c_i|f(t))\,)$
      - maximum of cumulated histogram $i(t)$



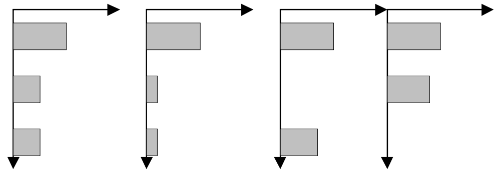
- Cumulated probability
  - $p(c_i) = 1/T \, \mathrm{sum}_t \, p(c_i|f(t))$
  - maximum of the cumulated probability $p(c_i)$



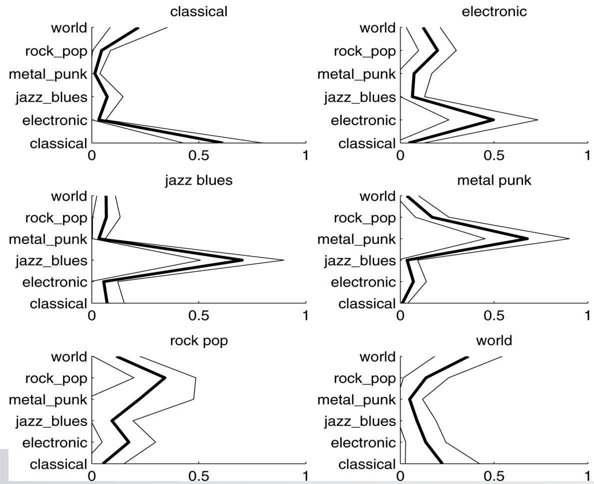
Generic audio indexing system

Description of the system

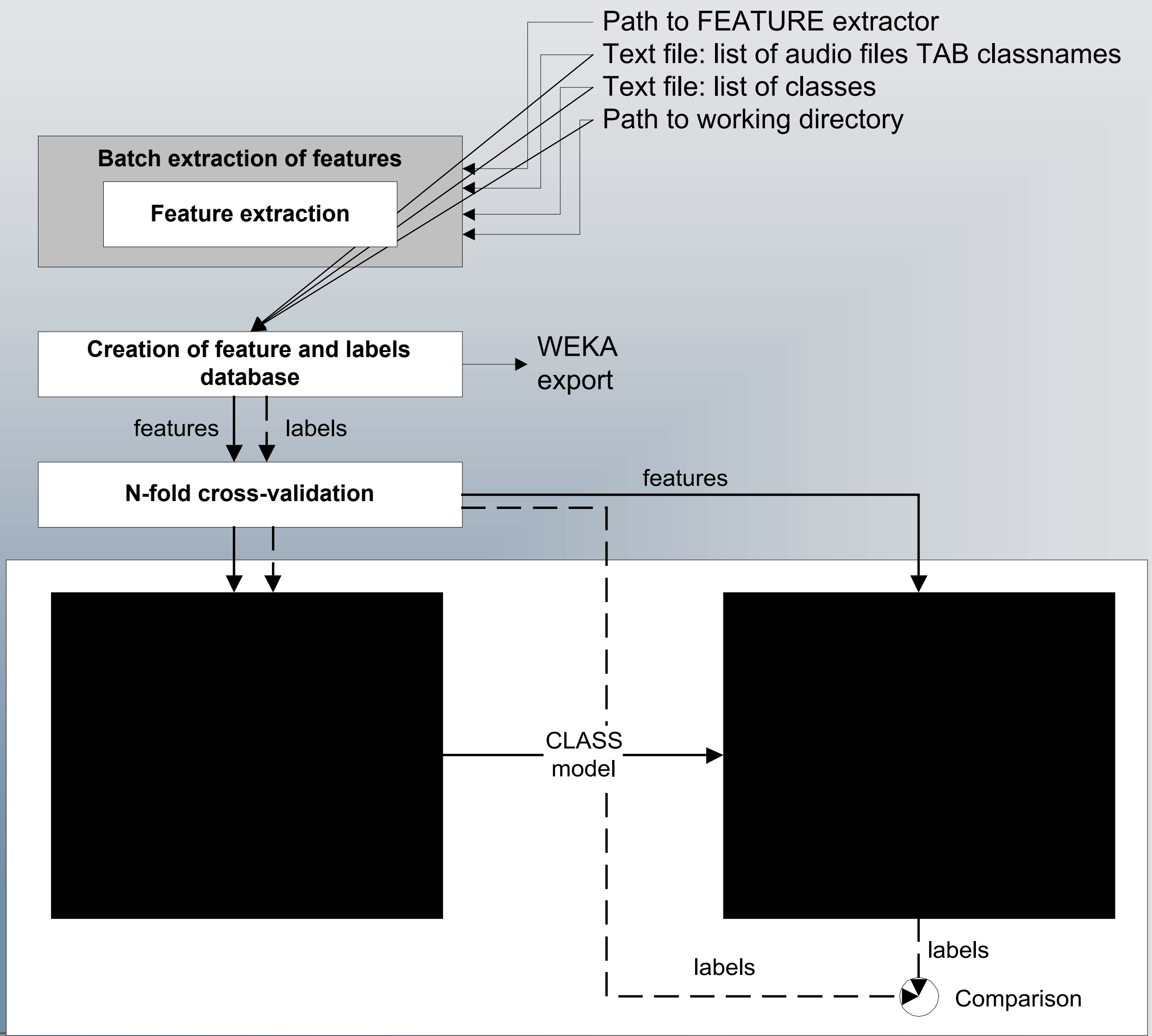
- Global indexing methods: case 2)
  - Methods:
    - Segment-statistical model
      - Notation: s_i a specific segment of the training set belonging to class i
      - for a specific class i
        - model the behavior of the bins c_i of $p_{si}(c_i)$ over all the segment si belonging to class i



