A STRUCTURAL SEGMENTATION OF SONGS, USING MULTI-CRITERIA GENERALIZED LIKELIHOOD RATIO AND REGULARITY CONSTRAINTS

Gabriel SARGENT, Frédéric BIMBOT, Emmanuel VINCENT
IRISA/INRIA Bretagne-Atlantique

Summary

Summary

Context

Music Information Retrieval :

- Collections of digital music grow over time
(web services : streaming, mp3...)

- Easy and quick access to their elements needed

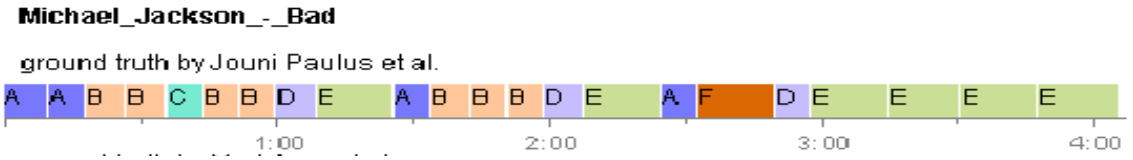
-> indexing songs from their musical content,
which permits song search from its melodies, genre...

Music structure...

We focus here on **western popular music**. Its musical content can be described according to different axes, like key, tempo, timbre, rhythm, harmony, melody and lyrics.

Musical structure is a high-level description of a song which can be characterized by **temporal segments** labelled according to the **similarity** of their musical content.

The span of these segments covers a group of musical bars.



(Peizer *et al*, Automatic Audio Summarization)

! Relying only on the audio, a variety of structural descriptions can be extracted from the same song...

Summary

Popular features

A song is described by sequences of features regularly extracted from its audio signal.

- **MFCC** (*Mel-Frequency Cepstral Coefficients*) : set of values (usually 13 or 20 coefficients) which describe the rough shape of the power spectrum's envelope of the audio signal. It gives information on the "timbre of the song"

- **Chroma vector** : set of 12 values which quantifies the energy of the frequencies associated to the 12 semi-tones of the chromatic scale (with no octave distinctions). It gives information on the harmony of the song.

Main methods for structure inference

2 "main categories" :

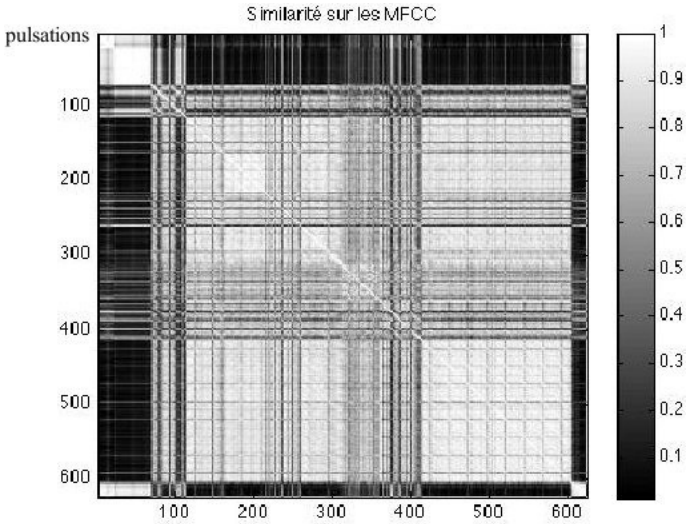
- finding homogeneous parts

- finding repeated sequences

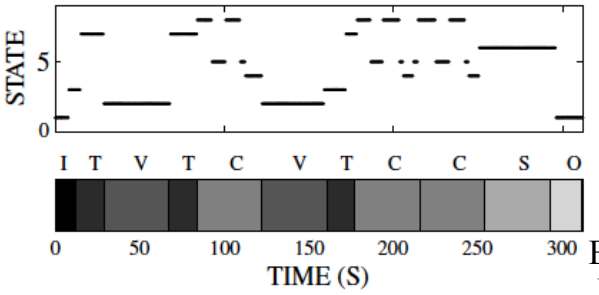
Finding homogeneous parts :

- MFCC sequence is usually considered.

