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

Gabriel SARGENT, Frédéric BIMBOT and Emmanuel VINCENT,
METISS project-team, INRIA-IRISA, Rennes, France

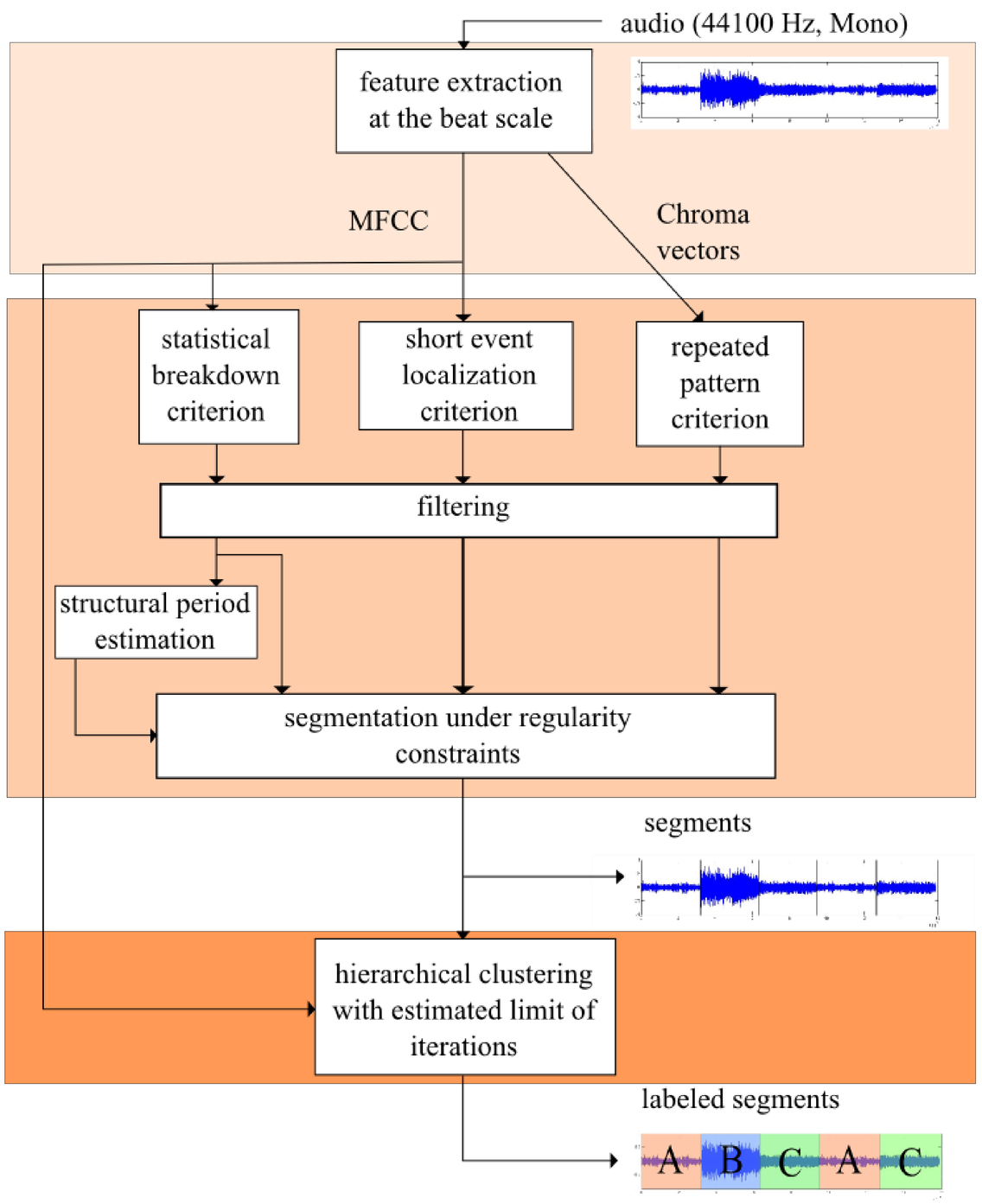
Abstract :

Music structure is a high-level description of a song in temporal segments with comparable musical content, which are sometimes refered as "musical sections" or "musical parts". Two segments related to comparable musical content share the same label (A,B,C…), and the span of these structural segments covers a group of musical bars.

