CLASSIFICATION OF CHORUS AND VERSE: PROJECT REPORT

ARTIFICIAL AND NATURAL MUSIC COGNITION (SOW-MKI55)

Thijs van den Hout*
s4597400
t.vandenhout@student.ru.nl

March 30, 2019

ABSTRACT

What distinguishes a chorus from a verse in songs may seem an elementary question to some people and a complex one to others. This project proposes a classifier which can discern between a chorus and a verse given raw audio segments and presents the features that prove important in this distinction. This information is useful in the automatic segmentation of songs into sections: a popular task in the field of music information retrieval. Furthermore, an experiment was conducted which revealed the human accuracy in the aforementioned task, as a baseline, and the particularities the subjects attended to during the task.

***K*eywords** Music cognition · Classification · Musical Features

1 Introduction

Music is increasingly important and accessible in present-day life. With the expanding collection of music, it is vital to organize it effectively in musical databases. Especially in the field of music research, it is useful to have songs structurally segmented, i.e. key structural sections in music identified and annotated. To this extent, a large portion of music research in the conference for music information retrieval MIREX is dedicated to automatic segmentation of music [1]. Recent advances in structural segmentation research show improvement in the automatic segmentation of music [2, 3, 4]. Conversely, relatively few of these advances regard the classification of segments into musical structures. A. Eronen published a paper in 2007 in which they segment the chorus from songs, which requires the ability to find repetition of segments. [5]. They propose a number of features that will be used in the current research. This project will focus on the classification of already segmented pieces of music. In particular, the distinction between chorus and verse will be made. With the ability to classify chorus and verse from musical segments, music databases will be enriched with this information, which may spark more ideas in the field of music research. The research question this project hopes to answer is the following.
To what extent is it possible to automatically make a distinction between a chorus and a verse and what characteristics enable this distinction?
Our hypothesis is that we can create a classifier that is able to distinguish chorus form verse with human-level accuracy.

*Member of project group 2, comprising Thijs van den Hout, Karen Beckers, Filip Slijkhuis and Bart van Tiel

2 Methods

In this section, we will discuss the data set that was used, how it was created and the methods for feature extraction and classification of this data.

2.1 Data set

For this task, a set of musical segments, annotated with corresponding label, is required. Such a data set did not exist prior to this project. However, textual annotations of a number of songs were available. The annotations were obtained from isophonics.net, a website hosting music research tools including data sets.[2] Annotations include songs by Michael Jackson, The Beatles, Carole King and Queen. In the light of this project, we filtered out all annotations that were not either chorus, verse or derivatives / synonyms of the two. Songs that did not contain both a chorus and a verse were also discarded. The corresponding songs were downloaded and segmented given these annotations. The fact that slightly different versions of the same songs may be used in the creation of annotations (e.g. longer or shorted leading silence, remastered versions, remixes, etc.), the extracted segments in some cases did not exactly match their label. In rare cases the annotation data was wrong altogether. These segments were removed whenever encountered upon selective examination. The resulting data set consisted of 170 pairs of chorus and verse, i.e. a segmented chorus and verse from 170 songs. The data set, as well as all the code used in the project and pre-extracted feature vectors are made publicly available via the GitHub page of the author.[3]

2.2 Feature Extraction

Classification tasks require features that describe the data well, yet are not too verbose. In the audio domain, a number of well known features exist that are found to be descriptive for many audio-related tasks. This section will describe the features that are used in the current project, and how they were extracted from the audio segments. For most of these features, the average value over time is taken. Since all segments differ in duration, the feature vectors would all have different shapes otherwise.

MFCC MFCCs, or Mel-Frequency Cepstral Coefficients, are the coefficients that make up the Mel-Frequency Cepstrum. This is a representation of the short-term power spectrum of all pitches on the mel-scale; a scale which transposes pitch to equidistant intervals approximate to to human perception. This non-linear scale therefore resembles the response of the human auditory system better than the frequencies on a linear scale.

Spectrogram The extensively used spectrogram is a representation of the intensity of frequencies over time. We take the mean value over time for each segment as features, since the matrix becomes very large quickly with increasing duration of audio. The frequencies are divided into 128 bins, resulting in 128 features.