- Similarity matrices (Foote):
→ zones with specific Textures ("timbres")
→ segmentation = border localization
+ labelling by clustering



- Hidden Markov Models (Logan)
→ musical parts = hidden states ;
find the best state sequence



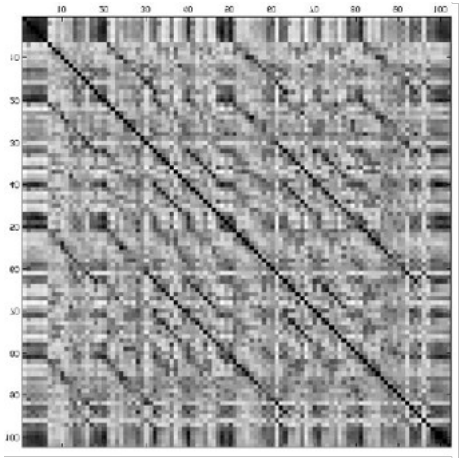
Extracted from Paulus, 2010

- Cost function optimization (Jensen):
→ find the structure with minimal number of homogeneous structural segments.

Finding repeated sequences :

- Chroma sequence is usually considered.

- Similarity matrices (Goto):
→ dark strips on sub-diagonals



Extracted from Peeters, 2002

- Hidden Markov Models (Rhodes and Casey):
→ pattern matching on the hidden state sequence

- Dynamic time warping (Martin):
→ pattern matching with tolerance on sequences repeated in "time-streched" versions.

a b c b d	e b	a b	b	c b d	f g h b...

Summary

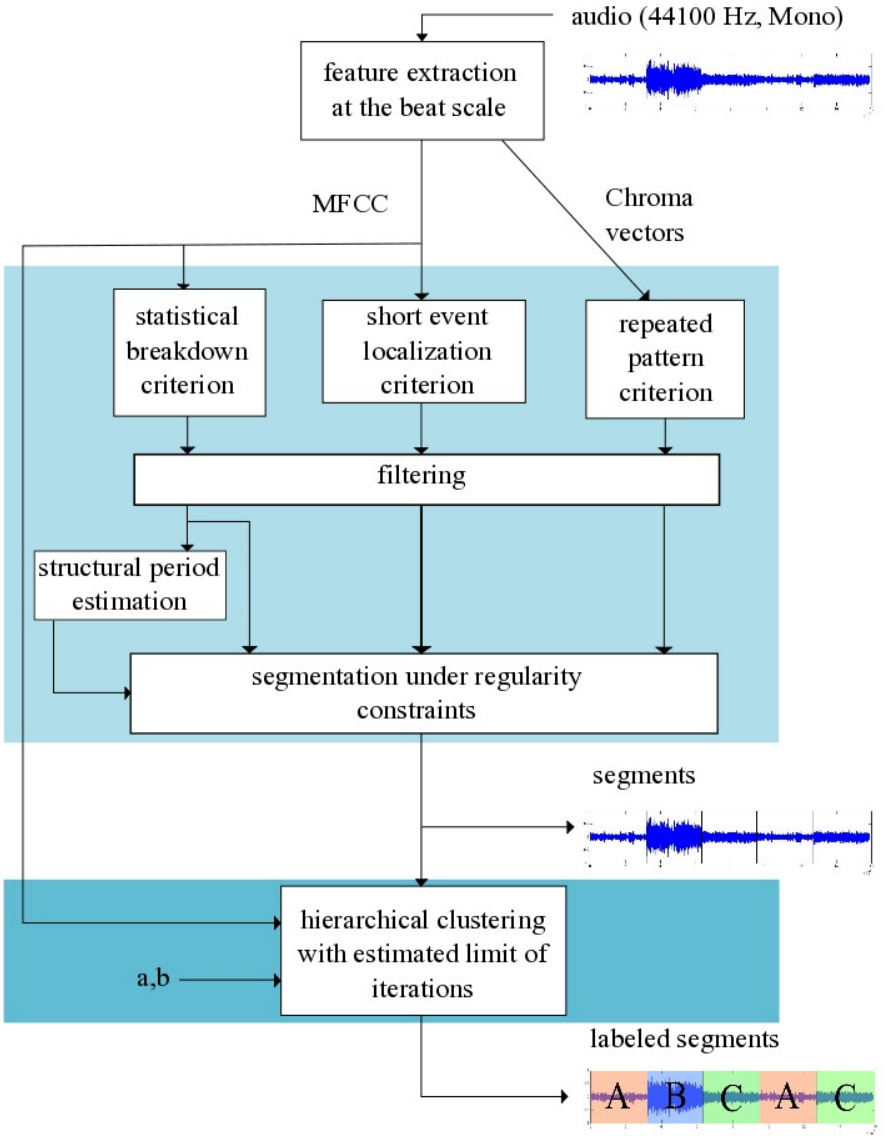
Algorithm :