Relying only on the audio, different music structures can be extracted from the same song by different listeners : for example, a variety of segmentation criteria can be chosen over time, such as the beginning of a melody or the changes in the instrumentation...

In this work, we aim at infering automatically the musical structure of songs according to their timbral and their harmonic content. Our algorithm consists in two parts :
- segmentation of the audio file, based on localization of timbral breakdowns, short audio events (using the MFCC features) and repeated harmonic progressions (using chroma vectors).
- labeling of the structural segments, grouped according to their timbral content by a hierarchical clustering with an adaptive number of clusters. Each group is then assigned to a different label.

Algorithm overview :



Features :

- MFCC (Mel Frequency Cepstral Coefficients), interpretable as timbre features (20 coefficients including the 0th one)
- Chroma vectors, interpretable as harmony features (12 coefficients for the 12 semi-tones of the tonal system)

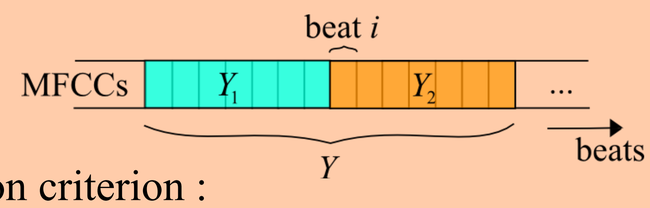
Segmentation process :

Three criteria were chosen to assume border locations. They are based on a Generalized Likelihood Ratio (GLR) approach, which estimates and compares the likelihood of an assumption H_0 to the likelihood of its contrary H_1.
Let Y be the sequence of features describing the song :

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

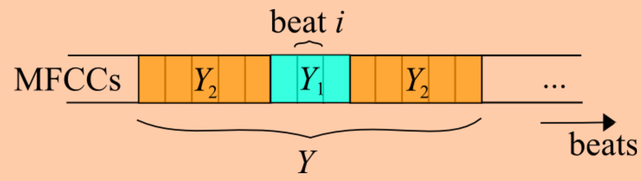
- Statistical breakdown criterion :

- 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.



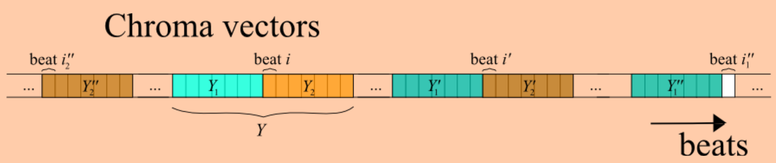
- Short-event localization criterion :

- 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.

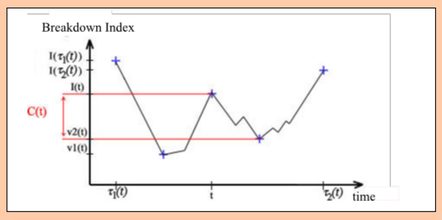


- Repeated pattern criterion :

- H_0 : Y appears entirely elsewhere in the song.
- H_1 : Y_1 and Y_2 appear separately in the song.



Filtering (from Seck [1])



Each peak is compared to its highest closest valley. The filtered criterion takes at each point the value of this difference in height :

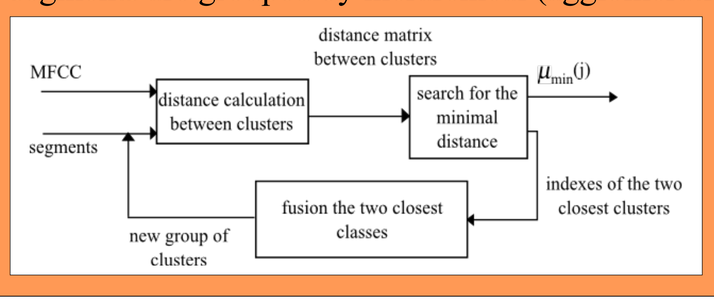
$$v_1(t) = \min_{\tau_1(t)<i<t} I(i) \qquad u(t) = \max(v_1(t), v_2(t))$$
$$v_2(t) = \min_{t<i<\tau_2(t)} I(i) \qquad C(t) = I(t) - u(t)$$

Criteria combination under regularity assumption :

- Estimation of a structural period [2] using the fast Fourier transform of the statistical breakdown criterion.
- Selection of borders 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.