      - Model ?
        - Gaussian modeling of the bins of $p(c_i)$ -> <u>segment-statistical model</u>
        - Cosine distance between the cumulated probability of the unknown segment $p_{\underline{s}}(ci)$ and the mean vector of each segment-statistical model
      - Example: case of music genre recognition
        - the bins i are the mg classes
        - each segment-statistical model represent the behaviour of the bin for a specific mg class

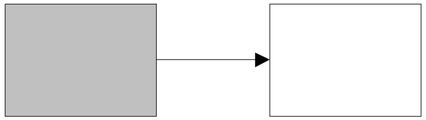


Generic audio indexing system
Description of the system

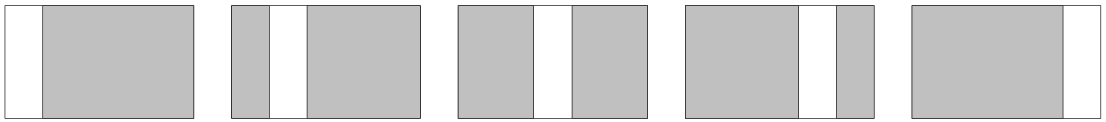


Generic audio indexing system
Description of the system

- Validation
  - Cross-database validation:
    - one database is used for training, the other one for evaluating the performances of the system