Audio described at the beat scale by :
- a sequence of MFCCs
- a sequence of Chroma

Two main parts :

- segmentation (light blue) using MFCCs and Chroma

- labeling (dark blue) using the MFCCs



II.1. Segmentation

3 criteria are used to infer the location of the boundaries :
- Statistical breakdown criterion
- Short event localization criterion
- Repeated pattern criterion

They are modeled by the **Generalized Likelihood Ratio :**

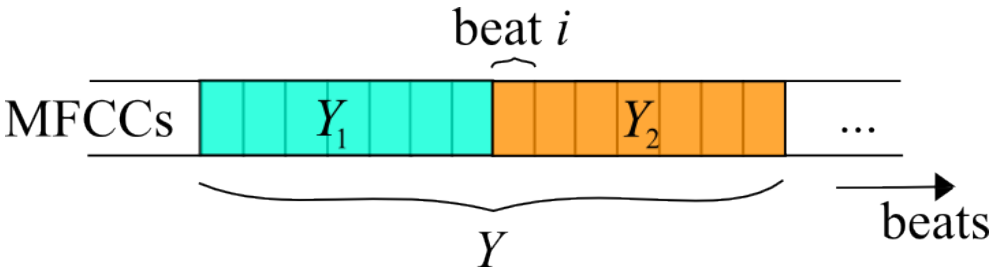
It estimates and compares the likelihood of a certain model H_0 of the data to that of an alternative model H_1.
Let Y be the sequence of features describing the song :

$$\mathrm{GLR} = \frac{P(Y|H_1)}{P(Y|H_0)}$$

II.1.a. Statistical breakdown criterion :

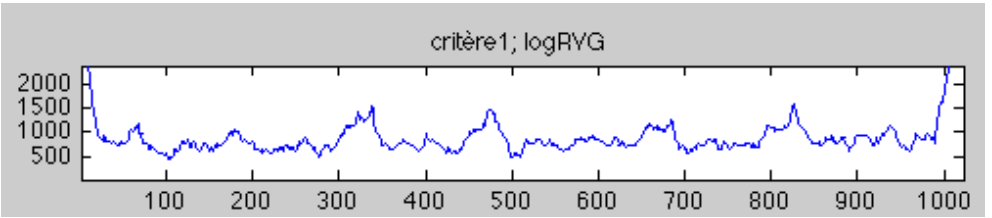
- MFCC sequence (beat scale) is used

- **Analysis window** of length 12s is centered on the current beat. It is divided in **two parts** : "close past" (6s) containing Y_1, and "close future" (6s) containing Y_2.



- H_0 : Y is well-modeled by a single Gaussian distribution.
- H_1 : Y_1 and Y_2 are well-modeled by two distinct Gaussian distributions.

