Research Proposal for Bachelor's Thesis (15 EC)

Leander van Boven, 6215637
l.m.vanboven@students.uu.nl
Utrecht University

May 2020

1 Introduction & Background

For this proposed research project we, Jesper Kuiper (6203299) and me, are going to do research in the field of musical structure analysis. One of the main tasks of musical structure analysis is the combination of dividing a musical piece into segments, such that each segment forms a musically semantic unit. Each segment is then grouped into meaningful parts. Due to the hierarchical properties[1] of music, the definition and application of musical structure analysis is very broad.
In Western music we can find a lot of high level patterns. The most common ones are showed in Figure 1. For our research we aim to find the patterns that occur most in popular music, for example the patterns depicted in Figure 1e and 1f. This pattern generally includes one or more *choruses* and *verses*, with often an *intro* part at the start of the song and/or an *outro* at the end of the song. Sometimes a *bridge* will connect two choruses near the end of the song.
To obtain these patterns there are two global approaches.
Distance-based segmentation and classification uses the distance between (consecutive) music samples to spot segment boundaries. If a part of music has a high distance value with relation to another part, they are probably two separate segments.
Segmentation by classification first divides the song up into multiple small parts (e.g. beats or segments of a few milliseconds), and classifies each small part. A series of parts that have the same classification is then treated as a big segment.

As these two approaches differ fundamentally from each other, we aim to evaluate and compare the performance of both. My main point of research will be the segmentation by classification approach, and Jesper will research the distance-based segmentation and classification approach.

[1] e.g. a segment that represents one couplet of a song can be subdivided into short melodies, which in their part can each be subdivided into beats, etc.

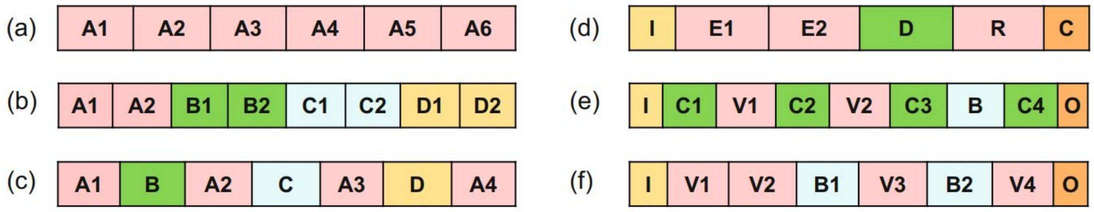

Figure 1: Examples for musical structures as encountered in Western music, taken from Müller [7]. **(a)** Strophic form. **(b)** Chain form with repetitions. **(c)** Rondo form. **(d)** Sonata form. **(e)** Beatles song "Tell Me Why". **(f)** Beatles song "Yesterday".

2 Scientific Embedding

There already is a lot of research regarding the structural segmentation of music. Every year, the Music Information Retrieval Evaluation eXchange (MIREX) is held. The most recent occurrence of music structure analysis at the MIREX was in 2017. Some approaches were submitted to one of the annual MIREX conventions [6], [3], which means that these approaches have been evaluated using the SALAMI data set as well because the SALAMI data set was one of the data sets available.
The segmentation by classification approach is already quite extensively researched. However this approach is most commonly used for classifying pieces of audio as being music, voice, noise, etc. [2, 10, 4, 11]. The basic idea in these applications is to split the audio into very small segments of a fixed time length (usually in the range of 1 millisecond to 50 milliseconds). Each small segment is then represented as a vector by using certain musical properties. These vectors are then used to classify each small segment. Because of the classes that need to be distinguished earlier applications used non-machine learning algorithms ([4, 11]) that used statistical musical properties embedded in these vectors, while newer work ([2, 10]) uses these in combination with deep learning. Common technique is to first differentiate between speech and non-speech audio, the with-speech audio is then classified as silence or pure speech, non-speech audio gets either a music or background noise label. The first step is often performed by support vector machines ([4, 10]), differentiating between music and background noise had better performance when deep learning was used ([2, 10])
Another use of segmentation by classification for musical structure analysis was performed by [5], the difference with above mentioned articles are the initial small segments. In these earlier articles segments with a fixed time-length were used, however [5] first performs note onset detection to create segments with a size proportional to the inter beat time interval of a song. This is done because the boundaries we are aiming to find are almost always at note onset, and will thus mean that there will almost never be two sections like chorus and couplet within one small segment.
The use of segmentation by classification also occurs outside musical structure analysis. [6] uses segments to enhance automatic chord transcription by combining chroma features from similar segments within a song. This further shows how musical structure analysis can be used in other parts of music information retrieval.
The references below also contain some other articles, including summaries, that we think are useful.

3 Research question & Methodology

I, for my part in this research, am going to focus on segmentation by classification. My research question is thus:
Main question: What is the feasibility of a method that uses segmentation by classification for structural segmentation and classification of popular Western music?
Sub question: How does the performance of such a method compare to a method that uses distance-based segmentation and classification?

