Unsupervised Online Concept Discovery in Structured Musical Streams

Duarte Teles, Instituto Superior Técnico

Abstract

A narrative piece can be accompanied by music as a way of emphasizing it. In this work, we approach the problem of modelling the thematic structure of music for film content in a streaming scenario. This is achieved through the mapping of relationships between sets of characters and locations in different time windows, elements that we claim to be narrative markers. By connecting similar situations based on the music that is associated with them, we relate narrative events though their thematic similarity.

We introduce a fully automatic method to generate, from one or more movies and their metadata material (script and subtitles), a quality version of the audio that is played together with a set of labels. These can be used to map the narrative of the movie and more generally, as a ground truth that can be applied to other studies, acting as the semantic to the material that they are associated with. We generalize the ground truth in terms of the co-occurrence of labels. This allows us to have a higher level overview of the different narrative guidelines that occur through the movie.

Clusters of similar events act as an anchor point in to which we associate the music that is played. By having a hard label to describe groups of musical features, we can build associations between these groups and give them names. We use these groups to build a global map of relationships between similar movie events, given their shared musical characteristics.

1. Introduction

Music that accompanies a narrative, such as the case of music in films or operas, carries a thematic structure that helps guide the visual content it is associated with. The music chosen for a given conceptual representation carries specific features that imply that there is a mapping between music and themes that underline a narrative. Building a structure that explains the different thematic occurrences throughout a score allows assigning the key narrative points to themes that most emphasise them. Music directors usually use their artistic sensibility to make or choose music in accordance with the dramatic guidelines and narrative of the visual piece, choices that often condition its success. The thematic structure created can vary in complexity depending on the choice of music. An example of this can be seen in classical music: it can be characterized by a broad spectrum of composition, not marked by a single beat that follows the song or by a fixed mode, in contrast with pop music, that is characterized by tonality and repetitions branching from popular music, that make it much easier to process computationally. These aspects make classical music more challenging than other musical genres and so the techniques that work for processing pop music may have worse performance when attempting to segment or to find relevant transitions (Chai and Vercoe, 2005).

Regarding the musical stream timeline, different sound concepts may emerge, depending on the director's choice, such as the introduction of a specific musical segment that is played every time a given character appears. Events like these motivate us to understand these patterns and to develop a method that is able to identify thematic concepts in a musical stream as well as to incorporate newly observed ones, producing a structure that explains these concepts and is able to relate similar patters across the timeline.

1.1. Objectives

We assume a context where there is no prior information regarding the number of themes in the music, choice that is motivated by the fact that thematic structure can vary greatly depending on the context of the narrative the music is following. Consequently, methods capable of dealing with an unknown number of concepts are required. The structural analysis is aimed at being done in an online setting with the goal of capturing the evolution and emergence of concepts in the music, and how these relate to the story that is being told through the main material. This aspect also allows to study: 1. how the observed concepts relate to previously observed ones; 2. to follow the themes that are recurrent during the movies; 3. where they are the most relevant; 4. to map the occurrences of similar musical events. Furthermore, when using online methods, as data arrives, the model should be updated in order to reflect changes made by new observations. This update, in cases where the volume of information received in stream is too large to be kept in memory, is required to be done on a representation of the data observed so far in the stream.

To achieve our goal, we aim at using density based methods as well as non-parametric statistical ones. These prove more complex than traditional segmentation models and will be used with the expectation of capturing patterns in the stream that other methods may not be able to model.

1.2. Musical Meaning in Film Scores

Films, among other forms of content, because they encase a fictional world where a narrative unveils, can have the dramatization of the its story complemented by music, as a way of elevating the narrative that it is being told. Some soundtracks are produced to create an atmosphere so the viewer can be immersed in the world of the movie. It has qualities especially well-suited to contribute to a films' narrative, as mentioned by Gorbman (1987), where "malleability, spatial, rhythmic and temporal values bond shot to shot, the narrative event to meaning, spectator to narrative and spectator to audience".

In the particular case of music composed to serve visual content, musical meaning becomes attached to the visual content and said meaning is retained by the observer when it is listened to in a detached manner from its original format. This detached semantic allows a piece of music, when paired with a piece of visual content such as a trailer or shot of a movie, different from what it was originally produced for, to deliver a similar semantic than when paired with its original counterpart. This property is used throughout many film instalments, whenever the main character, object or location, among others, are introduced in a scene or play a major role in it. There is an association between a given element in the film narrative and a corresponding track, that creates an expectation that that element will take some type of part in the narrative, whenever the music associated to that element is played.

1.3. Document Structure

The rest of this document is structured in the following way: Section 2 overviews the related work that connect to our goal. This includes types of pre-processing that can be applied to the musical stream as well as state of the art models used for semantic extraction. It also covers background on the density-based and non-parametric methods. Section 3 presents all the steps taken to obtain and prepare the dataset. Finally, Section 4 presents the experimental setup of our work, followed by Section 5 and 6 where we show and discuss the results of our experiments and present some closing remarks, respectively.

2. Background and Related Work

In this Section, we first present material in the literature related with the different challenges and goals of our work. We begin by exploring methods necessary to prepare the musical audio for further processing. These include methods for feature extraction targeted to our domain as well manners in which the audio can be portioned in similar segments.

We then cover work in the literature that has dealt with some of the obstacles that we faced, including, novel class discovery, unsupervised clustering, and leitmotif classification, as well as the background models proposed as solutions to model estimation in our work that cover some of the work that utilizes these models.

2.1. Audio Pre-processing

The audio format we are working with is digital audio. A pre-processing step is required with the goal of preforming feature extraction. We require these features to be able to capture the harmonic, rhythmic, timbrel and sequence aspects of the music, as these help characterize a given musical theme. The features can be used as individual frames of the audio or further processing can be done in order to build audio segments or obtain structures such as chordgrams, both more complex objects that carry information of the sequence of frames.

In the case where we are dealing with music produced by an orchestra, the changes in timbre can be informative when attempting to extract semantically significant transitions. Key instruments in the orchestra are predominant in some of the leitmotifs. Because this feature allows us to distinguish instruments, similar patterns in the data are expected to be observed with the recurrence of a leitmotif, when using this feature. Moreover, different fictional cultures, in some cases due to the complexity of one's culture, may use different scales, mirroring what happens in real cultures, as it is the example of Asian and Western music. For a given movie, music can be composed, for example, using the full chroma scale, on the opposite of music composed for another fictional cultures that may carry a different tonality of only 7 pitches, for example, (Rone, 2018). This suggests that this feature might be discriminate for modelling cultural aspects in music.

2.2. Methods for Semantic Extraction

The number of leitmotifs in the stream can grow indefinitely, leading to a problem of novel class detection. In cases where it is impossible to know the entirety of the domain we are dealing with, closed set methods cannot generalize well enough to model classes unobserved at training time and prove insufficient to deal with concept discovery and model sparsity, since observed patterns from a class may change or new concepts may emerge (Gama et al., 2014; Parker and Khan, 2015; Masud et al., 2010). Moreover, when dealing with streaming data, multiple problems require attention. In the first place, we cannot save the entire stream as practical memory issues would arise implying that an abstraction of the data is required. It should, as accurately as possible, capture the distribution of the observed domain. Secondly, a learning method should be able to update, using just newly observed samples and a representation of the previously observed data, also without the need of annotated data. In cases where data is generated at a large scale, annotation or prior information regarding the data may be unavailable and thus the model should be able to structure observations in an interpretable way, depending on the application.

One can look beyond this group of methods to others that provide actual characterization of the statistical distribution of the data. With the work of Rudd et al. (2017), both statistical information from the data and methods to deal with an unknown number of classes are used. The authors' goal is to perform image recognition (multiclass) in a open set environment using open world decision boundaries, where these are used to separate know classes from the unknown space effectively attempting to label samples from an underserved class as such. This approach is based on extreme value theory (Coles et al., 2001), that dictates the form of the functions for the radial probability of inclusion of a point with the respect of the class of another.