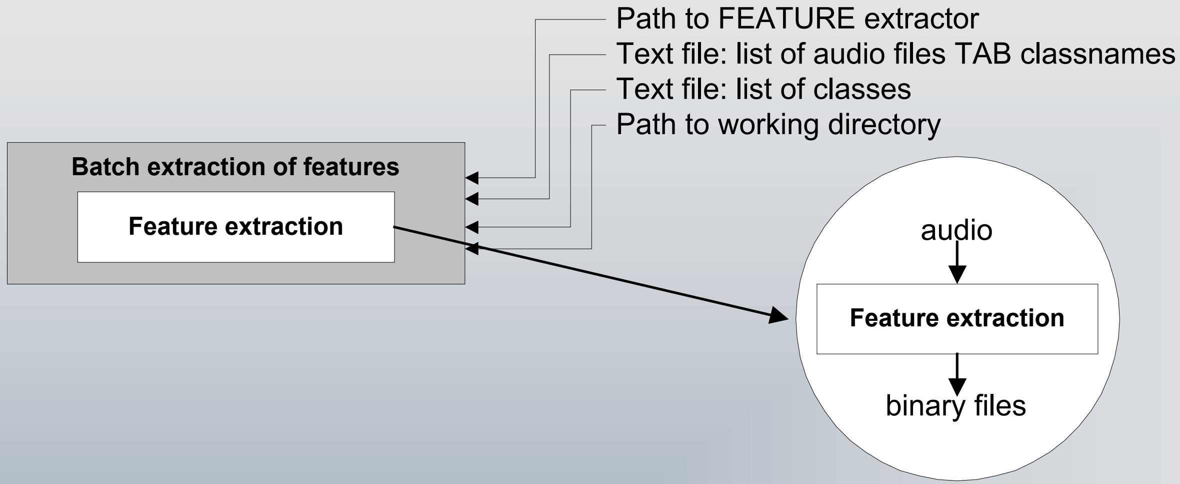


  - N-fold cross validation (Leave-One-Out)
    - the database is divided into N folds (as much independent as possible)
    - N-1 folds are used (in turns) for training, the remaining one for evaluation
      - ➔ Specific case: if N = the number of observation (or segments) = Leave-One-Out



Generic audio indexing system

Description of the system



Generic audio indexing system

Description of the system

- Features
  - <u>Dedicated</u> audio features set
    - the extraction of high-level concepts is
      - feasible: can be extracted considering current DSP limits
      - meaningful: has a meaning for the given signal
    - example: audio content= single mono. note (instrument sound sample recognition)
      - Attack-time, Fundamental frequency (assumption: time-extent of the signal, signal model)
      - Peeters 2004
  - <u>Generic</u> audio features
    - extraction of low-level concept because high-level concept extraction
      - can be difficult
      - can be meaningless
    - example: audio content= generic audio (any kind of audio: music, radio, talks, ...)
      - MFCC, SFM (no assumption on time-extent or signal model)
  - Too many data ! -> <u>Temporal modeling</u>:
    - 24h of radio program * 20ms hop size = 4 millions feature vector !
    - Temporal modeling = model the evolution of each feature over time (window length from 500ms to 2s)
    - Various models can be used over the window:
      - statistical moments (mean, variance),
      - histogram of cluster belonging,
      - spectral decomposition of feature evolution,
      - sub-band grouping

Applications

- Application of the generic system to two usual problems
  - Segmentation: speech/music segmentation of radio program
  - Recognition: music genre recognition (well-known MIR problem)

Speech/Music segmentation

- Goal:
  - develop a tool for the automatic segmentation of radio streams
  - developed in a real industrial framework
    - in coordination with a company that produces managing and archiving software for radio station
    - categories and corpuses directly defined and provided by their clients (real world data)

