SYMCHM: A COMPOSITIONAL HIERARCHICAL MODEL FOR PATTERN DISCOVERY IN SYMBOLIC MUSIC REPRESENTATIONS

Matevž Pesek, Urša Medvešek
University of Ljubljana
Faculty of computer
and information science
matevz.pesek@fri.uni-lj.si
ursa.medvesek@lgm.fri.uni-lj.si

Aleš Leonardis
Centre for Computational
Neuroscience and Cognitive Robotics
School of Computer Science
University of Birmingham
ales.leonardis@fri.uni-lj.si

Matija Marolt
University of Ljubljana
Faculty of computer
and information science
matija.marolt@fri.uni-lj.si

ABSTRACT

The compositional hierarchical model has been well explored for several tasks in the field of music information retrieval, including the automated chord estimation and multiple fundamental frequency estimation. This submission introduces the compositional approach to the symbolic representations of the music scores by using the model's structure which was applied to the audio domain. We adjusted the model for the task of pattern discovery in monophonic symbolic data.

1. INTRODUCTION

As an alternative to the existing deep learning architectures, a compositional hierarchical model (CHM) was introduced to the MIR field by Pesek et al. [3], based on the model in computer vision, developed by Leonardis and Fidler [2]. Its main difference between the model and other deep architectures is in its transparent structure, thus allowing representation and interpretation of the signal's information extracted on different levels.

This submission presents a novel application of the CHM for pattern discovery in symbolic music representations. While retaining the structure and the applied methods presented in [3], the model was adjusted for the two-dimensional symbolic input.

2. COMPOSITIONAL HIERARCHICAL MODEL

The compositional hierarchical model provides a hierarchical representation of the audio signal, from the signal components on the lowest level, up to individual musical events on the highest levels. The model is built on the belief of the signal's ability of hierarchical decomposition into atomic blocks, denoted as *parts*. According to their complexity, these parts can be structured across several layers from less to the more complex. Parts on higher layers are expressed as compositions of parts on lower layers — similarly as a chord is composed of several pitches, or a pitch represents a composition of several harmonics. A part can therefore describe individual frequencies in a signal, their combinations, as well as pitches, chords and temporal patterns, such as chord progressions.

The CHM was previously introduced to the MIR community [3]. The model with a three-layer structure was first used for the automated chord estimation task. The output was mapped to a chroma-like octave-invariant representation and used as an input to a Hidden Markov model. Due to the white-box approach, the model was later extended to the task of multiple fundamental frequency estimation using the same three-layer structure. The third layer part structures were observed as pitches or pitch-partial with activations indicating the location and probability of their occurrence. By using the same chroma-like approach described in [5], authors showed the same features can also be used for the mood estimation task similarly to the usage of MFCC and other features [3]. Furthermore, the authors showed the model's features outperform the chroma features in robustness to noise [4] for the automated chord estimation task. However, the evaluated tasks are all audio-input based. We adjusted the model to perform by using symbolic data as an input.

2.1 Input layer

The input layer $\mathcal{L}_0$ of the model is a symbolic representation of the music signal, consisting of a set of pitches, each defined by an onset and an offset. It contains a set of atomic parts (pitches), which are activated (is present in the signal) at any MIDI location and any given time. Similar to the original model, where any time-frequency representation can be used for the input layer, any two-dimensional representation — in this case pitch-time representation — can be used as an input to the adjusted version.

2.2 Subsequent layers

Higher layers $\mathcal{L}_n$ of the model contain sets of *compositions* - parts composed of parts from lower layers. Each composition contains two parts from the lower layers. A

composition can be part of any number of compositions on higher layers. A $\mathcal{L}_n$ part is a composition which consists of two structures $P_{n-1,j}$ and $P_{n-1,k}$ — parts on $\mathcal{L}_{n-1}$ layer. The relation is represented relatively as an offset in pitch, thus:

$$P_{n,i} = \{P_{n-1,j}, P_{n-1,k}, \mu_{n,i}\}. \quad (1)$$

The offset parameter $\mu_{n,i}$ is learnt for each composition from the input data during the learning stage of the model. Due to the discrete nature of the symbolic input, there is no need to model the offset by a Gaussian as it is done in the original CHM. Since the first subpart serves as the base for defining the relative offset between itself and the second subpart, we denote the first subpart ($P_{n-1,j}$) the central part.

The SymCHM is used in the same two-stage manner as the CHM. During the first, the build stage, the model is developed layer-by-layer. By composing atomic $\mathcal{L}_0$ parts, the model first produces compositions of two pitches ($\mathcal{L}_1$). To retain the compositions which cover the most information in the input layer, a statistical approach is employed. Based on the compositions' occurrence, the learning process retains the compositions which are more frequently activated.

2.3 Activations

A composition is *activated* (propagates output to higher layers) when all of its subparts are activated. For the $\mathcal{L}_0$, all input data is treated as a set of activations at given locations in terms of time offset and pitch. For consequent layers, there are two factors which limit the activations for a given part $P_{n,i}$:

- the pitch offset between the activations of $P_{n,i}$ subparts
- the time offset between the activations of $P_{n,i}$ subparts.