Work in the topic of novel class detection has also been done in the field of signal processing. Gharghabi et al. (2019) propose a domain-agnostic online segmentation model for multi-dimensional time series. In this work, time series extracted from motion sensors, for example, are analysed with the goal of identifying meaningful regime changes along a time series, such as detection of the transitions between walking and running or of certain patterns in heart rate. Hence, the semantics captured is shaped in the form of discrete classes. To achieve this, the authors use similarity-join metric for time series. It receives a time series T as input and a subsequence of length L, representing the size of the pattern, and returns two vectors. The first corresponds to the Euclidean distance between the subsequence and its nearest neighbour elsewhere in T (defined as MPValue). The second indicates the location of each of the nearest neighbour of each element of the subsequence in the time series T (defined as MPIndex). These two vectors lead to an annotated time series where one can derive the likelihood of a regime change.

Gjoreski and Roggen (2017) also focus on the discovery of activities such as running, walking or jumping, characterized by sensor signals. Their approach is based on agglomerative clustering and aims at exploiting the temporal information in the signal. The methodology consists on, at a given point in time, keeping a number of active clusters estimated by clustering the frames of a given time window, so that multiple deviations can be clustered into multiple temporally overlapping segments. The total number of clusters in the active pool does not represent the total number of clusters, which is open ended. Each of these clusters has a tolerance that gives the duration the cluster is allowed to exist without being updated (merged or deleted), with the goal of modelling short outliers and a minimum duration that discards the cluster if it was only present for a short period.

There as also been work done on bringing the modelling capabilities of neural networks, specifically CNNs, into the field of incremental learning and novel class detection. This class of models lack the robustness to deal with novel classes, due to the assumptions of closed world datasets with a fixed number of categories. The work of Wang et al. (2019) focuses on addressing this challenge by leaning a feature representation such that distribution of instances from the same class are discriminative enough in order to perform label prediction, novel class detection, and subsequent model adaptation.

Serra et al. (2014) propose a method for music structural annotation using time series structured features and segment similarity. They aim at annotating the structure of a music piece in an unsupervised way without employing explicit knowledge of previously annotated pieces, by detecting temporal locations of segment boundaries and to assess segment similarity based on repetitions. They achieve this by building a model that firstly extracts tonal and harmonic features from the audio. They then transform these into a time series of structured features from which they compute a novelty function whose peaks correspond to boundaries. Finally, the resulting segments are compared in a pairwise fashion and clustered.

Work that resembles our own from the musicological perceptive is the one from (Krause et al., 2020). In this paper, the authors conduct a case study on a dataset covering 16 recorded performances of Wagner's Ring of Nibelung, with annotations of ten central leitmotifs. They build a neural network classification model and evaluate its ability to generalize across different performances and leitmotif occurrences. These motifs constitute the classes of the classification task. Furthermore, all motif occurrences were annotated by a musicologist. In terms of the classification task, the authors define it problem of assigning a given audio excerpt to a class according to the occurring leitmotif, discarding segments where multiple occurrences of different leitmotifs happen in parallel.

This work shows major similarities with our regarding some of its difficulties and goals, specifically the identification of leitmotifs and the capability of producing a model that is able to generalize across multiple interpretations of the same motif or score. Despite these similarities, our task faces broader issues, as we do not posses such a fine grained annotation, nor the the multiple interpretations of the same score. Particularly, the leitmotif can suffer modifications through the material it supports, but further interpretations of these changes may help a learning model generalize each leitmotif better.

Effectively all the methods presented so far provide some insight on some of the methodologies available to us, either tackling the problem of novel class detection or the problem structural discovery. Although they face the same family of problems, them methods presented always make considerations regarding model adaptation, an incremental learning procedure or the use of annotated data, effectively circumventing one or more of these issues. To this effect, the following chapter present methods that can be directly used in our context, proposing solutions for the aforementioned problems simultaneously, without relaxing the problem in a way that allows to circumvent the issues present in our work.

2.3. Density estimation

Mixture models can a be a powerful tool to model uncertainty. They allows us to represent specific parts of a domain. A mixture model can be seen as the weighed sum of individual probabilistic functions and formally defined as:

$$p(x|\theta) = \sum_{i=1}^{K} \pi_i F(x|\theta_i) \tag{1}$$

where K is the number of components in the mixture, π_i is the weight of the component, θ_i is the set of parameters of the probability distribution and $F(.|\theta)$ is the probability distribution parametrized on θ. When using this model in an unsupervised setting, with the premiss that each components approximates a class present in the domain, each sample has only a given probability of belonging to each class, i.e., to a component of the mixture model.

There are two key aspects when using this model. First, there is a choice of the number of components of the mixture model. This number influences how well modelled a partition of the data is. Second there is a choice of the distribution F. How well the model is able to explain the data depends on this choice given the underlying shape of the distribution of the data. If the last is unknown, distributions that can model the variance in the observed data are recommended, hence the Gaussian or the Student-t distributions are good candidates.

One method for approximating a mixture model to a density function is Kernel density estimation (KDE). Generically, it is a method to estimate the probability density function of a random variable. It is defined as:

$$\hat{f}_h = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{2}$$

where K is defined as a non-negative function that integrates to one, usually modelled using a Normal density function. h corresponds to the smoothing parameter or bandwidth, where higher values may lead to oversmoothing.

Based on the described concepts, Kristan et al. (2011) introduce the multivariate online kernel density estimation method. Their goal is to approximate the distribution of the data, explained by a Gaussian Mixture Model (GMM), given an online setting where samples, after observed and processed, are discarded. The authors call this GMM the sample model. The proposed model is based on two key points: the first is that the model is non-parametric in the sense that the number of components is unknown a priori and can can grow given the observations. The second point is that each new observation corresponds to a Dirac-delta function and, during online operation, each new sample is added to the sample model in the form a new component.

There is a clear gain when using this model for our task. Since the number of leitmotifs is unknown a priori, the non-parametric nature of this model is capable of dealing with this aspect. Its explainability through the sample-model also proves adequate as the data is approximated through distributions. Another important point is that we want the number of components in the mixture model to approximate the number of leitmotifs in the musical stream, which we assume to correspond to different classes. The KDE approach gives us no guarantee of this approximation as the number of components may grow as much as needed in order to produce a more accurate explanation of the data.

2.4. Dirichlet Process

The KDE approach faced the unknown number of clusters through density estimation. In the case where we want to approximate the number of components in the mixture to the number of classes (leitmotifs), we propose the use of the Dirichelet process and Bayesian machinery. This model explicitly models the uncertainty of creating new clusters, in our case, new components. These have shown good results in other domains such as the cases of topic modelling and robotics (Nakamura et al., 2015; Nishihara et al., 2016).

The Dirichelet process (DP) is a member of the family of non-parametric stochastic processes. Let (Θ, β) be a measurable space, with G_0 a probability measure on that space. A Dirichelet process $\mathrm{DP}(\alpha_0, G_0)$ is a distribution of a random probability measure G over (Θ, β), where α_0 is a positive number, such that for any finite measurable partition of Θ, the random vector $(G_0(A_1), \ldots, G_0(A_r))$ is distributed as a finite-dimensional Dirichelet distribution with parameters $(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))$ Eq. (9):

$$(G_0(A_1), \ldots, G_0(A_r)) \sim Dir(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)). \tag{3}$$

The various representations of the DP are mathematically equivalent but their formulation differs because they examine the problem from different points of view. We focus on the CRP which provides a simple and computationally efficient way to construct inference algorithms for Dirichlet Process.