→ log(GLR) :

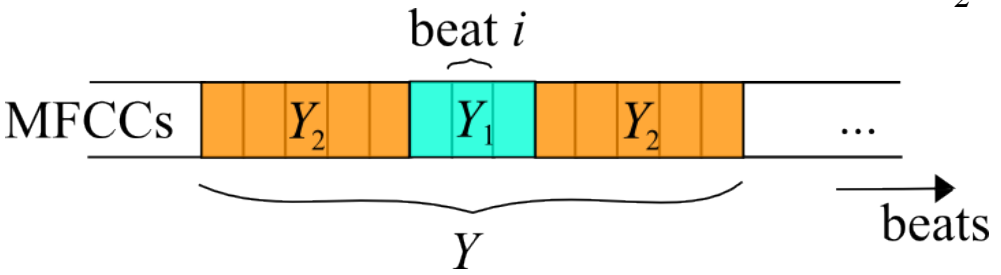


Pink Floyd,
Brain Damage

II.1.b. Short event localization criterion :

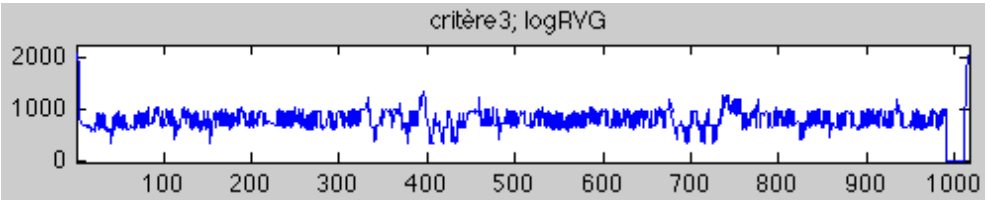
- MFCC sequence (beat scale) is used

- **Analysis window** of length 12s is centered on the current beat. **Two parts** : "close neighbouring" (2s) containing Y_1, and "environment" (10s in total) containing Y_2.



- $\boldsymbol{H_0}$: Y is well-modeled by a single Gaussian distribution.
- $\boldsymbol{H_1}$: Y_1 and Y_2 are well-modeled by two distinct Gaussian distributions.

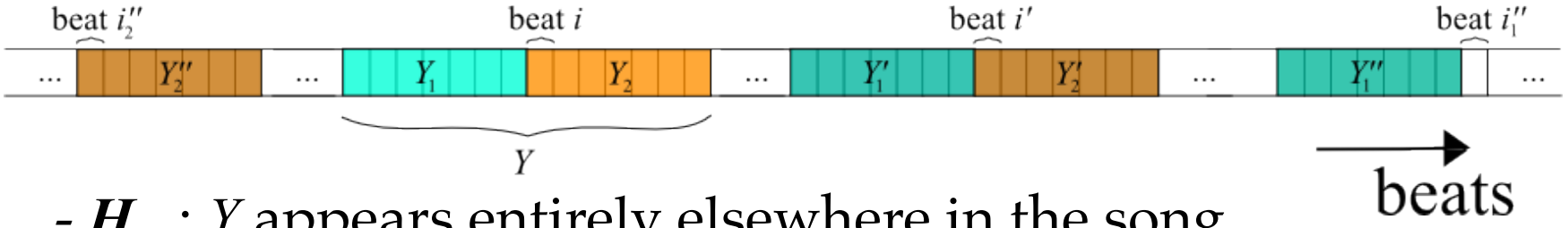
→ log(GLR) :



II.1.c. Repeated pattern criterion :

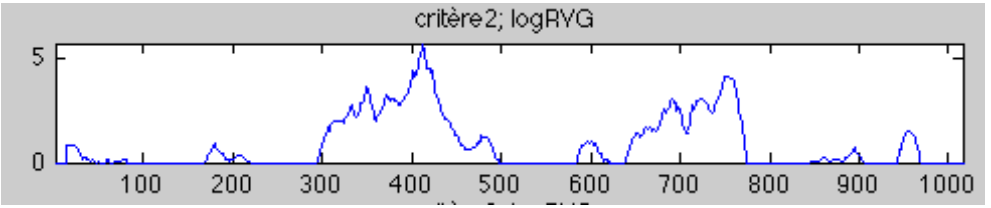
- Chroma sequence (beat scale) is used

- **Analysis window** of length 12s is centered on the current beat. **Two parts** : "close past" (6s) containing Y_1, and "close future" (6s) containing Y_2.