The pitch offset between the activations of subparts must be equal to the $\mu_{n,i}$, whereas the time offset is modelled by a threshold. The latter results in a window-like mechanism where two activations, each belonging to one of the $P_{n,i}$ subparts, are considered for a $P_{n,i}$ activation only if they occur inside the window. This limitation can be observed as a short-term memory-like mechanism. With each consequent layer, the time-offset threshold loosened with the factor of 2^n. Thus, the window grows exponentially with the granularity and complexity of the parts. Finally, the time-offset threshold also reduces the time complexity of the learning stage.

3. CHM FOR PATTERN DISCOVERY

Due to the statistical nature of the model's learning behaviour, more frequently activated parts are retained on each layer. The activations can be observed as locations of part's occurrences, thus, the amount of part's activations indicates the significance of the part's structure (i.e. a repeated pattern) in the signal. A part can thus be observed as a medium for aggregation of re-occurring patterns.

On the contrary to the spectral CHM, complex temporal patterns are not commonly shared between different musical pieces, with obvious exceptions of patterns shared between several musical pieces of the same artists, remakes or other influenced work. For this task, we therefore build a new model for each musical piece. After the learning process, the model's activations are produced through the output of the model where every part is observed as a pattern and each activation belonging to that part as a pattern occurrence.

4. RESULTS

The SymCHM was evaluated on the JKU PDD dataset for the symMono subtask. For each musical piece, the model was built independently and inferred with the same piece. We built a structure with six layers. The model's output, which was used during the evaluation, consists of activations of parts on layers $\mathcal{L}_4$, $\mathcal{L}_5$ and $\mathcal{L}_6$. We performed a simple pattern picking of all three layers. The latter is similar to the union (all three layers of patterns are merged into a single output), where the patterns which are subsets of other patterns are removed. For example, a 4th layer pattern is not included if the pattern represents a subpart of a 5th layer composition.

We built and inferred a separate model with each musical piece provided in JKU PDD dataset as an input to the model. We constrained the model's parameters to a handpicked combination. The parameters could be further optimized; however, the goal of this submission is to show the sheer ability to adapt the model to the pattern discovery task, thus we leave the parameter optimization for the future work.

Table 1. Inhibition an hallucination parameters of the learnt structure for the pattern discovery task. Though not optimised, the results show the adjustment of the model for the task is worthwhile. The parameters can be adjusted individually per layer.

Mechanism	layers 1 - 6
Inhibition	0.2
Hallucination	0.7

The results of the evaluation are shown in Table 2. For the evaluation and comparison of the output to the ground truth, we used the script, provided by Tom Collins [1]. The results display the commonly used metrics for the task.

5. REFERENCES

[1] Tom Collins. 2015:Discovery of Repeated Themes & Sections - MIREX Wiki, 2015.

Table 2. Result of merged CHM layers

Piece	P_{est}	R_{est}	F_{1est}	$P_{occ(c=.75)}$	$R_{occ(c=.75)}$	$F_{1occ(c=.75)}$	P_3	R_3	TLF
1	78.57	33.33	46.81	100.00	57.14	72.73	55.22	24.24	33.69
2	78.60	60.86	68.60	84.29	86.33	85.30	84.63	66.31	74.36
3	47.59	62.00	53.85	90.95	70.15	79.21	49.13	69.22	57.47
4	94.25	29.50	44.94	94.25	100.00	97.04	94.25	27.77	42.90
5	40.58	41.12	40.85	100.00	100.00	100.00	50.00	50.70	50.35
Average	67.92	45.36	51.01	93.90	82.72	86.85	66.64	47.65	51.75

Piece	$P_{occ(c=.5)}$	$R_{occ(c=.5)}$	$F_{1occ(c=.5)}$	P	R	F_1	$FFTP_{est}$	FFP
1	73.81	57.14	64.42	50.00	33.33	40.00	33.33	55.22
2	81.27	73.65	77.27	0.00	0.00	0.00	60.86	84.63
3	71.84	61.32	66.16	0.00	0.00	0.00	30.46	48.20
4	94.25	100.00	97.04	66.67	25.00	36.36	29.50	94.25
5	71.46	72.86	72.15	8.33	11.11	9.52	18.56	47.17
Average	78.53	72.99	75.41	25.00	13.89	17.18	34.54	65.89

[2] Aleš Leonardis and Sanja Fidler. Towards scalable representations of object categories: Learning a hierarchy of parts. *Computer Vision and Pattern Recognition, IEEE*, pages 1–8, 2007.

[3] Matevž Pesek, Aleš Leonardis, and Matija Marolt. A compositional hierarchical model for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 131–136, Taipei, 2014.

[4] Matevž Pesek, Aleš Leonardis, and Matija Marolt. A preliminary evaluation of robustness to noise using the compositional hierarchical model for music information retrieval. In *Proceedings of the Electrotechnical and Computer Science Conference (ERK)*, pages 104–107, Portorož, Slovenija, 2014.

[5] Matevž Pesek and France Mihelič. Hidden Markov model for chord estimation using compositional hierarchical model features. In *Zbornik dvaindvajsete mednarodne Elektrotehniške in računalniške konference*, pages 145–148, 2013.