Chromagram A chromagram, or pitch class profile, represents the intensity of the 12 chromas (i.e. C, C#, D, etc.). A pitch class is defined as all pitches that share the same chroma, regardless of their octave. The chromagram is also averaged over time and therefore results in 12 features, corresponding to the intensity of the 12 pitch classes in the segment. These values are found to be timbre invariant. [4]

Spectral centroid, contrast, flatness and roll-off These three features are derived from the frequency spectrum. The centroid is the frequency bin around which the highest amount of spectral energy is concentrated and is found to encode the perception of brightness in a musical piece [6]. The spectral contrast feature was first introduced by Jiang et al. in 2002 and represents the relative spectral distribution, in contrast to the average spectral envelope computed in for example the MFCs [7]. Spectral roll-off frequency encodes the spectrogram bin for which a set percentage of the energy in the spectrum (default 85%) is concentrated in and below this bin. It therefore describes the highest frequency bin such that less than 15% of energy is found in higher frequency bins.

Polynomial coefficients The coefficients of an n-th order polynomial function that is fit on the columns of the spectrogram. In this project we experimentally set n to 3, resulting in 4 features. These coefficients can be seen as distillation of the spectrogram into 4 values.

[2] http://isophonics.net/datasets

[3] https://github.com/tvdhout/ANMCproject

Tempo and length The final two features that were included in the classification model are the estimated tempo of the segment, and the length of the segment.

These features were extracted in Python using the Librosa library.[4] For the first experiment, classification of the order of chorus and verse in a combined feature representation, the feature vectors of the chorus and verse of the same song were combined to form one feature vector per song. The feature vector of one segment was subtracted from the other in random order, resulting in one feature vector per song. The resulting data set comprised the chorus features subtracted from the verse features, or vice versa, with their appropriate label. Appending the two feature vectors together obtained similar results but renders feature selection illogical, which is why the former format was used. The second classification task, where there was no reference to the song's other segment, naturally resulted in twice the number of data points. Chorus and verse segments were shuffled and labelled accordingly.

Finally, the 60% of features with the highest ANOVA F-value (variation between samples) were selected to promote conciseness and discard likely undescriptive features. This resulted in a feature vector of 168 features.

2.3 Classification

The classification task was carried out with a Random Forest Classifier, comprising 300 decision trees with a maximum depth of 80, provided by the Scikit-learn Python library.[5] These hyper-parameters were experimentally tuned. A classifier was trained on the combined feature vectors, containing both chorus and verse features. Another classifier was trained on the individual segments of either chorus or verse.

2.4 Human experiment

To form an idea of human accuracy in the classification of chorus and verse, as well as create a baseline for the classifier, we conducted two experiments on human subjects. The subjects in experiment 1 were mostly university students. Participants in experiment 2 were mostly adolescents (age 18-28, with a few outliers) with no particular education. Both groups were predominantly not musically trained and contained approximately equal numbers of male and female subjects.

Experiment 1 18 participants were presented with 11 audio fragments, each consisting of a chorus and a verse from the same song, in random order. The songs were picked to be large unknown to the participants, to exclude any prior knowledge of the song's structure. The participants were asked to indicate whether they thought the chorus segment preceded the verse of vice versa. Finally, they were asked to point out the features they paid attention to most when performing the task.

Experiment 2 Similar to experiment 1, 28 participants were presented with 11 audio segments from the same songs as in experiment 1. This time the segment was either a chorus or a verse, without a reference to the other. Subjects were asked to indicate whether they thought the segment was a chorus or a verse, and which features they paid most attention to. The participants were different from the participants in experiment 1 because the same songs were used, which would introduce a bias for these people.

3 Results

This section comprises the results of the experiments, the trained classifiers and the feature importance analysis. Note that results may differ between executions, since Random Forest Classifiers are not deterministic. This means the accuracy and feature importances may differ slightly between runs.