- $\boldsymbol{H_0}$: Y appears entirely elsewhere in the song.
- $\boldsymbol{H_1}$: Y_1 and Y_2 appear separately in the song.

→ log(GLR) :

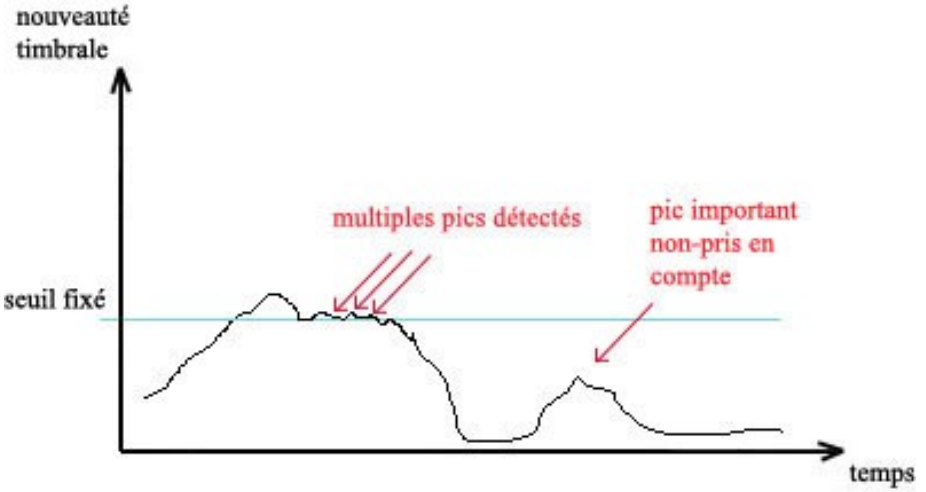


Here, exact matching is performed (euclidean distance is used).

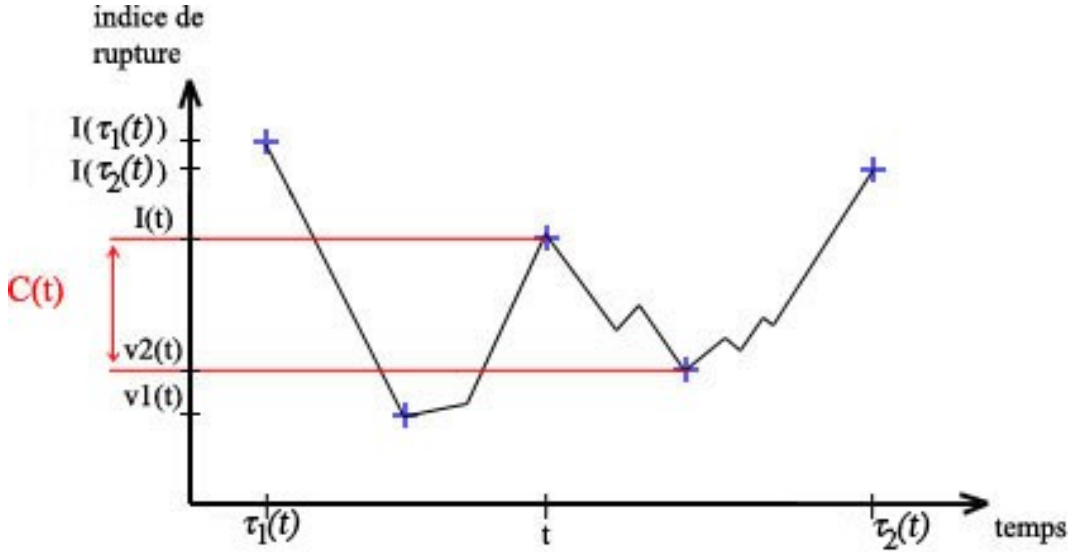
II.1.d. Filtering the criteria

Peak detection :

→ wrong detections, or no detection at all



Contribution : detection of dominant peaks (Seck)



Criterion : I.