The CRP is a preferential attachment model that directly reflects the clustering of draws from the DP. It is defined as a distribution over partitions and is explained by the following metaphor. Consider a Chinese restaurant with an unbounded number of tables. When the first customer arrives, he can randomly select one empty table (cluster), sit and order one dish. Then, the second customer can either join with the first customer and share the dish, or he can start a new table and order a new dish. In this way, when the n_{th} customer arrives, he can select one table from k occupied tables with probability proportional to the number of guests, m_k already seated there, or start a new table with probability proportional to α_0. Formally, the conditional probability can be written as :

$$\mathrm{CRP}(\theta_n|\theta_1, \theta_2, \ldots, \theta_{n-1}) = \begin{cases} \frac{m_k}{n-1+\alpha} & \text{if } \theta_k \text{ exists} \\ \frac{\alpha}{n-1+\alpha} & \text{if } \theta_k \text{ is new} \end{cases} \tag{4}$$

In this metaphor, the tables correspond to clusters and the dishes correspond to the parameters of the distribution of each cluster.

2.5. Language Based Methods

Another approach that can benefit us is the use of n-gram models. Sequences of musical structures, either low level structures, like frames or notes, or higher level structures like chords or segments, contain additional information, because they happen close to each other and more importantly, in sequence. With the premise that sequences of observations carry additional information, we are able to derive a symbolic representation of the audio and use these in the online context. The level at which we build the sequence will heavily impact what our model is learning. For example, sequences of MFCC frames will model changes in timbre along a short period of time.

More specifically, methods derived from topic modelling and language models, in the context of HDP have been used for speech segmentation. Raczyński and Vincent (2014) propose a genre dependent topic model, for modelling chords that aims at predicting a genre of a music using a distribution of chords.

Work has as also been done in the field of word segmentation from phoneme sequences by Takeda et al. (2018). This work aims at building systems that can acquire knowledge during their spoken interactions with human beings. Unknown or new words can frequently appear even if we carefully prepare a vocabulary set in advance. To combat this problem, the authors propose a model based on subword N-grams and subword estimation using a vocabulary set, and posterior fusion of the estimation results of a Pitman-Yor semi-Markov model (PYSMM) and their model. The PYSMM integrates both word-level and character- (phoneme) level N-gram language models and then estimates the segmentation labels of phonemes corresponding to word boundaries by updating both language models in an unsupervised manner.

3. Dataset Preparation

The aim of this section is to cover pre-processing steps required for the construction and preparation of the source materials associated with audiovisual content analysed in this work. Data was extracted from multiple sources, specifically, the audio from the movies themselves, their scripts, the subtitles, the chapter information and finally their respective soundtracks. We therefore do not have a single dataset, but a collection of distinct elements that make up the material related to the movies.

We begin by presenting the dataset, followed by work towards obtaining the musical audio played during the movies. The subsequent section then approaches how we obtain narrative characterization of the events in the movie at any given time and that can characterize the audio that is being played, from that point of view.

3.1. Dataset

We will use the Complete Recordings of the movie adaptation of Tolkien's The Lord of the Rings, by Peter Jackson, containing the complete score for the extended versions of the films. The Lord of the Rings score, composed by Howard Shore, accompanies almost entirely the films, where each track was produced for a given segment of the movie with a thematic background emphasizing how the movie tells the story, therefore enriching it. The score was selected because of the extensive work that has been done in the past analysing its compositional, structural, cultural, and literature background. It was produced solely for the movies, taking inspiration from the source material, the books. It offers around 13 hours of composed music that provide substantial data to work with.

Because of the extensive literature available, concept discovery that is done on this music can be interpreted with contextual story and cultural background, directly bridging the image with the musicological aspects. This aspect will allows to compare the quality of the structure created by our learning methods with the one agreed upon by the literature, giving us a validation tool.

Howard Shore uses similar thematic material through all his work on the trilogy, leading to an opportunity to study not only how the leitmotifs are used and related to each other but also to study how do these relate to the visual, emotional and cultural aspects shown in the movies, an analysis than has been done (not in a computation setting) by Young (2007); Adams (2010). The composer took inspiration from the descriptions that are present in the books, mainly the in depth descriptions of the inhabitants, the instruments used in each region (where each region is associated to a fictional culture, that has different leitmotifs associated with) as well as poetry that is sung by the characters, that show the importance of music within Tolkien's novels. There is also effort put into the novels, in order to deeply characterize the world which also leads to greater cultural background, later used to compose the soundtrack. For example, the exert "Doom, doom came the drum-beat and the wall shook . . . Another harsh horn-call and shrill cries rang out" is depicted in the movie and is accompanied by a musical rich in drum sounds showing a clear inspiration from Tolkien's descriptions.

3.2. Audio preparation

For the dataset used, we have access of two versions of the audio data. The first corresponds to the movie's soundtrack. The second corresponds to the music that is played in the movies, that for editing purposes, motivated by driving the story forward or other creative reasons, does not correspond directly to the music in the official soundtrack. The music present in the movie suffers distortions in terms of energy in scenes where the dialogue is to be more emphasized, for example, as well as being accompanied by other sound effects, such as character dialogue, battle scenes or world events. All these aspects contribute to a decrease in musical audio quality when working with the music present in the movies, compared to the use of the audio from the soundtrack. These aspects motivated to increase the audio quality of music in the movies.

An initial approach to this problem led us to experiment with source separation tools with the goal of isolating the musical audio from the the other audio components. However, as we were unable to successfully isolate the musical audio, we resorted to a solution based on audio alignment. This choice was made so that we can retrieve a high quality audio version of the music that accompanies the movies.

The tool implemented, based on Dynamic time warping (DTW), (Sakoe and Chiba, 1978), takes 20 second audio fragments from the movie and aligns it with the highest score audio fragment of the same length from the soundtrack. The algorithm's score is computed taking as input the chromagram of from the audio segment from the movie and the set of 20 second fragments from each track in the soundtrack. A semi exhaustive search is conducted to find the highest alignment score (a skip of 500ms was implemented to decrease the search space). This is repeated for each movie so that less comparisons have to be made.

For this dataset, prior information that the music present in the movies was played in an order that was respected in the soundtrack. Taking this into account, an heuristic was included when choosing the aligning segment from the obtained alinement's rank. A percentage of the total time of a track must be aligned before a segment from another track can be chosen.

With the alignment process established, each 20 second segment of the movie has a corresponding segment of audio from the soundtrack associated with it. Nonetheless, it is important to point out that this alignment is not perfect. Because of the noise that comes associated with the music in the movie, this can, in many cases, distort the shape of the feature we extract to perform the alignment, therefore negativity affecting it. The numeric values, metrics and features described where chosen based on empirical evidence, such that the audio alignment would be as accurate as possible.

3.3. Ground Truth and Metadata

We set as goal to capture relationships between narrative events that are similar in nature, given their musical audio counterpart. It is important to define what these event are and how we group them together, as these become the ground truth information from where we derive our conclusions. For this purpose, this section approaches how we can derive a ground truth from the metadata that accompanies the movies. Moreover, the approach that will be described is not bound to our dataset. As long as the required material is available, this extraction process can be applied to other audiovisual content.

We begin by extracting speaker and location information, for every instance through the movie, from the scripts and subtitles that are part of our dataset. This was achieved with the use of a tool for subtitle and script alignment, originally developed by Rosado (2016). For the alignment of the script with the subtitles, they use the Needleman–Wunsch DP algorithm, (Needleman and Wunsch, 1970). The algorithm finds an optimal path between two sequences, and then, detects an optimal alignment between them. To use the algorithm with the script and subtitles, first, the script's dialogue and the subtitle's dialogue are tokenized into words, and then, a similarity matrix is created to compare whether or not each word is the same. After the alignment, if the number of words matched between any two sentences is more than 50%, then those sentences are considered to be equivalent. An example of the aliment being computed can be seen in Figure 1.

