MODELING SONG SIMILARITY WITH UNSUPERVISED LEARNING

Matevž Pesek, Manca Žerovnik, Aleš Leonardis, Matija Marolt
University of Ljubljana, Faculty of Computer and Information Science
{firstname.lastname}@fri.uni-lj.si

ABSTRACT

The SymCHM, which was developed for the pattern discovery, was applied to the music similarity task. It was used as a feature extractor, unsupervisedly learning repeated patterns on pitch-time representation, eliminating any additional high-level information. The output was used in the retrieval, where the model achieved 74.4 % classification accuracy on the Dutch folk dataset. By the unsupervised aspect of model's training and the ability to perform similarity using only the most basic song representation, we find the results sufficient to further explore the use of the model on datasets with a low number of additional features and basic music representations.

1. INTRODUCTION

The concept of similarity in music has been studied in different research areas. The similarity in cognition plays a significant role in psychological accounts of problem solving, memory, prediction, and categorization (Holyoak & Morrison, 2005). Many research topics in musicology are inherently related to similarity and categorization in music, e.g. the study of motivic-thematic relations, comparison of musical motifs, categorization of songs into tune families and many others (Volk & Van Kranenburg, 2012). Understanding of the basic processes underlying perception of musical similarity is necessary for acquiring a deeper comprehension of music perception in general (Toiviainen, 2007).

The categorization of folk songs into tune families, where a tune family represents "a set of folk songs which have a common origin in history" (Bayard, 1950), also relies on the similarity. Many current approaches for this task rely on alignment algorithms. Mongeau & Sankoff (1990) were one of the first to use alignment algorithms for music, establishing the basis for several future approaches, employing the alignment algorithms and profile modeling for classification and retrieval tasks, e.g. using the pop and rock songs datasets Bountouridis & Van Balen (2014). Walshaw (2017) investigated enhancements of the well-established local alignment algorithms to also classify Dutch folk songs into tune families.

Bountouridis et al. (2017) used biologically-inspired techniques for MIR tasks. They identified several shared concepts between music and bioinformatics, such as melody (DNA), oral transmission (evolution), variations (homologues), tune families (homology) etc. and showed that bioinformatics algorithms are suitable for MIR tasks. (Savage & Atkinson, 2015) also used an adapted alignment algorithm from the field of bioinformatics to classify songs into four diverse tune families (two English, two Japanese).

Several developed approaches were evaluated on the Dutch folk song dataset compiled by Van Kranenburg et al. (2013). Among the most recent, the alignment approach by Van Kranenburg et al. (2016) produces the best classification accuracy on the Dutch folk song dataset. The approach models various features of music as substitution scoring functions, which are incorporated into the Needleman--Wunsch-Gotoh Gotoh (1982) algorithm. The model employs several 'viewpoints', such as pitch, duration, score time, time in bar, onset, current bar number, current phrase number, upbeat, current meter, free meter, accented, inter-onset-interval ratio, normalized metrical weight and the time position within phrase. Van Kranenburg had analyzed combinations of these attributes and had discovered the best results were given by using the pitch and position within phrase attributes. Despite the high accuracy, it requires a considerable amount of time to produce such attributes for each dataset. Consequently, the majority of these attributes are usually not available in music collections. To eliminate the need for expert knowledge, Velarde et al. (2013) classified Dutch songs using Haar-wavelet filters. The results are not on par with Van Kranenburg et al. (2013), but the approach does not require any encoded expert knowledge.

In this paper, we explore how unsupervised learning can be used for modeling tune similarities and classification into tune families. Specifically, we study the compositional hierarchical model that has been previously applied to several audio-based tasks, such as chord estimation and polyphonic transcription (Pesek et al., 2017a) and pattern discovery (Pesek et al., 2017b), using a modified symbolic version of the model (SymCHM).

2. METHODOLOGY AND EXPERIMENT

The compositional hierarchical model has been previously applied to several tasks, including spectral-oriented tasks of multiple fundamental frequency estimation (Pesek et al., 2017a) and automated chord estimation (Pesek et al., 2014), and symbolic-oriented tasks of pattern discovery (Pesek et al., 2017b). A model able to perform on symbolic music representations, denoted SymCHM will be used to tackle the tune family classification problem in this paper.

2.1 Model

The idea behind SymCHM lies in the organizing of frequently co-occurring events into compositions. Starting at its input, the model observes the statistics of events' pres-

ence and the relation between them. For example, if two events frequently co-occur in a given time window on a specific interval, both events can be joined into a composition. The composition is relatively encoded, meaning, should two events co-occur at one pitch location and again at a different one, the same composition would be formed. This procedure is repeated on consecutive layers. In contrast to the first layer, instead of observing the input, the model observes co-occurrences of compositions and forms new relatively encoded compositions on the next layer, based on the previous layer. In the SymCHM, we name all compositions *parts*, similar to nodes in other models.