$v_1(t)=\min_{tau1(t)<i<t} I(i)$

$v_2(t)=\min_{t<i<tau2(t)} I(i)$

$u(t)=\max(v_1(t),v_2(t))$

Filtered version C :

$C(t)=I(t)-u(t)$

II.1.e. Segmentation using combined criteria and regularity constraints :

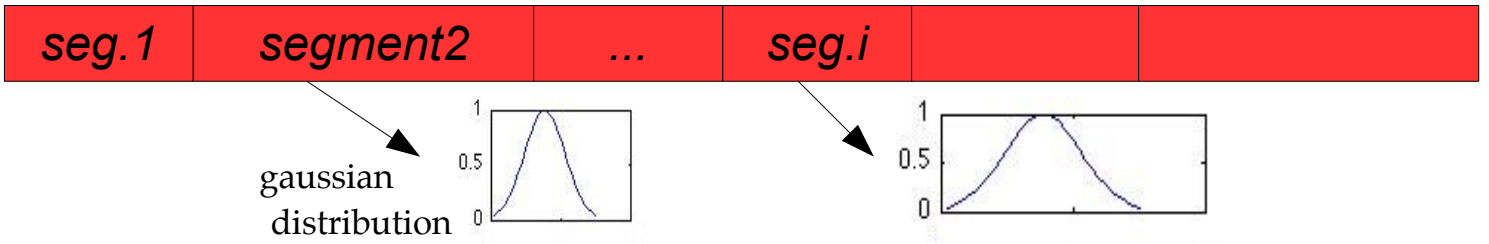
- estimation of a structural period (preferred length for structural segments) using the fast Fourier transform of the statistical breakdown criterion.

- selection of boundaries by minimizing the amplitude of the 3 criteria between segment borders combined with a penalty function which increases when the length of a segment departs from the estimated structural period.

II.2. Labeling

II.2.a Modeling the segments with their timbral content:

We assume that each structural segment can be modeled by a Gaussian distribution of its MFCCs (Cooper).

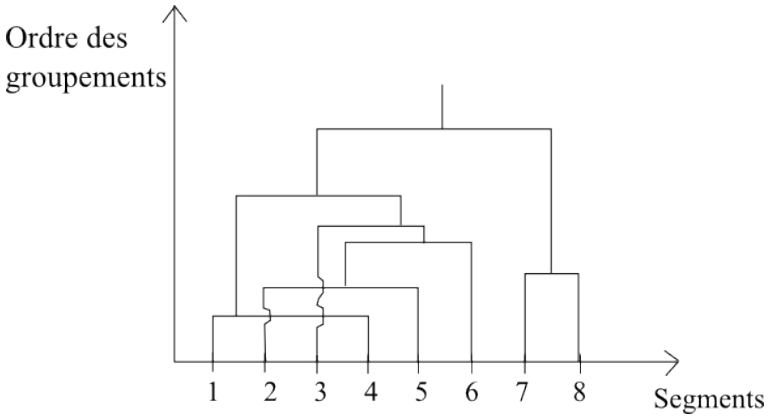


These Gaussian models are compared thanks to the *symmetrised Gaussian likelihood measure* used in speaker identification.
It quantifies if the features of a segment *i* are well modeled by the gaussian model of another segment *j*.

Contribution : hierarchical (agglomerative) clustering with adaptive number of iterations.

II.2.b. Hierarchical clustering

- 1st iteration :
Each segment belongs to a different cluster



- i^{th} iteration :
All the distances between the clusters are computed.
The two closest clusters ($\mu_{min}(i)$) are grouped in a new cluster, and its gaussian model is computed.