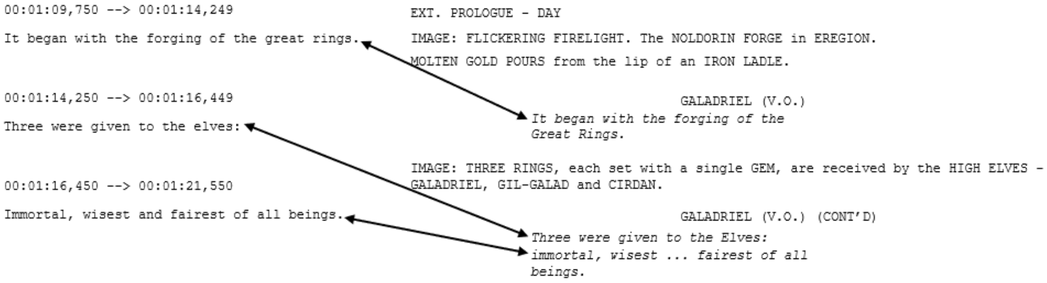
Since subtitles are time-indexed, any alignments with the subtitles stream is implicitly shared with other time-indexed data, such as audio and video streams. This gives us a tool to automatically retrieve pertinent information from the movie scrip. Originally, the tool would only output the speaker for each line of dialogue but throughout this work, it was extended so that the location information present in the script could also be retrieved. For clarification, speaker information corresponds to the name of the characters that are speaking and the locations correspond the the fictional places present in the scenes of the movie.

It is important to mention that, as this process is automated, it suffers from algorithmic errors that affect the quality of the produced match. This in turn, affects the quality of the ground truth produced. Through testing of the tool, it was concluded that it's output had a high recall per alignment, meaning that there is a correct alignment between the subtitles and the script, although the tool misses some of the matches resulting in an empty alignment, which implicates that there are some characters or locations that may not appear as frequently due to this error in alignment.

Another key point in the construction of our ground truth is the use of **scenes**. Originally introduced by Lopes (2017), these are defined as the sequence of utterance or lines of a single speaker and their use is aimed towards capturing the dialogue (consisting of one or more subtitles) of a single character or set of characters that are very close in time, based on the timestamps of the subtitles. We based the scene construction criteria on the previous work, as the authors expensively tested which values would fit scene creation best. The values used were 500ms as the maximum distance between two subtitles than belong in the same scene.

Rather than associating character and location at the subtitle level, to the corresponding music segment, we use the scene as a way of aggregating this information. This higher abstraction level, allowing for the association of a longer piece of audio.

Having the metadata information associated with each scene, another key decision was scene aggregation based on either the repetition of the exact same character(s) in distinct adjacent scenes or the repetition of location in the



```
00:01:09,750 --> 00:01:14,249

It began with the forging of the great rings.

00:01:14,250 --> 00:01:16,449

Three were given to the elves:

00:01:16,450 --> 00:01:21,550

Immortal, wisest and fairest of all beings.
```

```
EXT. PROLOGUE - DAY

IMAGE: FLICKERING FIRELIGHT. The NOLDORIN FORGE in EREGION.
MOLTEN GOLD POURS from the lip of an IRON LADLE.

                    GALADRIEL (V.O.)
          It began with the forging of the
          Great Rings.

IMAGE: THREE RINGS, each set with a single GEM, are received by the HIGH ELVES -
GALADRIEL, GIL-GALAD and CIRDAN.

                    GALADRIEL (V.O.) (CONT'D)
          Three were given to the Elves:
          immortal, wisest ... fairest of all
          beings.
```

Figure 1: Example of alignment between subtitles and script.

same conditions, therefore establishing two distinct ways of grouping information, each giving more weight to their specific aggregation key (character or location).

The decision to aggregate scenes was based on the knowledge that there are pieces of audio playing in the movie were the are no speakers present, implying that there is no annotation available for these segments. In order to leverage the annotation present per scene to other segments not covered by any subtitle and subsequently by any scene, we aggregate them as mentioned above, so that segments of audio between scenes that share the same metadata information may also share that annotation, thus increasing the overall amount of music that has ground truth information associated with it.

Given the established aggregation possibilities, we opted towards aggregating scenes by characters, as described above. We found this solution to be a good balance in terms of granularity of ground truth information being grouped. Aggregation by location resolved in very long sets of scenes with different characters, as the narrative in the movie can take place in a single location for a long period of time.

Finally, we point out a processing step in regards to location information present in the ground truth. The locations present in the script and subsequently in the annotation correspond to geographical locations from the fictional world. Due to the observation of fine grained locations present through the scripts, that we considered to be of too much detail, we opted to encase some of these locations into broader corresponding ones, always referring to the map to make such decisions. Moreover, it is important to note that this change was had-doc for this particular dataset and possible to the existence of a detailed map of the world, that allowed us to make informed decisions. Such change is possible for other datasets as long as geographical information is present.

3.4. Clustering Labels

This processing pipeline, described previously, leaves us with a set of metadata information that is directly associated with each audio instance, i.e., for each set of aggregated scenes, we have the corresponding music that is played in that time interval. Specifically, location and character information. The two elements combined help characterize the narrative of the movie at any given moment, thus being possible to view these as elements that characterize a situation in the movie. There can then be an association between the audio and any of these items, either one-to-one or one-to-many.

The leitmotifs and musical audio in general suffer modifications through the narrative. This implies that the audio that is associated with a character or a location is not uniform throughout the movies. For example, the music from Minas Tirith used in the first movie, where Gandalf is present is very different in theme from the one present in the third movie, with the same character. I can also be seen that the same pair charter-location hold different narrative meaning in these two occasions and because of this difference, the underlying audio also changes tone. These two aspects motivates to go beyond the association of a piece of audio to pair character-location, as we find this association insufficient.

To cover this issue, we follow the approach of Chollet (2016), where the author relies on matrix factorization to reduce the dimensionality of the target labels. This method makes use of co-occurrence of the target labels, projecting the high-dimensionality target vectors. Formally defining this technique, let M be the binary matrix of aggregated scenes I and labels L where $m_{ij} = 1$ if i_i contains label l_j and $m_{ij} = 0$ otherwise. We then use matrix M to compute the Pointwise Mutual Information Gain (PPMI) for the set of labels L, that we will denote as matrix X. Let L_i be the set of scenes associated with label l_i, the PPMI is defined as:

$$X(l_i, l_j) = max\left(0, \log\frac{P(L_i, L_j)}{P(L_i), P(L_j)}\right) \qquad (5)$$

where $P(L_i, L_j) = |L_i \cap L_j|/|I|$ and $P(L_i) = |L_i|/|I|$. Intuitively, the PPMI gives us the association measure between a pair of discrete outcomes x and y. In our case it measures the association between aggregated scenes and a context by calculating the log of the ratio between their joint probability and their marginal probabilities.

X is then factorized using Singular Value Decomposition (SVD) in the form $X \approx U\Sigma V$. Let Σ_d be the diagonal matrix containing the the top d singular values, and let U_d be the matrix obtained from selecting the corresponding d columns from U, we build the matrix $C_d = U_d \cdot \sqrt{\Sigma_d}$

that corresponds to the label factors in d dimensions. The item factors are obtained in similar fashion defined by the matrix $F_d = M^T \cdot C_d$. These two matrices encode the aggregated scenes and labels in the same projected space, respectively. Thus, a distance measure can be used to aggregate scene embeddings and labels. Similar labels are grouped in space, and at the same time, scenes with similar sets of labels are close together. In Figure 2, we can how both matrices project in a common space.

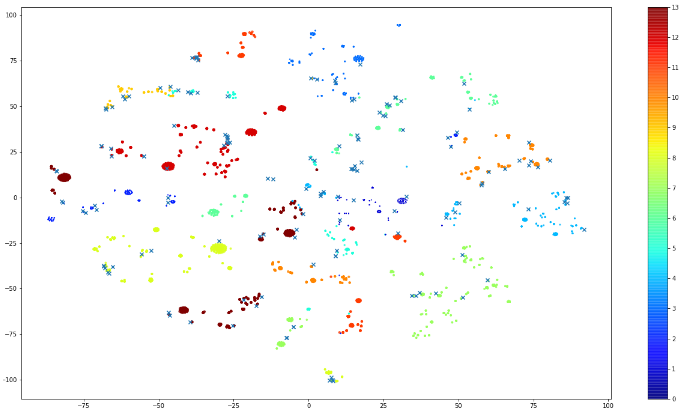


Figure 2: Crosses correspond to elements of the matrix C_d. Dots correspond to items of matrix F_d and are coloured given the cluster centroids they are closer to.