3.1 Human experiment

Both experiments contained one fragment that was known to the majority of the subjects. This fragment was not included in the assessment to avoid knowledge bias. In experiment 1, the participants achieved an average accuracy of 84%. This accuracy is broken down in the confusion matrix in figure 1. Experiment 2 proved more difficult. Human subjects scored an average accuracy of 71%, which is dissected in the confusion plot in figure 2. Humans seem to be just as good in detecting a chorus as a verse. As previously mentioned, subjects also indicated the particularities they listened for when deciding between chorus and verse. Often subjective characteristics such as "catchiness" and

[4] http://librosa.github.io

[5] https://scikit-learn.org

"story telling" were noted. Many also indicated the chorus to have more energy, be more up-tempo and have higher perceived volume. Finally, many participants listened to the lyrics for a clue. Repetition of words, title-like phrases and memorable lyrics were particularly strong indicators of choruses.

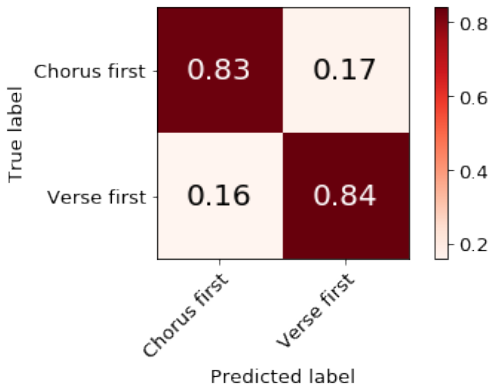

Figure 1: Confusion matrix of the human performance on experiment 1 (N=18)

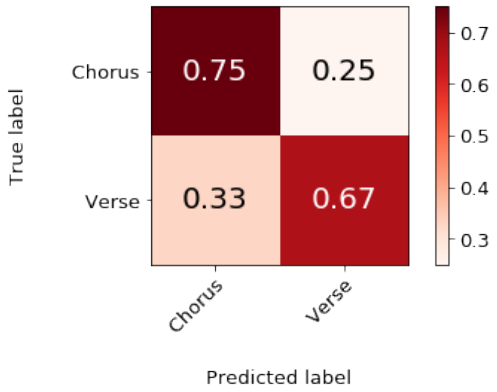

Figure 2: Confusion matrix of the human performance on experiment 2 (N=28)

3.2 Classification accuracy

The first classifier was trained on the feature vectors containing both the chorus and verse (with reference). Its task was to determine the order of chorus and verse, analogous to experiment 1. This classifier performed with 78% accuracy ($p<0.0001$): somewhat worse than the human baseline. The second classifier was trained on the individual segments (without reference) and performed surprisingly well, with 90% accuracy ($p<0.0001$). Both results are broken down in figures 3 and 4 respectively. Furthermore, training the second classifier on half the number of data points, the same number as was used to train the first classifier, resulted consistently in accuracies equal to the first classifier. The classifiers were evaluated using leave-two-out cross validation: training the classifiers on all but two data points and predicting these two with the classifier. This is repeated with different splits until every data point is evaluated, after which the average of these evaluations is taken.

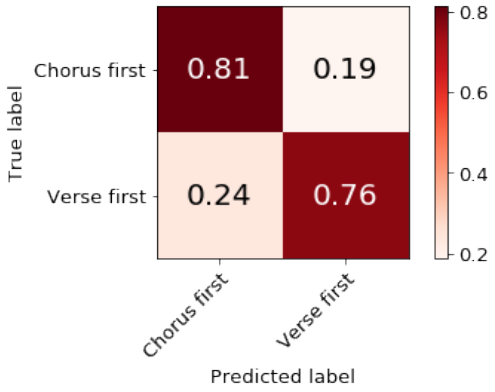

Figure 3: Confusion matrix of the classifier's performance on the classification of chorus and verse *with* reference

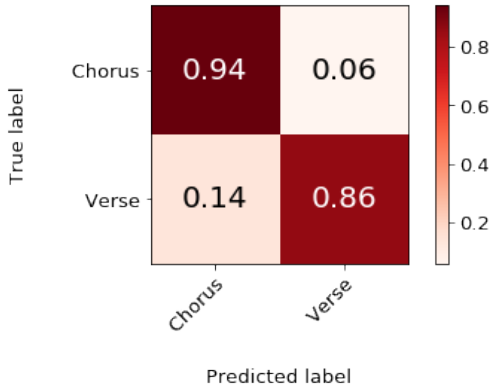

Figure 4: Confusion matrix of the classifier's performance on the classification of chorus and verse *without* reference

After inspecting which segments were often misclassified, a somewhat stable pattern was found which was similar in both tasks. Choruses were often misclassified when they were quiet, monotone or similar to the verse. Segments with bad audio quality or that were wrongly aligned and segmented due to erroneous annotations were also misclassified more often.