The process is iterated for a certain adaptive number of iterations.

II.2.c. Adaptive number of iterations

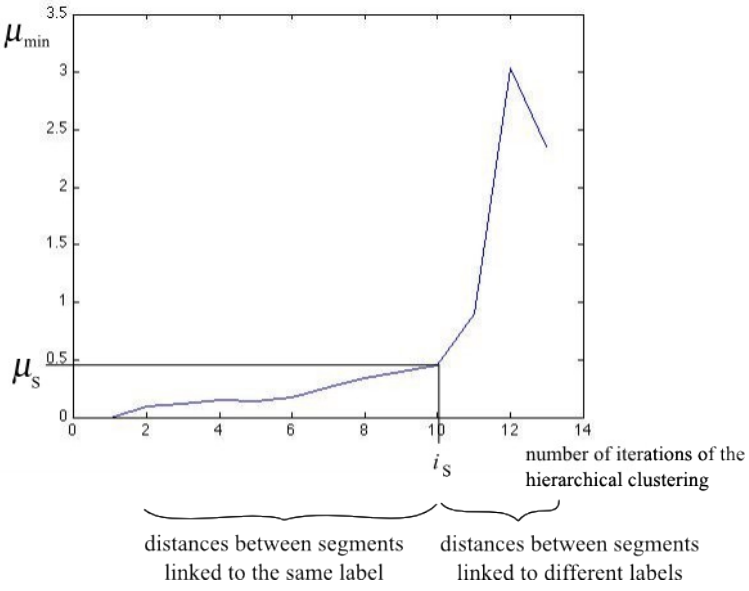
→ The evolution of $\mu_{\min}=\{\mu_{\min}(1),...,\mu_{\min}(N)\}$ is analysed

N = total number of segments

Assumption :

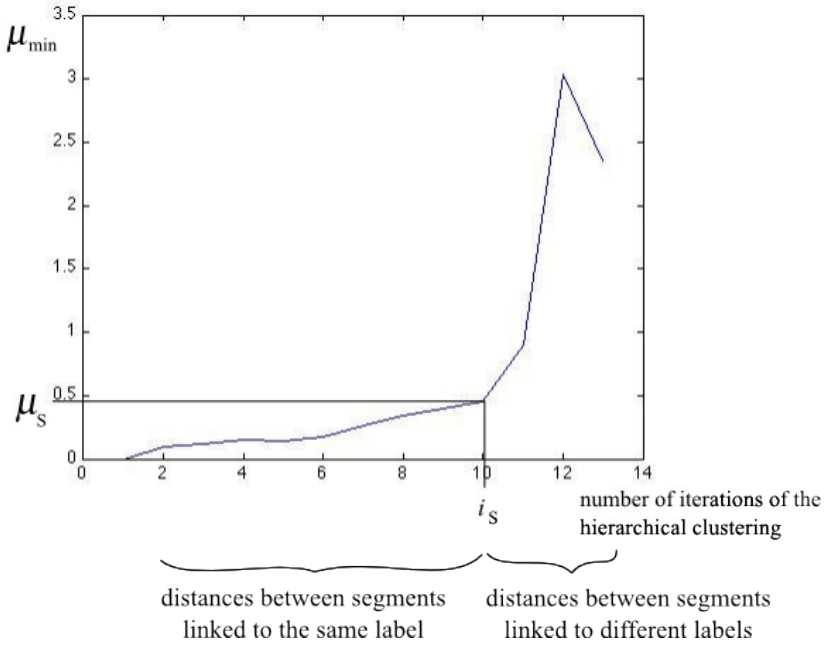
There are two types of distances :
- intra-cluster (distances between 2 segments with the same label)
- interclusters (distances between 2 segments with different labels).