Clustering can be done in two distinct manners. The first is done by clustering aggregated scene embeddings, corresponding to the rows of matrix F_d. Aggregations of these items can be viewed as which aggregated scenes share similar labels. The second and more interesting type of aggregation made possible in the cluster of label embeddings, belonging to the matrix C_d. These clusters represent the metadata information that co-occurs, hence, that can be related explicitly. In our work, these labels can correspond to locations, characters and temporal information in the form of movie chapters or a fixed temporal window. If one is to include these types of information as labels, we can encode which different aspects of the narrative and merge them for a more meaningful representation of the narrative of the movies.

The introduction of the temporal window allows us to differentiate the same charters at the same locations in different periods of the movie. This differentiation is wanted as due to narrative aspects, the audio being played, in the same region can evolve pushing us to make this type of distinction. Thought development, we experimented with the introduction of time via chapter information and by fixed size temporal segments. We concluded experimentally that 10 minute window segments created clusters where the items present were more compatible. Moreover, this type of information is common to segment the narrative of movie. In disk format of distribution of movie, it is common to have the notion of chapter information built in. In the Lord of the Rings movies, these average around five minutes, length that we found to be too short. By having temporal segments greater that chapter length we effectively grouping this information.

Finally, the quality of the clusters produced is focal for the rest of this work. The number and shape of the label clusters is very affected by the method used, thus it requires attention. Clustering can be performed directly on the items of matrix F_d or on the labels of matrix C_d. By doing this, we explicitly use the label embeddings to group aggregated scenes, as these are only implicitly shaped by label co-occurrence.

Two approaches were considered for clustering the labels on the matrix C_d. The first was to use KNN (k-nearest neighbours algorithm) with a number of clusters chosen empirically. This corresponds to our hard clustering approach, in the sense that the only metric in question is the distance between the label points. Because of these points, it presents two clear limitations: the choice of k, the number of clusters, and the absence of modelling variance or uncertainty of grouping labels together. To mitigate these limitations, a second clustering approach based on the KDE model was considered. By training one KDE with label data, the number of clusters grows has needed in order to build better explanation of the underlying data. Like wise, has the mixture components that make up are Gaussian, the uncertainty of grouping labels is being taken into account when building the clusters.

As the clusters were built using matrix C_d, it was necessary to cluster the items of matrix F_d given the computed sets of labels. In the case of KNN model, the centroids were used to label each item from F_d. In the KDE case, we assigned the cluster whose component yield the highest likelihood.

Regarding the clusters obtained with KDE model, we found that the number of clusters was correlated with the dimensionality d of the SVD decomposition. The number of clusters grew the higher the dimensionality. Moreover, label clusters that were more representative of overall number of aggregated scenes preset in the movie presented a higher weight in the corresponding mixture component.

As a final clustering approach, for the case on KNN clustering, we set the dimensionality of matrix factorization to 20, and number of clusters to 14. In the case of KDE approach, the dimensionality of the matrix factorization was set to 5, yielding 21 clusters.

4. Experimental Setup

In this chapter, we present our approach towards leaning relationships between situations with the same narrative context, in an unsupervised fashion, given the soundtrack that accompanies the movie narrative. We assume that different situations, defined in Chapter 3, as characters in a given location and point in time, because of the co occurrences of these elements, share similarities from the narrative point of view, thus sharing similar music. It is with the use of these similarities that we can map, depending on the perceptual features used, how two event scattered across the movies relate based on the audio that accompanies them.

In figure 3 a diagram of the pipeline of our work is showed. The last chapter covered the initial steps displayed, specifically the steps of music alignment, metadata extraction, and label clustering. This chapter covers the subsequent steps.

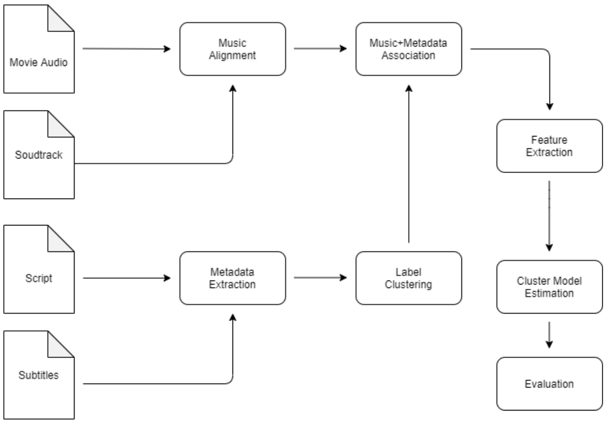


Figure 3: Diagram of the pipeline of our work.

The first part of this chapter covers the feature extraction step, necessary for all subsequent work processing the audio features. The rest of the chapter covers our approach to the construction of the explanation of these audio relationships. To infer these, for each label cluster, we fit a model to the corresponding audio features in order to obtain a statistical representation of each cluster's audio. This is done in two different manners: 1 a baseline approach using multivariate Gaussian distributions; 2 A construction on the baseline using KDE model. Finally we describe how evaluation process was done.

4.1. Audio Setup

A pivotal choice in our work is the choice of which label clusters to use. The construction and analysis of audio relationships are built on top of these and so, the method used to generate the clusters must be weighted. Chapter 3 presented two alternatives for cluster computation. One based on hard clustering where the choice of the number of clusters was made empirically (so that the sets of labels inside the clusters and meaningful based on the knowledge of the movies present) and one obtained via an unsupervised method that accounts for variance of the labels present in each cluster. Based on these these two types of clustering approaches, the experiments presented here take into account both clustering scenarios.

Another important choice is how we chose to model the musical audio. In the last chapter, it was shown that the end result of of dataset preparation in terms of scene aggregation and it's respective audio were acoustic segments of distinct lengths with a set of narrative labels as a direct correspondence. These segments, in many cases, were too large to process, being in the order of minutes, aspect that motivated their split into smaller ones. By working with smaller sized segments, we aimed at capturing pieces of melodic content with fewer notes, which reduces both variability and complexity of the musicals segment and increases precision when extracting feature information. We set the size of each audio segment to 5 seconds.

The audio from each aggregated scene was then split into 5 second segments, with padding of zeros being added to segments that were smaller that the decided size. With audio segment size stipulated, all feature extraction and processing was done at this level.

4.2. Feature Extraction

The features selected were motivated by both the literature and the ability to distinguish, orchestral music in terms of timbre and tonal aspects. Specifically, the features extracted were MFCC, Chroma, chords, and a combination of MFCC+Chroma. Mfcc and chroma features were extracted with a fast Fourier transform (fft) window size of 2048, a hop length of 1024 between frames and a sampling rate of 22050 htz. Regarding MFCCs, the number of coefficients was set to 19: setting the number of coefficients to 20 and removing c0 coefficient (which indicates the average power of the input signal). In terms of the chroma feature, 8 octaves were considered and the number of ”chromas” per octave was set to 12. Normalization was applied to the combination of Chroma and MFCC with respective variance, setting the mean to zero and variance to one, since the numerical range of each feature is different.

For chord extraction, we followed the approach of Müller et al. (2012). The first step, consisting of chormagram extraction, used the setup just described. The following steps generated, for each 5 second audio segment, the set of chords from the from the corresponding twelve major and twelve minor triads. As these correspond to discrete features, and to leverage additional information from the set of chords in each audio segment, we computed the TF-IDF (term frequency–inverse document frequency (Salton and McGill, 1986)). This gives us how important a chord is to an audio segment given the collection of segments available.