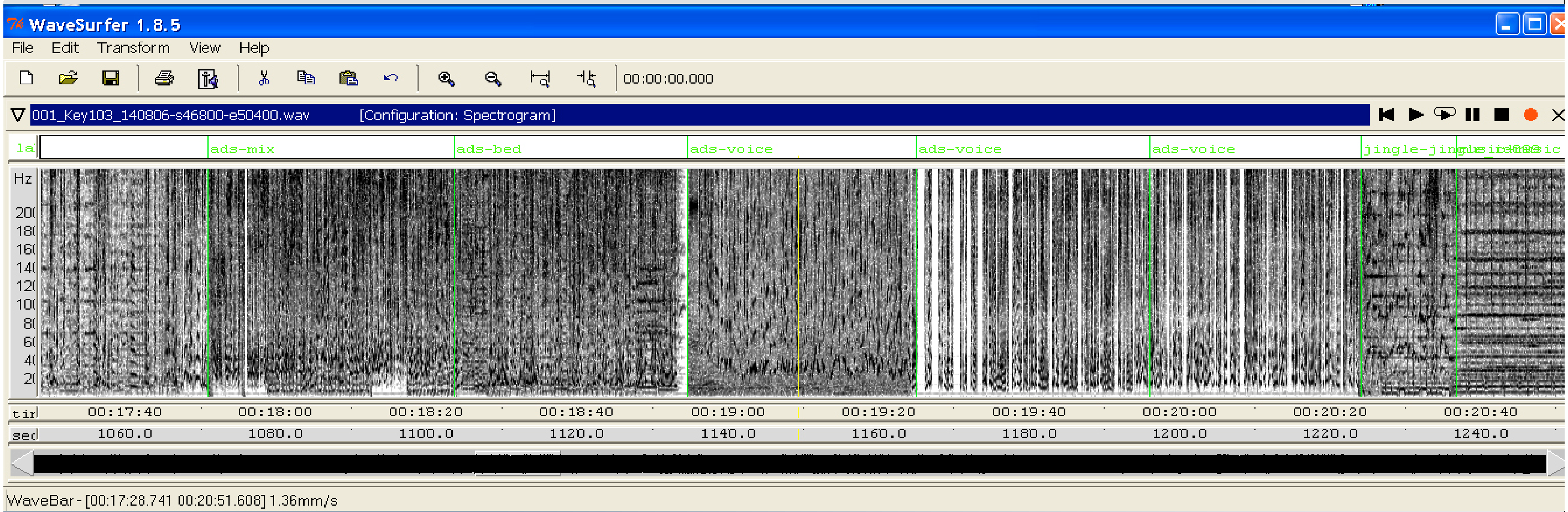
- Related works in speech/music segmentation
  - Large amount of research in the last two decades
  - Usual methods:
    - low-level features (ZCR, 4Hz energy modulations, MFCC, entropy)
    - KNN, GMM, SVM
    - References: Scheirer97, Saunders96, Carey99, Harb03, Pinquier06, Richard06
  - Evaluation protocols:
    - DARPA (USA), ESTER (France)

Speech/Music segmentation

- Considered categories
  - acoustical categories
    - music
    - speech
    - mix: speech and music exist but do not overlap continuously over time (succession)
    - bed: speech and music overlap regularly over tome (introduction of radio news)
  - industry categories
    - music
    - talks
    - ads
    - jingles

Industry / Acoustical	Music	Jingle	Voice	Mix	Bed
Music	music-music				
Jingle		jingle-jingle			
Talk			talk-voice	talk-mix	talk-bed
Ads			ads-voice	ads-mix	ads-bed

Speech/Music segmentation



Speech/Music segmentation

- Corpus
  - Corpus Radio France
    - speech: a subset of MPEG-7 corpus (RadioFrance July 1998)
    - music: ISMIR2004 “song excerpts” dataset + a private music genre database
  - Corpus UK
    - 24h of recording of a major commercial radio group in the UK
    - Description: high rate of audio compression, many ads, jingles, talks and music
  - Corpus SUD
    - 24h of recording of a regional radio station in France

Corpus name		**RadioFrance**		**UK**	**SUD**
Description		french speaking		english speakin	french speaking
Total duration		622m		1375m	1333m
Classes:	**music-music**	74% (460m)	**music-music**	57% (788m)	70% (945m)
	speech-clean	26% (162m)	**talk-voice**	16% (222m)	23% (312m)
			talk-mix	8% (111m)	1% (13m)
			talk-bed	3% (41m)	1% (10m)
			ads-voice	4% (51%)	1% (10m)
			ads-mix	6 (89m)	3% (35m)
			ads-bed	5% (70m)	1% (8m)

