Diplomarbeitspräsentationen der Fakultät für Informatik

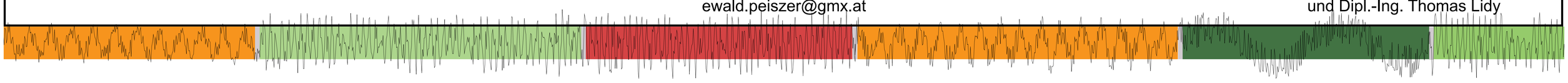
Automatic Audio Segmentation: Segment Boundary and Structure Detection in Popular Music

Ewald Peiszer
ewald.peiszer@gmx.at

Technische Universität Wien
Institut für Softwaretechnik und Interaktive Systeme
Arbeitsbereich: Information and Software Engineering Group
Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber
und Dipl.-Ing. Thomas Lidy

FAKULTÄT FÜR !INFORMATIK

Magisterstudium:
Intelligente Systeme



Abstract

Automatic Audio Segmentation aims at extracting information on a song's structure, i.e., segment boundaries, musical form and semantic labels like verse, chorus, bridge etc. This information can be used to create representative song excerpts or summaries, to facilitate browsing in large music collections or to improve results of subsequent music processing applications like, e.g., *query by humming*.

This thesis features algorithms that extract both segment boundaries and recurrent structures of everyday pop songs. Numerous experiments are carried out to improve performance. For evaluation a large corpus is used that comprises various musical genres. The evaluation process itself is discussed in detail and a reasonable and versatile evaluation system is presented and documented at length to promote a common basis that makes future results more comparable.

Algorithm

22,050 Hz mono audio files

Onset detection

Beat synchronized feature extraction delivers better performance

Feature extraction

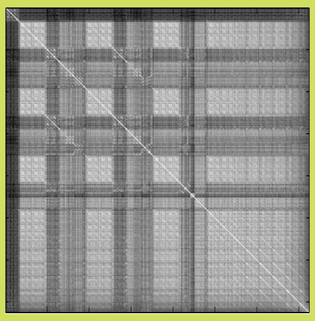
Feature sets used:

Spectrogram, Mel Frequency Cepstrum Coefficients (MFCC), Rhythm Patterns (RP), Statistical Spectrum Descriptors (SSD) and Constant Q Transform (CQT)

Self-similarity matrix $\boldsymbol{S}$

Similarity values of each pair of feature vectors

Repeated segments are visible as bright lines parallel to the main diagonal.



Similarity matrix of *Chumbawamba: Thubthumping*

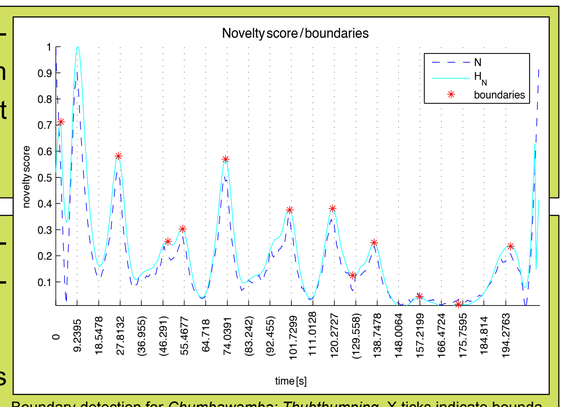
Novelty score N extraction

Gaussian tapered kernel moves along main diagonal of $\boldsymbol{S}$ to detect abrupt changes.

Boundaries $\mathcal{B}$ detection

Low-pass filter is applied ($\mathrm{H_N}$), some heuristics are employed.

Peaks are regarded as segment boundaries.

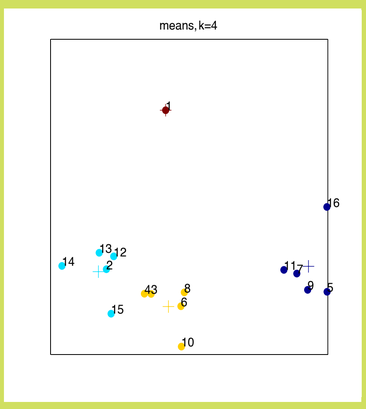


Boundary detection for *Chumbawamba: Thubthumping*. X ticks indicate boundaries from the ground truth. Non-distinct boundaries are in parenthesis.

Musical form extraction (Clustering)

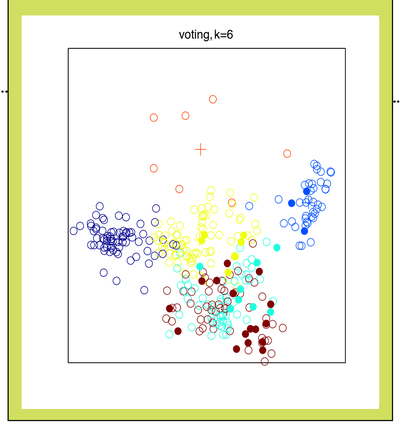
1.) **Means-of-frames**

Segments are represented by means of frame feature vectors.



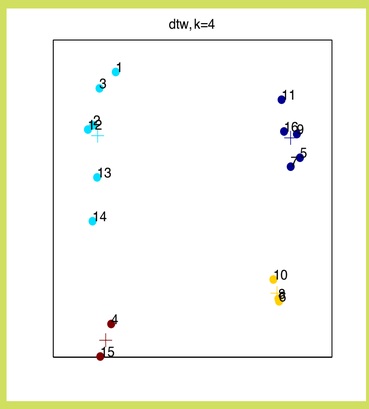
2.) **Voting**

All frame feature vectors are clustered. One segment's cluster is cluster where most of its frames lie.