The aforementioned features, however, do not take into account melody information as they do not model sequence in any form. This issue was approached in two different manners. The first was to take the mean and variance of the frames from each five second segment. The second was to encode the audio sequence into an embedding with the use of an autoencoder tool. Regarding the first choice, the mean poses as a relevant mechanism to model the overall aspects of the audio segment. The variance was added in order to account for variation of information in the audio segment. Both were computed by fitting a multivariate Gaussian to each the set of frames that for the segments being modelled. In terms of dimensionality, the diagonal of the covariance matrix of the fitted Gaussian is used and it's concatenated to the mean value of the feature for the given segment.

Regarding the encoding of audio segments, the approach previously described in Chapter 2 was followed. For this setup, for each five second audio segment, the mel-spectrogram was extracted (with the same parametrization of the above features and 320 bins as suggested by the authors) and fed to the autoencoding network, with the base parametrization proposed. After training was complete, for each audio segment, the embedding was extracted from the hidden layer connecting the encode and decode layers.

4.3. Computing Relationships

Three different scenarios were considered in this work, using the features pointed above and the two types of clustering labels presented. As a baseline approach, for each label-cluster, we fit a multivariate Gaussian to the audio features associated with the cluster, thus obtaining the mean and variance values for the corresponding features, for each cluster. Relationships between clusters were computed as distance between the distributions that model the audio features of each cluster. We used the Bhattacharyya distance (Bhattacharyya, 1946) to compute this association. This procedure was implemented for each set of features presented above and with both types of label cluster generation.

A step up from these baselines was the replacement the Gaussian distribution with KDE model. By doing this, we are effectively changing how we build the audio explanation for each label cluster. To measure the distance between KDE model of each label cluster, we used the Hellinger distance (Hellinger, 1909).

4.4. Evaluation

The evaluation procedure looked at the distance between the underlying models of each label cluster in order to assess if the distances reflect acknowledgeable relationships. For each setup, we computed a distance matrix between all label clusters and built a chord diagram. In the chord diagram, each entry corresponds to a label cluster and holds three outgoing edges corresponding to the top three most relevant relationships. The width of each edge is proportional to it's relevance. Each edge was computed as the inverse of the distance between two label clusters, so that closer edges, i.e, ones that share stronger relationships, appear with wider edges in the chord diagram.

The work presented in this document is evaluated in a qualitative fashion, as it is challenging to quantify if the associations captured by the top are meaningful without a ground truth of said associations. It is because of this challenge that evaluation requires knowledge of the movies and score of the Lord of the Rings ,in order to assess if the relationships captured are relevant. Furthermore, the assessment is also dependent on the features being used to characterize the audio of each cluster. Depending on these, one can look at the audio segments closest to the mean in the Gaussian case, or of higher probability function values, in the case of the KDE model, to complement the relationship assessment and to answer what each model is capturing. Therefore we also listen to these audio segments as part of our evaluation, to determine, depending on the features used, what is being grouped together.

An example of chord diagram analysis can be seen in Figure 4. The diagram displays a set of highlighted connections. These correspond to the most relevant relationships between cluster 3 and all other clusters.

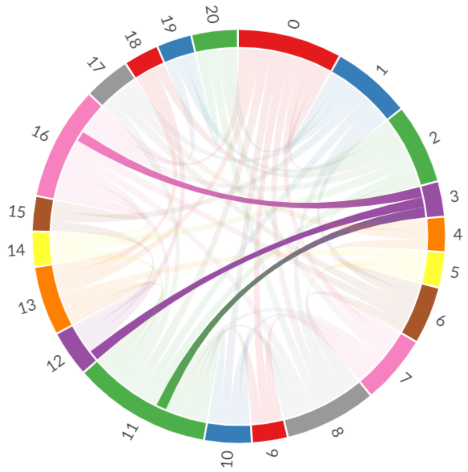


Figure 4: Top relationships computed between cluster 3 and all others. Thickness of edge represents more weight.

5. Experimental Results and Discussion

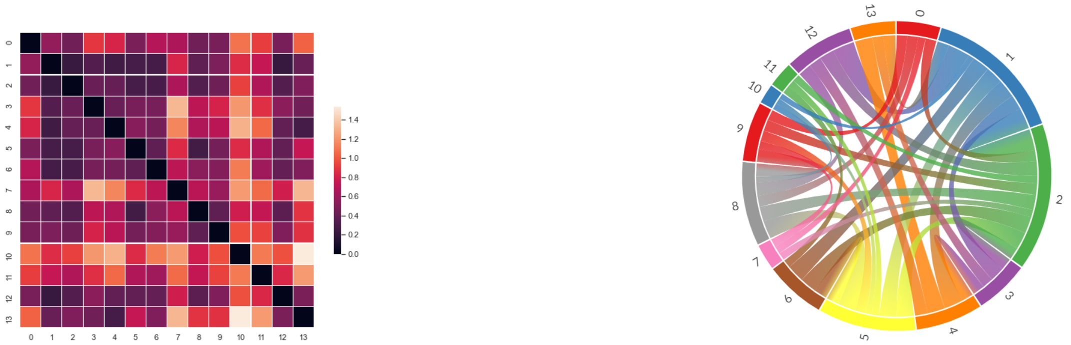
We present all the relevant experiments and corresponding results, highlighting the impact of using different features for our task as well as how changing the method used to model the underlying audio features of each label cluster influences the identification of relationships between these clusters. We begin my mentioning some of implementations necessary to build and evaluate the experimental setup.

5.1. Results

This section showcases the computed distance matrices and respective chord diagrams for the most relevant results, following the experimental setup previously described. We start by presenting some of the baseline scenario. First of all, Figure 5 shows a baseline experiment done using KNN clustering on the label side. The audio features of each label cluster, in this case Chorma, were modelled by fitting a multivariate Gaussian.

Figure 6 shows a increment to the other baselines experiment. The clustering on the label side was done using KDE clustering. The audio features of each label, in this case Chorma+MFCC, were modelled by fitting a multivariate Gaussian to the data of each label cluster.

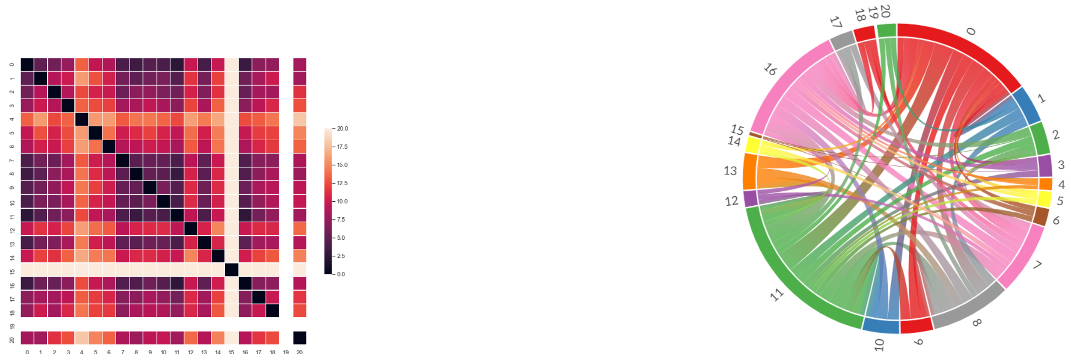
For the setups where KDE model is used for clustering in the label space and as a model of the audio of each



(a)

(b)

Figure 5: (a) Distance matrix computed using chroma features and with KNN clustering on the label space. (b) Chord diagram of top relationships between clusters.



(a)

(b)

Figure 6: (a) Distance matrix computed using Chroma+MFCC features and with KDE clustering on the label space. (b) Chord diagram of top relationships between clusters.

cluster, the distance matrices do not contain the full scale of the Hellinger distance $(0 - 1)$. The choice to shorten the presented range was done so that relevant differences between clusters would be more evident.