Speech/Music segmentation

- System configuration
  - Signal: 11KHz, mono, 40ms Blackman, 20ms hop size

  - Features
    - 13 MFCC + Delta + Delta-Delta
    - 4 SFM + Delta + Delta-Delta
    - Temporal modeling: mean+variance 2s / 1s

  - Classifier (best configuration found ...)
    - Feature selection: IRMFSP first 40 selected features
    - Feature space transform: Linear Discriminant Analysis
    - Class modeling: GMM with 20 mixtures and full-covariance matrix, training set highly unbalanced -> no use of prior information

Speech/Music segmentation

- Results
  - 7 classes problem (random=14.28%)
    - results for the UK corpus (the most difficult), ten-fold CV

		Found							
		'music-music'	'talk-voice'	'talk-mix'	'talk-bed'	'ads-voice'	'ads-mix'	'ads-bed'	
Real	'music-music'	**79,4**	0,5	1,7	2,9	0,9	8,5	6,1	
	'talk-voice'	0,5	**71,8**	8,1	5,0	12,4	1,3	0,8	
	'talk-mix'	2,6	8,3	42,6	22,2	6,3	9,1	8,9	
	'talk-bed'	4,1	3,9	34,9	39,8	5,3	6,4	5,6	
	'ads-voice'	1,1	10,0	5,8	3,1	**66,2**	9,2	4,4	
	'ads-mix'	12,1	2,3	9,4	5,8	10,4	41,7	18,3	
	'ads-bed'	6,8	0,8	6,0	5,3	6,1	14,4	60,6	
									57,5

    - largest confusion between
      - the non-pure catgories (mix and bed),
      - talk and ads
  - 2 classes problem (random=50%)
    - we only consider the pure categories (no mix, no bed),
      we merge talk-voice and ads-voice into "speech", results for ten-fold CV
      - UK: Rmusic=96.7%, Rspeech=94.4%
      - RadioFrance: Rmusic=96.48%, Rspeech=96%
      - SUD Rmusic=95.8% Rspeech=92.1%
    - Conclusion: music tends to be more easily recognized than speech

Results

- Cross-database validation: one corpus for training, two remaining for evaluation)

	Mean Recall (Mean F-Measure) Music Recall - Speech Recall	Evaluation: Radio-France	Evaluation: UK	Evaluation: SUD
Training	Radio-France		**86,5 (87,9)** 99 - 73,9	**95,2 (96,4)** 90,9 - 99,5
Training	UK	**92,1 (57,6)** 84,3 - 99,9		**89,4 (92,7)** 79,1 - 99,8
Training	SUD	**95 (78,1)** 96,9 - 93,2	**90,2 (91,3)** 99,1 - 81,3	

- Conclusion:
  - Best results: RF->SUD, SUD->RF
  - Worst results: RF->UK, UK->SUD
  - RF / SUD very close , UK different
  - Assumption: does the difference come from the language ? No
    - → UK speech -> RF speech / SUD speech = good
  - Assumption: does the difference come from the music ? Yes
    - → UK music -> RF music / SUD music = bad
  - Best corpus for training the music ?
    - → SUD music

Speech/Music segmentation

- Conclusion
  - Good results considering that we did not perform any modification of our system for the problem of speech/music segmentation
  - Problem for the "mix" and "bed" categories
    - need new features for these classes :
  - Good results for the 2 "pure" classes speech and music
  - Choice of the training set is important for the generalization of the system
    - a different training set may be used for the classes speech and music

- Comment on the F-measure and Precision factor
  - both depend strongly on the distribution of the test set (which is highly unbalanced in our case)
    - Recall: s->s / s
    - Precision: s->s / (s->s + m->s)
    - F-measure: 2RP/(R+P)
  - example:
    - UK-> RF : Recall speech =99.9% but Precision = 13.4%
    - It looks like a large part of music has been classified as speech ?
    - this part is small in comparison to the number of music data (only 15.6% of the music data)
    - but the number of music data (m=48382) is large compared to the number of speech data (s=1175)
    - Therefore the Precision drops
      - → Precision = 0.999s / (0.999s + 0.156m)