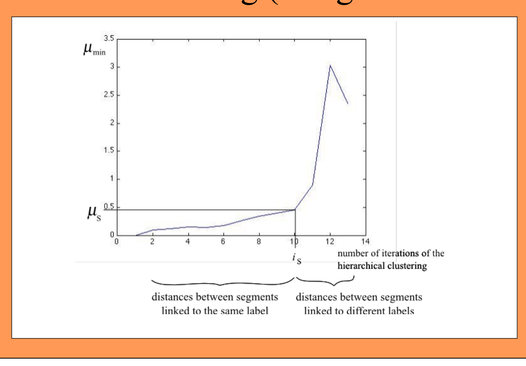
MFCC-based labeling :

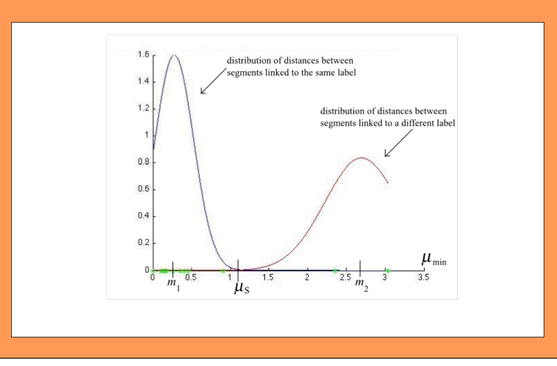
- each segment is modeled by a Gaussian distribution of its MFCC
- models are compared by a symmetrized gaussian likelihood measure μ [3]
- segments are grouped by hierarchical (agglomerative) clustering



$\mu_{min}(j)$ is the minimal distance between two clusters of segments at the j^{th} iteration of the hierarchical clustering

- optimal number of clusters estimated using a bi-Gaussian model of the minimal μ for each iteration of the clustering (using the *K*-means, *K=2*)





Evaluation / Results :

Corpus :
we use 20 songs from various styles (popular, rock, house) chosen and annotated by IRCAM in the framework of the QUAERO project [4]. Q1 and Q2 are the 2 halves of this corpus (Q).

Evaluation metrics :

modeling score (Peeters [5]) : considers the best match between estimated structure and ground truth, and measures the percentage of badly matched (or unmatched) parts of the signal.
precision, recall, F-measure : quantifies the match between annotated borders (ground truth) and estimated ones (tolerance : 3 s).

Results :

The algorithm (1) is compared to another system (2) which has the same structure : the segmentation is done by selecting peaks of a filtered novelty function extracted from a MFCC similarity matrix; the labeling step is the same as (1).

algo.	learning	test	precision	recall	F-measure
(1)	x	Q1	0.550	0.668	0.592
(2)	Q2	Q1	0.572	0.635	0.584
(1)	x	Q2	0.595	0.664	0.615
(2)	Q1	Q2	0.596	0.703	0.627

Modeling score for labeling only, with corpus Q : 0.571
Modeling score for segmentation and labeling : 0.521

Conclusion :

This system 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.

[1] : M. Seck, R. Blouet, and F. Bimbot, "The IRISA/ELISA Speaker detection and tracking systems for the NIST'99 evaluation campaign", *Digital Signal Processing, Volume 10, Issues 1-3*, pp.154--171, 2000.
[2] : F. Bimbot, O. Le Blouch, G. Sargent, E. Vincent, "Decomposition Into Autonomous and Comparable Blocks: A Structural Description of Music Pieces", to be published in *Proc. of the ISMIR Conference*, 2010.
[3] : F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, "Second-Order Statistical Measures for Text-Independant Speaker Identification", *Speech Communication, Vol.17, No 1-2*, pp.177--192, 1995.
[4] www.quaero.org/ [5] G. Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum likelihood approach", *Proc. of the ISMIR Conference*, 2007.