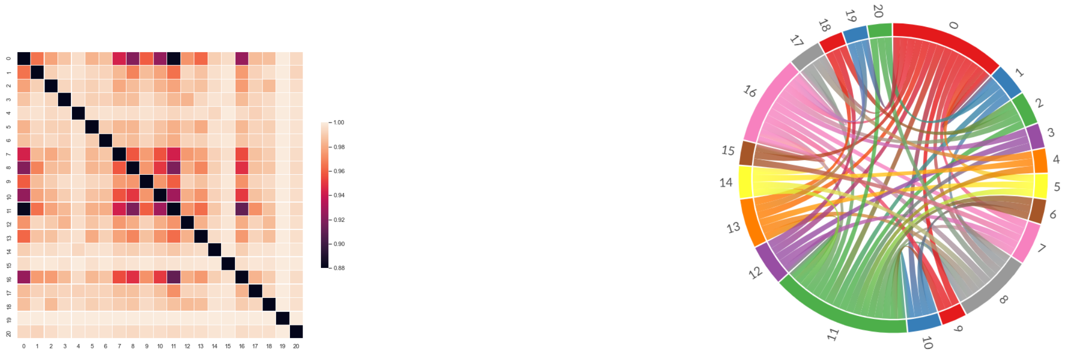
Figure 7 shows the first experiment done changing both the label cluster approach and the way the audio features of each cluster are modelled. KDE clustering was used on the label side. The audio features of each label, in this case Chorma+MFCC, were modelled by fitting a KDE model.

5.2. Discussion

It is always possible to compute distances between clusters and the challenge of this evaluation is to comprehend if these relationships in fact suggest that the music that is present in each of label clusters is able to relate similar situations across the movies. The presented results goal is to capture the differences that the clustering technique on the label side, together with how the audio from each cluster is modelled, help improve the ability to map these relationships.

We begin by looking at our most simple setup, where hard clustering is used to produce the label clusters and a multivariate Gaussian models the audio features in each one. The distance matrix shows us that the majority of the connections between label clusters do not differ greatly in size. Similarly, when using the KDE clusters and a Gaussian to model each cluster, the same behaviour is noticed. This lead us to conclude that increasing the complexity of the method used to create the ground truth is insufficient to obtain sufficient characterization of the label clusters.

The agglomeration of relations on some of the clusters, as it is the case of clusters 0, 11 and 16 in 6 can be explained by reading into the cluster composition. Although the situations that were aggregated share similarity from a narrative point of view, if the music range inside the cluster is too large, it will bring distribution that model the cluster closer to all others. As a concrete example, Figure 8 gives a higher insight into the content of cluster 0. The power law behaviour observed was found amongst all clusters and is another indicator of the statement above. Cluster 0 presents elements from both the fellowship (Aragorn, Gimli, Legolas) and secondary characters that interact with them. These three character are predominant throughout the three movie instalments and the audio that is shared is varied in theme. The same point occurs in the case of Galadriel. The top occurrence, in the prologue location, contain music from different themes, including "The Ring", "The Ringwraiths", and "The Fellowship of the Ring", showing the vast variability inside a



(a)

(b)

Figure 7: (a) Distance matrix computed using Chroma+MFCC features and with KDE clustering on the label space and KDE used model the underlying audio of each cluster . (b) Chord diagram of top relationships between clusters.

single label.

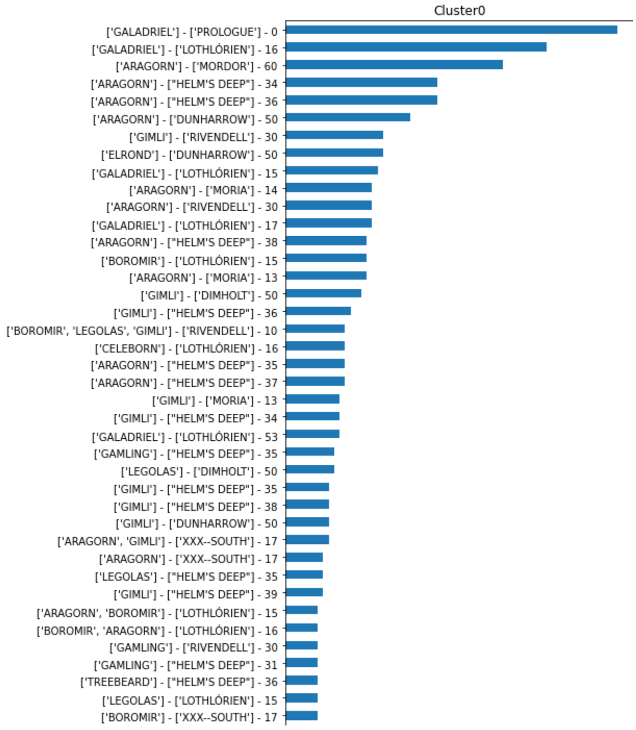


Figure 8: Content of label cluster 0. Labels are sorted by frequency. Longer bar indicates higher frequency.

A mention can be made towards the difference in colour when inspecting the distance matrix regarding the two lighter clusters, 11 and 12. Because their size is reduced compared to other clusters such as cluster 1, the amount of differentiation is a lot smaller, thus pushing the distance to diverge from other in the distance matrix. This observation is also corroborated by to top audio segments on each of these clusters. Cluster 11 is located in a narrow section of the third movie, where the music is predominately themed after battles (the scenes of the battle of Minas Tirith). The specificity in theme is what sets these clusters aside from others.

In comparison to the other setups, the clusters obtained with KDE clustering on both the label side and the audio side revealed to be more differentiable. In regards to the distance matrix, very dark spots, isolated from others, can be clearly seen, implying that the added complexity when explaining the underlying audio of each clusters helps discriminate difference in the underlying audio. This can be explained by the fact the each mixture component as well as the underlying mixture contained in (composed of two Gaussian that work as a memory factor) are able to model different areas of the feature space, not losing the relative importance of some areas, compared to a single Gaussian.

When evaluating the chord diagrams for this setup, we can see that in comparison with the baselines, the number of connections with some of the clusters diminishes. In Figure 6, clusters 1 and 2 are very predominant and display connections with almost every other cluster, something that is less apparent when using the KDE model. One example of label clusters that share particular relevant connection is the case of clusters, 1, 8 and 20, in Figure 7, that share relationships among them. When inspecting the top labels inside the cluster, we can see a connection between a cluster that heavily groups situation in Hobbiton, connected to a cluster that groups events where Gandalf and Frondo co-occur, also connected to events where Sam is present. With knowledge from the source material, we can say that these charters and locations are highly connected and more importantly they share a lot musical themes, in particular, the ones connected to the Shire. The fact that these relations all occur in the top relations displayed by the chord diagram, imply that we can relate groups on similar narrative situations using the underlying music that follows them.

As a final point, regarding the use of audio embedding, these caused numerical problems when estimating both the multivariate Gaussian and the KDE model, not enabling us to follow the proposed experimental setup using these features.

6. Conclusions and Future Work

Given the results discussed in Section 5, we now overview the major contributions of our work, as well as discuss the limitations of the solutions proposed. We conclude by pointing future work that can aim towards improving some of the limitations found.

6.1. Conclusion

We defined the goal of modelling the thematic structure of music for film content in a streaming scenario through the mapping of relationships between sets of charters and locations in different time windows, elements that we claim to be narrative markers. By connecting similar situations based on the music that is associated with them, we relate narrative events though their thematic similarity.

We presented the musicological aspects that give music the narrative properties and the leitmotif as a key role in this aspect. The contributions of this work consist on the introduction of a fully automatic method to generate, from one or more movies and their metadata material (script, subtitles), a quality version of the audio that is played together with set of labels that can be used to map the narrative of the movie and more generally, as a ground truth that can be applied to other studies. We proposed a method to generalize the ground truth in terms of the co-occurrence of labels. This allows us to have a higher level overview of the different narrative guide lines that occur through the movie.

We then use these clusters of similar events as a way as an anchor point in to which we associate the music that is played. By having a hard label to describe groups of musical features, we can build associations between these groups and build a network of relationships.

We found that the KDE model allowed us to build a more detailed and complex representation of the audio since since its parametrization is more complex compared to the two parameters that shape a Gaussian distribution. Therefore, the relationships found were less prone to noise from the mean, when compared to baseline results.

