# FEATURE ANALYSIS OF REPEATED PATTERNS IN DUTCH FOLK SONGS USING PRINCIPAL COMPONENT ANALYSIS

Iris Yuping Ren[1], Hendrik Vincent Koops[1], Dimitrios Bountouridis[1], Anja Volk[1], Wouter Swierstra[1], and Remco C. Veltkamp[1]

{*y.ren, h.v.koops, d.bountouridis, a.volk, w.s.swierstra, r.c.veltkamp*}*@uu.nl*
[1]Utrecht University

## 1. INTRODUCTION

Oral transmission plays a significant role in folk music. Through this often imperfect communication process, certain parts of melodies remain stable, variations are created, repeated patterns emerge [1]. Formulated in ethnomusicological studies, the concept of tune family describes the structures in this stream of transformations: folk songs that are supposed to have a common ancestor in the process of oral transmission are grouped into a tune family. Local structures within the melodies, namely characteristic motifs, or prominent, nonliterally repeated patterns, are detected to be useful in determining music similarity and classifying tune families [2]. Subsequently, in an annotation study on the influence of different musical dimensions on human similarity judgements of melodies belonging to the same tune family, repeated patterns between melodies turned out to play the most important role for similarity among all considered musical dimensions [3]. Therefore, algorithms which can extract these repeated patterns automatically would be useful for tune family classification.

Different pattern discovery methods have been introduced, such as sequence-based approaches [4, 5, 6, 7], geometric approaches [8]. Unfortunately, patterns extracted by state-of-the-art algorithms are not yet capable of replacing human annotations when we attempt to apply the patterns to classification and discovery tasks [9, 10].

This paper uses Principal Component Analysis (PCA) to better understand characteristics of musical patterns and to further use this information for designing and evaluating future pattern discovery algorithms. We show what features can summerise the data variance in musical patterns and propose using feature selection and extraction methods to improve pattern discovery algorithms.

There exists research that uses patterns in analysing tune families, modelling similarity, improving compression and retrieval tasks [10]. In this setting, it is common to either take the features of the whole song or the raw data of pitch and duration pairs of the patterns. We do not know of existing studies that focus on investigating the features only within patterns in music.

## 2. DATA AND SETUP

**Dataset** The corpus data we use is the Dutch folk song dataset MTC-ANN [11]. Three experts have annotated the prominent patterns in each song which could best position the song into one of the 26 tune families. The dataset consists of 360 Dutch folk songs with 1657 annotated patterns.

**Feature calculation** We calculate features from the patterns by using a common feature extraction tool: the jSymbolic2 toolbox in the jMIR toolset [12]. jSymbolic2 takes MIDI files as input and computes 155 musically meaningful features in six categories: texture, rhythm, dynamics, pitch, melody and chords.

**Feature selection** We perform a feature selection step and retain 64 features as follows: (1) Eliminating the features which are constant across all patterns; (2) Eliminating the features which are irrelevant to the music content of time and pitch, such as the dynamics features and artefacts introduced by MIDI conversion.

**PCA** After feature selection, we further combine and transform features to make new combined features, which is known as the feature extraction step. PCA is a well-known feature extraction and dimension reduction method. PCA gives new combinations of features which form orthogonal principal components. The principal components are in the same directions as the directions of the largest variances of the dataset. By examining the resulting principal components, we gain insights as to which features are of more significance in explaining the spread of the data points. PCA has been employed and shown to be effective in many MIR tasks [13]. We take a similar approach in the PCA analysis as [13] in which the author investigated audio features in popular music.

## 3. RESULTS

In Table 1, we report the prominent features and the weights in the first three PCA components. We make the following observations: (1) The most significant feature of the first component is *the number of strong rhythmic pulses*. Since rhythmic pulses are derived from beat histogram, it shows the importance of metric structures in the patterns. [1] More specifically, although there are both pitch and rhythmic features in the first principal component, we have three rhythmic features and two pitch features. In the second component, although there are more pitch features, the repeated notes feature is relevant both to pitch and duration. In the third component, we only have rhythmic features.

Furthermore, to give a fuller picture than the first five features in each component, we calculate the total weight sums of rhythmic and pitch related features. In the first component, the pitch related features have a total weight sum of 48.89% and the rhythmic features have a total weight

[1] For the details of other features, please refer to [12]. Given that we have 40 pitch related features, 20 rhythmic features and 4 features related to both pitch and duration, it is non-trivial that we have rhythmic features top-ranked in the first three principal components.