Since each part may occur at several locations in a single input, such occurrences must be defined by its location in time and pitch (a single part may occur at two different pitch heights at the same time). We denote such occurrences *activations* and define their position by their time, and pitch. The parts learned by the model can be observed as melodic patterns and their activations as pattern occurrences.

Once the model is built, it can be inferred over another (or the original input). The inference may be exact or approximate, where in the latter case biologically-inspired hallucination and inhibition mechanisms enable the model to find variants of part occurrences with deletions, changes or insertions, thus increasing its predictive power and robustness. The hallucination mechanism provides means to activate a part even when the input is incomplete or changed. In symbolic music representations, such changes often occur in melodic variations and ornamentation. The hallucination enables the model to robustly identify patterns with variations. The inhibition mechanism is also essential in the SymCHM for removal of redundant co-occurrences. As the model does not rely on any musicological rules, parts may produce a large number number of competing patterns. Inhibition may be used to reduce the number of activations and find the patterns that best correspond to the learned hierarchy.

The SymCHM therefore learns a hierarchical representation of patterns occurring in the input, where patterns encoded by parts on higher layers are compositions of patterns on lower layers. The inference produces part activations which expose the learned patterns (and their variations) in the input data. Shorter and more trivial patterns naturally occur more frequently, longer patterns less frequently. On the other hand, longer patterns may entirely subsume shorter patterns.

2.2 Experiment

We tackled the melody classification using the SymCHM. The MTC-ANN annotated dataset [1] was used. The dataset consists of 360 Dutch folk songs, accompanied by tune family annotations. Similar to the experiment presented in Van Kranenburg et al. (2013), we classified the folk tunes in to tune families using features. To gather the features, we employed the symbolic version of the compositional hierarchical model for pattern discovery for this task. The SymCHM, shown in Fig. 1, was presented by Pesek et al. (2017b) and was evaluated for the MIREX discovery of repeated patterns and sections task.

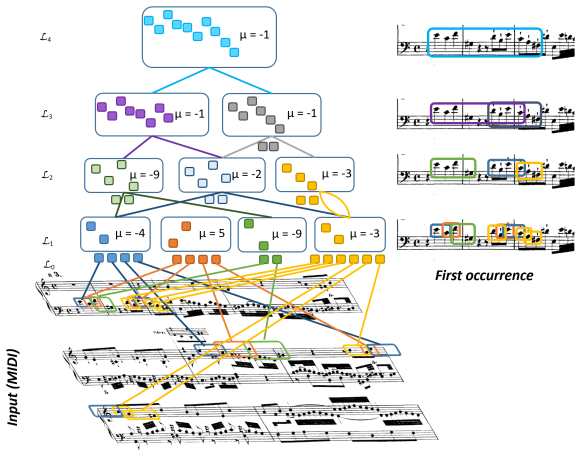

Figure 1: An abstraction of the SymHM's performance over a symbolic music representation. Each part has multiple activations (the number of activation is expressed under each part). Composing parts can partially overlap. While composing, the parts are joined into a composition by a relative offset, represented by μ. The activations for the first pattern for the selected example are shown on the right side per layer.

The SymCHM can be used for intra-opus task, such as the mentioned MIREX task, by building a model for each music piece individually. The model is first built and later inferred on a single symbolic representation. In the intra-opus task, the statistical drive behind the model's building procedure reflects the frequency of co-occurrences within a single music piece. It is therefore difficult to compare the learned patterns due to the difference in models' hierarchies, when comparing multiple music pieces. The music similarity task is, on the other hand, an inter-opus task. We have therefore built a single model on several songs. The compositions therefore still reflect the frequent proximity of events (i.e. pattern occurrences) within each single music piece, but is also regulated by the frequency of such occurrences across the given input dataset. The model accepts multiple inputs separately and analyzes them piece by piece. The statistical nature in the model is invariant to the length or the number of inputs—it calculates the co-occurrence of events within a single input and produces compositions of such events. If a similar co-occurrence of events occurs in another input, its statistic is added to the existing composition reflecting the occurrence.

Any symbolic music representation with the following two features is accepted as an input to the model: as a set of note onsets (e.g. in seconds) and note pitches (e.g. MIDI pitch). MIDI format may be used, extracting only these two attributes; all other attributes, which can be extracted from the MIDI format (e.g. meter, bar, phrase number

[1] Dataset accessible here: `http://www.liederenbank.nl/mtc/`

etc.), are discarded.