Multiple challenges were identified: the complexity of the object we want to model and how it can be captured with different levels of features; novel class detection without the use of labelled data; what representations we can use to model the observed clusters.

In regards to the audio alignment tool, this can be seen as both a contribution and a limitation to the work developed. Aligning the musical audio with the soundtrack brought us the benefit of being able to work with high quality audio but at the same time introduced error in the work pipeline. Through inspection of the audio segments aligned, the error was introduced when the volume (translated to energy) of the music being played in the movie was, in many cases, distorted by other audio events or lowered, resulting in a poor alignment.

The method proposed for ground truth construction shares similar aspects. On one hand, it brings the contribution of generating a set of labels that act as anchor points to the narrative of the movie. On the other hand, it brigs the limitation where only musical segments that are covered by the sets of aggregated scenes have an associated set of labels. Music fragments outside these windows are not considered, which is leading us to not take full advantage of our dataset.

In regards to the experimental setup conducted, we were able to map meaningful relationships between groups of similar events and improvements were observed from the baselines, showing a more complex explanation of the audio of each cluster did improve the overall understanding of the relationships present.

One of the goals established in the beginning of this work was the online setup. The work done so far can be seen as a baseline towards the online setup. The construction of the ground truth poses a particular issue, since, as the musical stream grows, so due the narrative contexts.

Regarding preliminary experiments in online setup, using the musical stream, the interpretation of the results proved challenging. It is easy to point out that the clusters produced are dependent on the features extracted from the audio, however, after inspection, the groups constructed and the relations between them were not meaningful to describe relationships between similar narrative events.

A similar conclusion was made when attempting to use more descriptive models to explain the relationships between label clusters. We sought to use topic modelling techniques for this purpose, as they give a richer description of relationships through the statistical importance of each topic. Much like the previous scenario, the interpretation of each topic from the mixture of topics is very challenging and the experiments conducted lacked this very interpretation in order to build a coherent connection between narrative entities given the corresponding music, thus motivating future work in this direction.

Finally, a point not addressed in the experimental setup was the use of Dirichelt process methods. Although we take from these the advantage of the unbound number of cluster and the ability to learn these in an online setting buy changing the sampling scheme, these methods share the same challenges as topic modelling techniques (they can be seen as an extension of models such as the LDA when we want to grow the number of topics). When building clusters using just the audio, either in a stream scenario or not, the problem of cluster evaluation emerges. As less complex setups such as the ones described above proved very challenging to evaluate, this technique was not pursued in this work but is left as a possible tool to solve the unbound number of clusters in future work.

A straight comparison to the state of art cannot be established, as to our knowledge, there is no work in the same setup as ours. Although we can not find a direct comparison, we can consider future work to several challenges faced through the development of our study, as we will discuss bellow.

6.2. Future Work

The automatic alignment tool between the movie's music and the corresponding soundtrack poses as an element is this work that can have further improvement. The errors in alignment introduce noise on posterior analysis using this data, something that needs to be further mitigated. Preliminary experiments were done with work with Spleeter, proposed by Hennequin et al. (2020), that contains pre-trained models for vocals/accompaniment separation, four stems separation (vocals, bass, drums and other) and five stems separation with an extra piano stem (vocals, bass, drums, piano and other). Another system tested was InaSpeechSegmenter (Doukhan et al., 2018). It is a CNN-based audio segmentation toolkit. It splits audio signals into homogeneous zones of music,speech and noise. Both system were tested as a way of circumventing the music noise removal problem, but yielded poor results. These methods were not adapted to our particular domain (western classical music), which motivates further work in source separation tools as a possible solution to the increase in quality of the musical audio played.

As it was previously mentioned, music fragments with no corresponding labels associated are discarded in our work. This aspect requires further attention, as to find a manner that facilitates the inclusion of all music played throughout the movie.

Throughout this work, one of the ways to capture melody information from the musical audio was via network encoding. There has been much work done in this area, although specifically for our task, we were limited in terms of the dataset available. It can be seen as one of the limitations of this work, the methodology used to capture sequence information. The results using the autoencoder did not distance themselves from the other setups using different features. We can argue that the architecture used was to general for the problem at hand and that further work should be done trying to better capture the melodic aspects of classical music. The work of Zalkow and Müller (2020) was considered for this purpose but unfortunately, the dataset (composed of a vast number of scores of western classical music), is not available, not allowing us to reproduce the results and use their setup and architecture to encode sequence information.

By improving the way we encode melody information at the audio segment level, we can come closer to capture leitmotifs present, thus improving how situations across the movie relate based on the music played.

References

Adams, D. (2010). *The Music of the Lord of the Rings Films: A Comprehensive Account of Howard Shore's Scores.* Alfred Music Van Nuys, CA.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406.

Chai, W. and Vercoe, B. (2005). Detection of key change in classical piano music. In *ISMIR*, pages 468–473.

Chollet, F. (2016). Information-theoretical label embeddings for large-scale image classification. *ArXiv*, abs/1607.05691.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44.

Gharghabi, S., Yeh, C.-C. M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., Kaplan, A., Crouter, S. E., and Keogh, E. (2019). Domain agnostic online semantic segmentation for multi-dimensional time series. *Data mining and knowledge discovery*, 33(1):96–130.

Gjoreski, H. and Roggen, D. (2017). Unsupervised online activity discovery using temporal behaviour assumption. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 42–49. ACM.

Gorbman, C. (1987). *Unheard melodies: Narrative film music*. Indiana University Press.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1909(136):210–271.

Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154. Deezer Research.

Krause, M., Zalkow, F., Zalkow, J., Weiß, C., and Müller, M. (2020). Classifying leitmotifs in recordings of operas by richard wagner.

Kristan, M., Leonardis, A., and Skočaj, D. (2011). Multivariate online kernel density estimation with gaussian kernels. *Pattern Recognition*, 44(10-11):2630–2642.

Lopes, A. L. V. d. S. (2017). Natural language generation for open domain human-robot interaction.

Masud, M., Gao, J., Khan, L., Han, J., and Thuraisingham, B. M. (2010). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):859–874.

Müller, M., Goto, M., and Schedl, M. (2012). *Multimodal Music Processing*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Germany.

Nakamura, T., Ando, Y., Nagai, T., and Kaneko, M. (2015). Concept formation by robots using an infinite mixture of models. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4593–4599. IEEE.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

Nishihara, J., Nakamura, T., and Nagai, T. (2016). Online algorithm for robots to learn object concepts and language model. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3):255–268.

Parker, B. S. and Khan, L. (2015). Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In *Twenty-ninth AAAI conference on artificial intelligence*.

Raczyński, S. A. and Vincent, E. (2014). Genre-based music language modeling with latent hierarchical pitman-yor process allocation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):672–681.

Rone, V. (2018). Scoring the familiar and unfamiliar in howard shore's the lord of the rings. *Music and the Moving Image*, 11(2):37–66.

Rosado, L. C. C. (2016). Cinema at the service of natural language processing.

Rudd, E. M., Jain, L. P., Scheirer, W. J., and Boult, T. E. (2017). The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):762–768.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Salton, G. and McGill, M. J. (1986). Introduction to modern information retrieval.

Serra, J., Müller, M., Grosche, P., and Arcos, J. L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5):1229–1240.

Takeda, R., Komatani, K., and Rudnicky, A. I. (2018). Word segmentation from phoneme sequences based on pitman-yor semi-markov model exploiting subword information. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 763–770. IEEE.

Wang, Z., Kong, Z., Changra, S., Tao, H., and Khan, L. (2019). Robust high dimensional stream classification with novel class detection. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1418–1429. IEEE.

Young, M. D. (2007). *Projecting Tolkien's Musical Worlds: A Study of Musical Affect in Howard Shore's Soundtrack to Lord of the Rings*. PhD thesis, Bowling Green State University.

Zalkow, F. and Müller, M. (2020). Learning low-dimensional embeddings of audio shingles for cross-version retrieval of classical music. *Applied Sciences*, 10(1):19.