3.) **Dynamic Time Warping**

Uses temporal information and allows for slight variation of tempo.



Evaluation

Corpus

The corpus contains 108 songs of various genres (dance, rock, pop, R&B, etc.). Experiments and parameter selection have been conducted on 93 songs, the final evaluation was carried out on 108 songs which is the largest corpus used so far in Automatic Audio Segmentation studies.

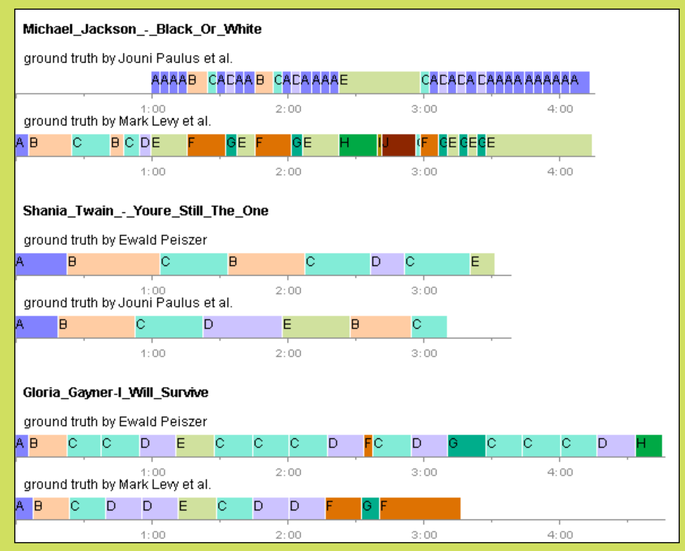
(Michael Jackson, Madonna, The Beatles, The Roots, ABBA, Eminem, Shania Twain, Britney Spears, The Police, Cranberries, Faith No More, R.E.M., Portishead, Scooter, etc.)

Ambiguity

Musical structure is ambiguous. Thus, it is not trivial to evaluate algorithm outcome against "ground truth" annotations.

The right figure shows three songs each with two "ground truth" annotations that have been carried out by two different subjects. Note that the two segmentations of the same song differ to a certain degree.

I decided to carry out evaluation against two-level hierarchical ground truth annotations.



SegmXML

This newly introduced XML annotation file format can contain a two-level-hierarchy of segmentations, as well as alternative labels for song segments. Various metadata information can be stored. Conversion routines to and from an application specific format (Wavesurfer) as well as a XML schema definition file are provided to facilitate the possibleuse by fellow researchers.

Results

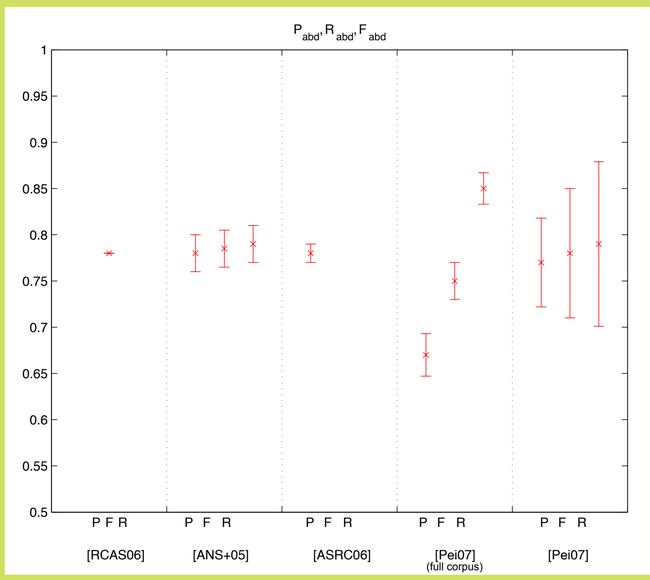
Boundary detection

The figure compares my evaluation results [Pei07] to those of other studies. A small corpus of 14 songs is used (except for the fourth column where my results for the full corpus are visible).

Errorbars indicate 5 % confidence intervals.

Note that there is no significant difference between the results of the 14 song corpus.

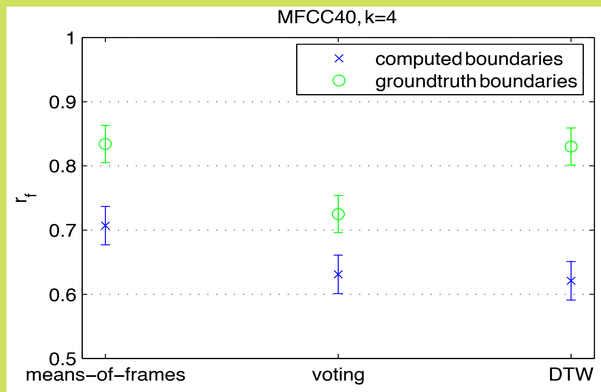
My full corpus results, however, are worse than those of the small 14 song set. This shows that mean performance also depends on the underlying corpus.



Musical form extraction

The figure illustrates the mean performance using different clustering approaches. Green circles indicate results where the correct segment boundaries have been taken from the ground truth.

It can be seen that the means-of-frames approach produces the best results. Note that DTW approach is very sensitive to correct boundaries.



Conclusion

Both boundary detection and musical form extraction are quite acceptable, yet improvable.

The algorithm, however, proved to be robust in a negative and positive sense: Many experiments conducted with various parameter settings and heuristics applied did not lead to a statistically significant improvement of the mean performance.

On the other hand, cross validation and the performance on an independent test set did not show any decline in performance either. Thus, the algorithm presented seems suitable to be applied to a wide range of songs and genres.