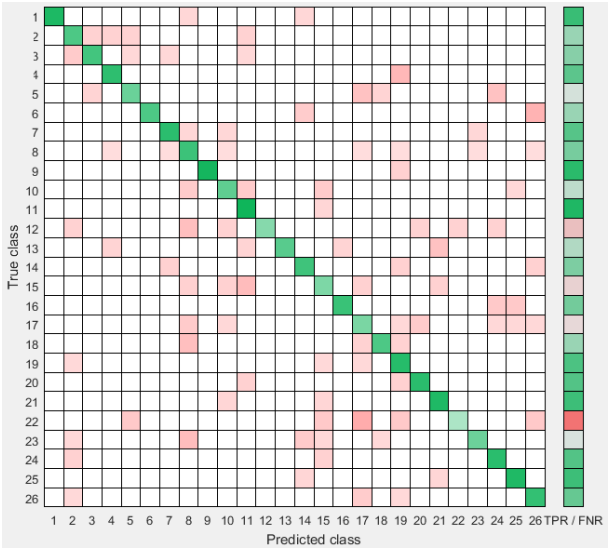

Figure 2: The confusion matrix of the tune family classification. The reference annotations are represented in rows (left) and the predicted classes in columns (bottom).

The model was built on the 360 songs of the MTC-ANN dataset. The songs are categorized into 26 tune families. No annotations were used during the training. The built model produced a list of discovered patterns across the input songs. The output was encoded into a feature vector where each part represented a vector element and its value represented the sum of activations for the represented part. The output was encoded into a feature vector where each part was mapped onto a vector element, and the value represented the sum of the part's activations, as described in the first experiment. The model generated 3750 parts across layers 3–7. The vector values were adjusted as described in (Van Kranenburg et al., 2013). For each element, the values were scaled to have zero mean and standard deviation of 1. As described by (Van Kranenburg et al., 2013), the cosine distance was used for comparison of vectors. The result of the vector comparison was a 74.4 % classification accuracy. The confusion matrix is depicted in Figure 2.

3. CONCLUSIONS

The results are about 20 percent lower when compared to Van Kranenburg et al. (2013). However, we believe the results are interesting, considering the fact that a pattern discovery model, relying only on onset-pitch notation, was used for this task. The model was not specifically trained or parameter-tuned for this task and was applied to the dataset without any dataset-specific adjustment. The model provided compositions of relatively-encoded melodic patterns learned in an unsupervised manner. In contrast to several approaches applied to this dataset, no know-how about the dataset or folk and western music in general was used in the procedure of patterns which were used for classification. Nevertheless, such incorporation could also be beneficial to the proposed model's results and will therefore be further explored in our future work.

4. REFERENCES

Bayard, S. P. (1950). Prolegomena to a study of the principal melodic families of british-american folk song. *The Journal of American Folklore*, *63*(247), 1–44.

Bountouridis, D., Brown, D. G., Wiering, F., & Veltkamp, R. C. (2017). Melodic similarity and applications using biologically-inspired techniques. *Applied Sciences*, *7*(12), 1242.

Bountouridis, D. & Van Balen, J. (2014). The cover song variation dataset.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, *162*(3), 705–708.

Holyoak, K. J. & Morrison, R. G. (2005). *The Cambridge handbook of thinking and reasoning*. Cambridge University Press.

Mongeau, M. & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*(3), 161–175.

Pesek, M., Leonardis, A., & Marolt, M. (2014). A compositional hierarchical model for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (pp. 131–136)., Taipei.

Pesek, M., Leonardis, A., & Marolt, M. (2017a). Robust Real-Time Music Transcription with a Compositional Hierarchical Model. *PLoS ONE*, *12*(1).

Pesek, M., Leonardis, A., & Marolt, M. (2017b). SymCHM—An Unsupervised Approach for Pattern Discovery in Symbolic Music with a Compositional Hierarchical Model. *Applied Sciences*, *7*(11), 1135.

Savage, P. E. & Atkinson, Q. D. (2015). Automatic tune family identification by musical sequence alignment. In *Proceedings of the 16th ISMIR Conference*, volume 163.

Toiviainen, P. (2007). Discussion Forum 4A - Editorial. *Musicæ Scientiæ*, *1*(1), 3–6.

Van Kranenburg, P., Janssen, B., & Volk, A. (2016). The meertens tune collections: The annotated corpus (mtc-ann) versions 1.1 and 2.0. 1.

Van Kranenburg, P., Volk, A., & Wiering, F. (2013). A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies. *Journal of New Music Research*, *42*(1), 1–18.

Velarde, G., Weyde, T., & Meredith, D. (2013). Wavelet-filtering of symbolic music representations for folk tune segmentation and classification. In *Proceedings of the Third International Workshop on Folk Music Analysis (FMA2013)*, (pp.5̃6).

Volk, A. & Van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, *16*(3), 317–339.

Walshaw, C. (2017). Tune classification using multilevel recursive local alignment algorithms.