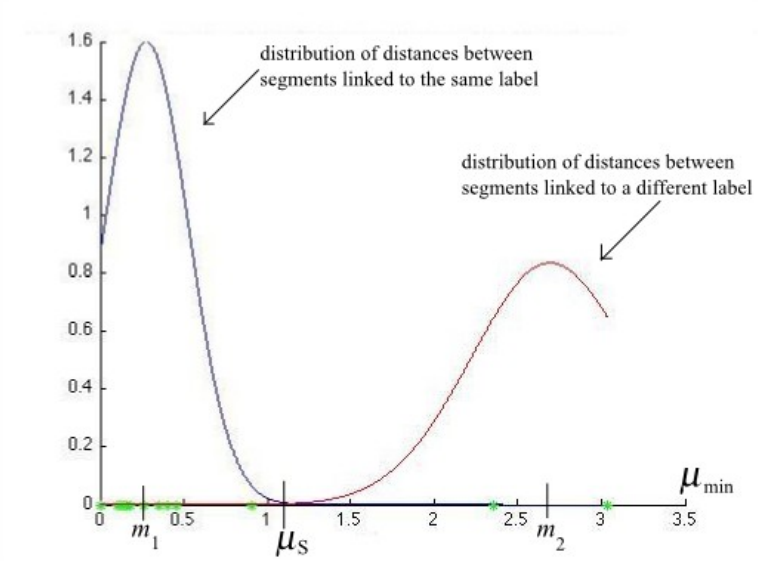
→ which separation between small/large distances?



II.2.d. Adaptive number of iterations

Gaussian modeling of the two distance types :





The separation is estimated as the crossing of the two Gaussians which model each type of distance. N_{estim} is then inferred.

Summary

III.1. Evaluation measures :

- precision, recall, F-measure : quantifies the match between annotated borders (ground truth) and estimated ones (tolerance : 3 s or 0.5s)

- frame-pair clustering precision/recall/F-measure rate : Quantifies the match between pairs of frames which have the same label in the estimated structure with these pairs in the reference structure.

- under-segmentation scores : quantifies the amount of reference structure which is missing in the estimated structure.

- over-segmentation score : quantifies the amount of reference spurious information

→ Tend to 1 when the estimated structure and reference structure match perfectly.

III.2. Evaluation : MIREX - Structural segmentation task

Corpus : MIREX 2009 (297 songs, mostly from *the Beatles*, collected by Paulus, Peiszer and C4DM)

Evaluation (segmentation and labeling):

Participants	F-measure(3s)	F-measure(0.5s)	Pairwise F-measure	Normalized conditional entropy (NCE) based over-segmentation score	NCE based under-segmentation score
Mauch *et al.*	0.6074	0.3246	0.6126	0.7631	0.6101
Sargent *et al._1*	0.5667	0.2172	0.5016	0.5885	0.6920
Sargent *et al._2*	0.5593	0.2193	0.4928	0.7139	0.4284
Martin *et al.*	0.5079	0.1853	0.5546	0.6795	0.5357
Peeters	0.5014	0.1813	0.5359	0.6025	0.6798
Weiss *et al.*	0.4753	0.2004	0.5440	0.6668	0.5406

Corpus : MIREX 2010 (100 songs from RWC POP database, and annotated by Bimbot et al.)

Evaluation (segmentation only):

Participants	F-measure(3s)	F-measure(0.5s)
Sargent *et al._1*	0.6101	0.2433
Sargent *et al._2*	0.6060	0.2521
Mauch *et al.*	0.6051	0.4408
Weiss *et al.*	0.5819	0.3618
Peeters	0.5708	0.2325
Martin *et al.*	0.4864	0.3163

(http://nema.lis.illinois.edu/nema_out/mirex2010/results/struct/mirex10
http://nema.lis.illinois.edu/nema_out/mirex2010/results/struct/mirex09)

Summary

The performance has to be improved, and it implies the choice of common criteria (and ground truth).

Perspectives :
The proposed approach can be improved in two ways :

- songs can have more than one structural period, and a method to estimate the number and the value of these periods has to be developed to improve the segment detection.

- other models for the minimal distances between clusters have to be tested for labeling.

Thank you!