To find the answer to these questions I will first start by doing more research into the currently used implementations of segmentation by classification in general and musical structure analysis in particular. As described in the Scientific Embedding a lot of current applications are used to classify audio as music, speech, noise, etc, while my application will classify parts of only music. This means I will probably use more of the techniques discussed by [5] for creating the initial small segments (e.g. note onset detection). After this I will do research on the conversion of raw audio of each small segment to feature vectors that can be used by a classifier, this will probably be a combination of the Mel Frequency Cepstral Coefficients together with some other features.
Then I am going to do research into a classification method that uses the context of the music (thus earlier and/or later smaller segments) when classifying each small segment. After all, if the features of the current segment are relatively equal to the features of the previous segment, they will probably have the same class/are part of the same global segment. Recurrent neural networks are gaining popularity for these kind of tasks because of their good performance on temporal data.
These steps will require a lot of prior research, testing and tweaking. If enough time is left I will also try to look into making this approach work in real time, since prior applications were especially designed and used to be in real time, however due to difference between the classes I want to identify and the classes used in prior applications, I expect this to be very challenging if it is even possible.

To optimally compare our results of each approach we will be using the same data set. This is the SALAMI [9] data set, which provides structural analysis for more than 1000 songs. For each song, the SALAMI data set contains high-level segments (denoted A, B, C, ...) along with what this segment represents (e.g. intro, verse, chorus, ...). Additionally, SALAMI provides lower-level segments (denoted a, b, c, ...), such that a high-level segment contains one or more lower-level segments and a lower-level segment can occur in multiple high-level segments. The songs are provided in the MP3 format, which we will convert to WAV.

We noticed that, from our current findings and proposed methodology, we both differ not only in our approach to the same problem but also in the implementation of our respective approach. While I will probably use some kind of machine learning for the classification part of my research, Jesper is currently very focused on non-machine learning / explainable methods for each part of his research. If this will indeed turn out to be our respective implementations of our approaches, we are going to see why this implementation seemed to be the best implementation for its approach and what benefits or drawbacks come with them.

4 Relevance for AI

Our idea to do research in this field started with a colloquium from Richard Evans, about Machine Apperception. Evans and his team are working on a model that is able to understand the rules that underlie simple games like Pacman. We then imagined how cool it would be to apply this theory to music, i.e. that an agent would be able to fathom the internal patterns that underlie a piece of music to an arbitrary level of complexity. You could then see the structural analysis of a piece of music as one of the first levels of complexity.

Finding and analyzing the structure of a musical piece is an important part of music information retrieval and an important step in music processing, both for humans and computers. The structure of a song can be used to parse a song to get more insight in how certain songs are equal to each other, or how certain structures are used for certain kinds of songs or in certain genres. This thus may be used as additional information when finding for example the genre of a musical piece.

Additionally, knowledge about the structure of songs that are produced by humans can be used for generating songs using computers. This information can for example be used as additional information during the generation phase, or as validation information for already generated songs to validate whether these generated songs comply to human musical structure.

For humans, finding the musical structure is quite trivial, because they constantly and often unconsciously adapt themselves to the musical and acoustic properties of what they listen to. However the amount of different musical structures make computational structure analysis a challenging problem.

References

[1] Chris Cannam, Emmanouil Benetos, Matthew E. P. Davies, Simon Dixon, Christian Landone, Mark Levy, Matthias Mauch, Katy Noland, and Dan Stowell. Mirex 2018: Vamp plugins from the centre for digital music.

[2] Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. *EURASIP Journal on Audio, Speech and Music Processing*, 2020. URL `https://doi.org/10.1186/s13636-020-00172-6`.

[3] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:318–326, 2008.

[4] Lie Lu and Stan Li. Content-based audio segmentation using support vector machines. 06 2001. doi: 10.1109/ICME.2001.1237830.

[5] Namunu C. Maddage. Automatic structure detection for popular music. *IEEE MultiMedia*, 13: 65–77, 01 2006. doi: 10.1109/MMUL.2006.3.

[6] Matthias Mauch, Katy Noland, and Simon Dixon. Using musical structure to enhance automatic chord transcription. *International Society for Music Information Retrieval*, 10:231–236, 2009.

[7] Meinard Müller. *Fundamentals of Music Processing.* Springer International Publishing Switzerland, Erlangen, Germany, 2015.

[8] Ewald Peiszer, Thomas Lidy, and Andreas Rauber. Automatic audio segmentation: Segment boundary and structure detection in popular music. Technical report, 2007.

[9] Jordan B. L. Smith, John A. Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. *International Society of Music Information Retrieval*, pages 555–560, 2011.

[10] Saadia Zahid, Fawad Hussain, Muhammad Rashid, Muhammad H. Yousaf, and Hafix A. Habib. Optimized audio classification and segmentation algorithm by using ensemble methods. *Mathematical Problems in Engineering*, 2015. URL `https://doi.org/10.1155/2015/209814`.

[11] T. Zhang and C. . J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4):441–457, 2001.