| PC (Percentage of variance explained) | Features | Weight (Percentage) |
|---|---|---|
| PC1 (22.51) | Number of Strong Rhythmic Pulses | 5.18 |
| | Pitch Variety | 5.15 |
| | Number of Relatively Strong Rhythmic Pulses | 5.07 |
| | Number of Common Pitches | 5.07 |
| | Number of Moderate Rhythmic Pulses | 5.07 |
| | Other Features | 74.46 |
| PC2 (12.42) | Repeated Notes | 8.24 |
| | Relative Prevalence of Top Pitches | 8.06 |
| | Relative Prevalence of Top Pitch Classes | 7.58 |
| | Prevalence of Most Common Pitch | 6.32 |
| | Prevalence of Most Common Pitch Class | 5.98 |
| | Other Features | 63.82 |
| PC3 (8) | Combined Strength of Two Strongest Rhythmic Pulses | 10.58 |
| | Polyrhythms | 9.98 |
| | Rhythmic Variability | 9.27 |
| | Strongest Rhythmic Pulse | 7.26 |
| | Strength of Strongest Rhythmic Pulse | 7.14 |
| | Other Features | 55.77 |

**Table 1**: The first three principal components of PCA and the weights of features. We omit the rest of $64 - 3 = 61$ components since they do not contribute significantly ($< 7.5\%$) to the variance and, for visualisation purposes, it is common practice that only the first three dimensions of PCA are considered.

sum of 46.38%. In the second component, pitch and rhythmic features have 64.45% and 27.89% weight sums respectively. The weight sums are 25.2% and 68.0% for the third component. In summary, looking at the first three dimensions of PCA, we see a balanced contribution from both the pitch and rhythmic features.

## 4. CONCLUSION AND DISCUSSIONS

Using PCA, we show the prominent features of MTC-ANN patterns. The pitch related and rhythmic features contribute together to the first PCA component; the second and third component is consist mainly of pitch related features and rhythmic features respectively. Despite the fact that we have less rhythmic features computed using the jSymbolic2 toolbox, the rhythmic features do not contribute less in the first three principal components. One might argue it is obvious that both pitch and rhythmic features are important, but it is remarkable that the two together contribute to each of the first few PCA dimensions.

The prominent features also give hints on potential improvements to current existing pattern discovery algorithms. Although metric structures have been considered in musical pattern research [14, 15, 16], many pattern discovery algorithms do not explicitly consider metric structures imposed by musical punctuations such as bar lines and measures. According to what PCA shows, in designing and evaluating pattern discovery algorithms, we should take metric structures into consideration as well as the repetitions and pitch related features in the patterns.

This investigation is a starting point for future work on using extracted pattern features for pattern classification and discovery. More concretely, we can further use the features to cluster and classify the patterns into tune families; using other metadata in the annotations, we can also correlate the features to the descriptions of annotators and motif classes; the features after PCA transformation can be used to explore, evaluate and compare algorithmically extracted patterns with human annotations.

## 5. REFERENCES


**[1]** Berit Janssen. *Retained or Lost in Transmission?* PhD thesis, University of Amsterdam, 2018.

**[2]** James R. Cowdery. A fresh look at the concept of tune family. *Ethnomusicology*, 28(3):495–504, 1984.

**[3]** Anja Volk and Peter Van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *MusicæScientiæ*, 16(3):317–339, 2012.

**[4]** Olivier Lartillot. Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal of New Music Research*, 34(4):375–393, 2005.

**[5]** Darrell Conklin. Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554, 2010.

**[6]** Iris Yuping Ren. Closed patterns in folk music and other genres. *Proceedings of the 6th International Workshop on Folk Music Analysis*, pages 56–58, 2016.

**[7]** Matevž Pesek, Aleš Leonardis, and Matija Marolt. Symchman unsupervised approach for pattern discovery in symbolic music with a compositional hierarchical model. *Applied Sciences*, 7(11):1135, 2017.

**[8]** David Meredith, Kjell Lemström, and Geraint A. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

**[9]** Iris Yuping Ren, Hendrik Vincent Koops, Anja Volk, and Wouter Swierstra. In search of the consensus among musical pattern discovery algorithms. *Proceedings of the International Society for Music Information Retrieval*, pages 671–680, 2017.

**[10]** Peter Boot, Anja Volk, and W. Bas de Haas. Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research*, 45(3):223–238, 2016.

**[11]** Peter van Kranenburg, Berit Janssen, and Anja Volk. The Meertens Tune Collections: The Annotated Corpus (MTC-ANN) versions 1.1 and 2.0.1. *Meertens Online Reports*, 2016(1), 2016.

**[12]** Cory McKay. *Automatic Music Classification with jMIR*. PhD thesis, McGill University, 2010.

**[13]** Jan Van Balen. *Audio description and corpus analysis of popular music*. PhD thesis, University Utrecht, 2016.

**[14]** Peter Van Kranenburg and Darrell Conklin. A pattern mining approach to study a collection of dutch folk-songs. *Proceedings of the 6th International Workshop on Folk Music Analysis*, pages 71–73, 2016.

**[15]** Darrell Conklin and Christina Anagnostopoulou. Representation and discovery of multiple viewpoint patterns. In *Proceedings of the 26th International Computer Music Conference*, pages 1–7. Citeseer, 2001.

**[16]** Darrell Conklin and Mathieu Bergeron. Feature set patterns in music. *Computer Music Journal*, 32(1):60–70, 2008.