3.3 Feature importance

In figure 5, the ten most important features are plotted with their contribution. MFCC and spectrogram features seem to dominate the plot. MFCC features are difficult to interpret but can be thought of as describing timbre. The spectrogram bins are more interpretable, as they correspond to different frequencies. In the plot, mostly lower to medium frequencies are observed, implying the intensity of lower and medium frequencies are more important in the distinction between chorus and verse. One of the polynomial coefficients is also important in this distinction.

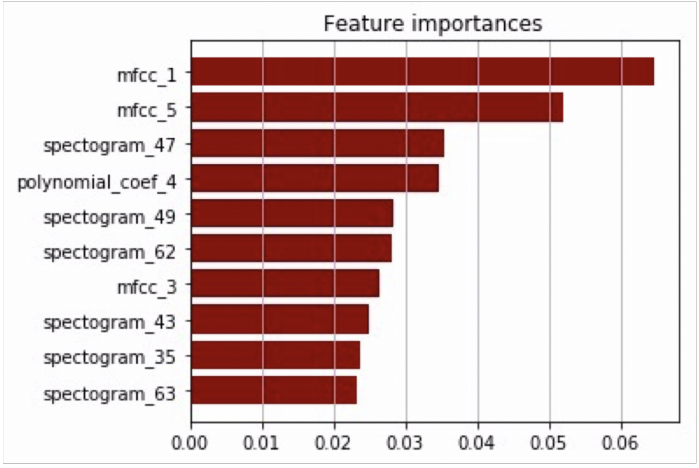

Figure 5: Feature importance plot. Larger values mean greater contribution to the decision.

4 Conclusion

From the presented results we can conclude a number of things. humans are better at determining the order of chorus and verse when presented with both segments than classifying either one without a reference. The Random Forest Classifier performed better on the latter task, due to the increased number of data points. The classifier without reference performed equally well as the classifier with reference when trained on the same number of data points. This indicates the classifier does not significantly improve when presented with a reference to both segments, unlike humans. Our hypothesis can be accepted with regard to experiment 2 (no reference), and can likely be accepted for experiment 1 with a larger data set.

5 Discussion

In this section, we will reflect on the project in terms of its progress, findings, implications and future research.

Qualitative data in music research is difficult to obtain. A large portion of this project's schedule was dedicated to creating a suitable data set for the problem. The resulting data set conformed to our expectations, but could be extended and rid of errors for future research. Moreover, the scope of the used data set was rather narrow, consisting only of (relatively) old pop and rock songs. In the future, it would be interesting to apply the same methods on a data set including different genres and more modern music.

Many participants in the two experiments stated they paid attention to lyrical features such as repetitiveness, memorability and snappy sentences. Future research may improve on our results by including textual features in the feature vectors. In particular, repetition, part-of-speech tags and sentence lengths should be observed. This would require extending the data set to include this information.

Future research may also look into integrating our findings in adjacent fields like structural segmentation to expand segments with more or more accurate information.

References

[1] Mirex 2019: structural segmentation. https://www.music-ir.org/mirex/wiki/2019:Structural_Segmentation. Accessed: 2019-03-28.

[2] Jordan BL Smith and Elaine Chew. A meta-analysis of the mirex structure segmentation task. In *Proc. of the 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil*, volume 16, pages 45–47, 2013.

[3] Jonathan T Foote and Matthew L Cooper. Media segmentation using self-similarity decomposition. In *Storage and Retrieval for Media Databases 2003*, volume 5021, pages 167–176. International Society for Optics and Photonics, 2003.

[4] Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Guenter Rudolph. *Music data analysis: Foundations and applications*. Chapman and Hall/CRC, 2016.

[5] Antti Eronen and F Tampere. Chorus detection with combined use of mfcc and chroma features and image processing filters. In *Proc. of 10th International Conference on Digital Audio Effects*, pages 229–236. Citeseer, 2007.

[6] John M Grey and John W Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.

[7] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116. IEEE, 2002.