A STRUCTURAL SEGMENTATION OF SONGS USING GENERALIZED LIKELIHOOD RATIO UNDER REGULARITY ASSUMPTIONS

Gabriel SARGENT
Université Rennes 1,
IRISA(- UMR 6074)
gabriel.sargent@irisa.fr

Frédéric BIMBOT
CNRS,
IRISA(- UMR 6074)
frederic.bimbot@irisa.fr

Emmanuel VINCENT
INRIA,
Centre INRIA Rennes
- Bretagne Atlantique
emmanuel.vincent@inria.fr

ABSTRACT

This document presents the algorithm submitted to the "Structural segmentation" task at MIREX 2010. It consists in three parts. First, feature extraction (beat, MFCC, Chroma) from the song is achieved using existing scripts. Second, a segmentation is done according to three criteria for localizing statistical breakpoints, repeated feature sequences, and short events, using a filtered version of generalized likelihood ratio. The segment borders are then selected according to the amplitude of these criteria and a regularity constraint about the length of the structural segments searched. Third, the segments are gathered into similar classes using a hierarchical (agglomerative) clustering. The number of steps of this clustering is estimated separately for each song.

1. FEATURE EXTRACTION

In the framework of MIREX, the songs' sample rate is 44100 Hz. Three features are extracted for each input song :

- beats (estimated by Daniel Ellis's scripts [1])
- MFCCs : 20 coefficients (including the 0th coefficient), with window size = 23.2 ms and window hop size = 11.6 ms (using scripts from MA toolbox by Beth Logan and Malcolm Slaney [2])
- chroma vectors : 12 coefficients, with window size = 92.9 ms, and window hop size = 23.2 ms (using Daniel Ellis's scripts [3])

2. STRUCTURAL SEGMENTATION OF THE AUDIO SIGNAL

It is based on the calculation of three criteria extracted from the audio signal at the beat rate :

- criterion 1 evaluates the presence of statistical breakpoints in the MFCC sequence extracted from the audio file
- criterion 2 evaluates the presence of repeated sequences of features, using the chroma vector sequence extracted from the audio file
- criterion 3 evaluates the presence of short events, which uses the same MFCC description as the criterion 1

These criteria are calculated by a Generalized Likelihood Ratio, which compares the likelihood that the sequence of extracted features Y follows a particular assumption (let's note it H_0), or its opposite (H_1), at each beat :

$$\text{GRL} = \frac{P(Y|H_1)}{P(Y|H_0)} \quad (1)$$

For criterion 1, the assumption H_0 is that the MFCCs contained in a 12 s window centered on the current beat can be well modeled by a single Gaussian distribution. On the contrary, H_1 assumes this group of MFCCs is well modeled by two Gaussian distributions (i.e. one Gaussian distribution models the sub-group of MFCCs located before the current beat, and another one models the MFCCs located after this beat). We therefore evaluate at each beat if it is better to assume H_1 rather than H_0 (it corresponds to high values taken by criterion 1) and this indicates possible assumptions on the location of the structural segment borders (a high value corresponding to a high border probability).

Criterion 2 evaluates if every sequence of chroma vectors contained in a window of size 12 s centered on each beat is completely repeated in the rest of the song (H_0), or if the two halves of this sequence are repeated separately (H_1). The comparison between 2 sequences is made with the Euclidean distance.

Criterion 3 considers for every beat a long window (12 s) and a short window (2 s) centered on each other. H_0 assumes to model the MFCCs contained in the whole long window with only one Gaussian distribution, and H_1 models these MFCCs by two Gaussian distributions : one modeling the MFCCs contained in the long window only - those contained in the small window are excluded - and another one modeling the MFCCs of the small window only.

[1] http://labrosa.ee.columbia.edu/projects/coversongs/
[2] http://www.ofai.at/~elias.pampalk/ma/documentation.html
[3] http://labrosa.ee.columbia.edu/projects/coversongs/

Our assumptions on the location of short events is shown with high peaks on the criterion 3 (which is evaluated the same way than criterion 1).

The lengths of the different analysis windows have been tuned on a development set of 10 popular songs. These values have to be adjusted in the future on a wider corpus.

The three criteria are filtered using the method proposed by Seck in the context of the segmentation of an audio flow into speech and music segments [3]. The peaks of the resulting criteria are used as assumptions on the location of the borders of the structural segments.

The structural pulsation period [2] is estimated by Fourier Transform of criterion 1 (filtered version). Then, the selection of borders is made by dynamic programming, minimizing the amplitude of the 3 criteria between segment borders, combined with a penalty function which increases when segments' length moves away from the estimated structural pulsation period.

3. LABELLING THE STRUCTURAL SEGMENTS

For each segment obtained in the previous process, its MFCC sequence is modeled by a Gaussian distribution. We use the Gaussian parameters of the segments to compute a symmetrized Gaussian likelihood measure [1] in order to compare their timbral content.

The segments are grouped using a hierarchical clustering algorithm [4]: First, each segment is assigned to a different class. At each iteration, the segments of the two classes which contain the most similar Gaussian models are grouped (the new class is modeled by the fusion of the 2 Gaussian models). The clustering is stopped at the iteration i_{S}, which is estimated by modeling the set collection of minimal symmetrized Gaussian likelihood measure for each iteration by a bi-Gaussian model. It is assumed that these measures belong to one of the two following classes :

- the measures resulting from the comparison of two segment classes associated to the same structural label, and
- the measures resulting from the comparison of two segment classes associated to different structural labels.

Classes are determined using a K-means algorithm (with $K = 2$). i_{S} is estimated as the number of elements of the first class.

i_{S} is adjusted by two parameters a and b. We assume that the optimal i_{S} (we note i^*_{S}) and the estimated i_{S} are linked by the following linear expression:

$$i^*_{\mathrm{S}} = a * i_{\mathrm{S}} + b \qquad (2)$$

a and b are tuned by learning on a corpus which contains 20 popular songs, annotated by IRCAM in the framework of the QUAERO project [4] .

[4] www.quaero.org/

4. REFERENCES

[1] 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.

[2] F. Bimbot, O. Le Blouch, G. Sargent, and E. Vincent: "Decomposition into autonomous and comparable blocks : A structural description of music pieces," *to be published in Proceedings of ISMIR 2010.*

[3] 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.

[4] G. Sargent, F. Bimbot, E. Vincent: "Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique," *Proceedings of Journées d'Informatique Musicale 2010*, pp. 177–186, 2010.