Applications

- Application of the generic system to two usual problems
  - Segmentation: speech/music segmentation of radio program
  - Recognition: music genre recognition (well-known MIR problem)

Music Genre recognition

- Goal:
  - develop a tool for automatic recognition of the music genre of an audio track
- Music genre categories ?
  - Fuzzy and hill-defined concept
  - still it is important for understanding the underlying features of music similarity
- Related works in music genre recognition
  - Usual methods:
    - 1) low-level features: MFCC, spectral contrast, loudness, roughness
    - 2) high-level features: tempo, beat histogram, chroma, pitch contours
    - References: Aucouturier03, Jiang02, Burred03, Tzanatakis02, McKay04
  - Evaluation protocols:
    - ISMIR2004, MIREX05/06/07

Music Genre recognition

- Corpus and categories
  - ISMIR2004 music genre contest
    - training, development parts (not the evaluation parts)
  - 6 categories
    - Highly unbalanced in favor of Classical music

Music Genre	Classical	Jazz / Blues	World	Electronic	Metal / Punk	Rock / Pop	Total
Training set	320	26	106	115	45	101	713
Development set	320	26	122	114	45	102	729

Music Genre recognition

- System configuration
  - Signal: 11KHz, mono, 40ms Blackman, 20ms hop size
  - Features
    - 13 MFCC + Delta + Delta-Delta
    - SFM + Delta + Delta-Delta
    - Temporal modeling: mean+variance 4s / 2s
  - Classifier (best configuration found ...)
    - Feature selection: no
    - Feature space transform: Linear Discriminant Analysis
    - Class modeling: GMM with 5 mixtures and full-covariance matrix, training set highly unbalanced -> no use of prior information

Music Genre recognition

- **Results**
  - frame-based decision method (87039 frames): 62.2% (14.3%)
- Comparison between the global decision methods
  - cumulated histogram: 76.2% (18.9%)
  - cumulated probability: 77.4% (16.8%)
  - segment-statistical model (cosine distance method): 78.7% (14%)

- Largest confusions
  - classical <-> world
  - metal-punk -> pop-rock

		Found						
		classical	electronic	jazz_blues	metal_punk	rock_pop	world	
Real	classical	90,6	0,0	0,3	0,0	0,0	**9,1**	
	electronic	1,8	73,7	0,9	2,6	**9,6**	11,4	
	jazz_blues	0,0	0,0	96,2	0,0	3,8	0,0	
	metal_punk	0,0	0,0	0,0	84,4	**15,6**	0,0	
	rock_pop	0,0	4,9	2,9	**16,7**	67,6	7,8	
	world	**16,4**	4,9	4,9	0,8	13,1	59,8	
								78,7

- Comparison to state of the art
  - results during ISMIR2004
    78.78%, 67.22%, 58.60%, 55.70%, 51.48%

- Conclusion:
  - Good results considering that no specific ...
  - But purely timbre-related features (such as MFCC or SFM) is not enough to distinguish high-level concept such as music genre

Conclusion and future works

- Generic system
  - easy to use
  - applicable to a wide range of indexing problems
- Application to speech/music segmentation (in a real industrial framework)
  - good performances when considering only the pure categories speech/ music
  - system can be generalizable across radio channels (cross-database valid.)
  - performances drop when considering the non-pure categories (mix and bed)
- Application to music genre recognition problem
  - we have proposed the use of segment-statistical model
    -> allows improving the results
  - results close to previous state-of-the-art algorithm (ismir2004)
- Future works
  - new features required for
    - bed and mix categories
      - observe separately the various parts of the spectrum
    - music genre
      - higher-level features such as rhythm patterns, chord succession
    - extend the current set of features on which the automatic selection is performed
  - real-world data-sets often unbalanced
    - take into account this unbalancing in our algorithm
      - modify feature selection and feature space transform algorithms