THE COMPOSITIONAL HIERARCHICAL MODEL FOR MUSIC INFORMATION RETRIEVAL

Matevž Pesek
Univ. dipl. inž. rač. in inf.

Supervisors:
assoc. prof. dr. Matija Marolt
prof. dr. Aleš Leonardis
Parts of presentation

• Music information retrieval field (MIR)
  – Deep architectures in MIR

• Motivation for this research

• Compositional hierarchical model – structure
  – Transparent structure and mechanisms

• CHM for time-frequency representations
  – Chord estimation\(^1\), transcription\(^2\)

• CHM for symbolic representations
  – Pattern discovery\(^3\), tune family identification

• CHM for rhythm modeling

• Conclusion
Introduction

Music

• “The science or art of ordering tones or sounds in succession, in combination, and in temporal relationships to produce a composition having unity and continuity.” [www.meriam-webster.com]

• “There is no noise, only sound.” [John Cage - interview]

Several research fields

• Musicology [Lerdahl1983, McDermott2008] (rules)
• Psychology [Gelfand2004, Tirovolas2011] (perception and cognition)
• Neuroscience [Amitay2006, Peretz2003, Werner2012] (mechanisms)
• Computer Science - signal processing and music information retrieval (analysis, understanding, retrieval)
Music information retrieval

- **Interdisciplinary science** of retrieving information from music

- Relatively young field (1970’s / late 1990’s) [Orio2006]

- **Popular problems** [Downie2008, Downie2010]:
  - Extraction of high-level features:
    - Melody extraction [Ryynanen2008, Salamon2014]
    - Rhythm and beat tracking [Schmidt2013, Pikrakis2013, Bock2015]
    - Mood estimation [Laurier2009, Dixon2013]
  - Music creation [Huang2012, Dean2014]
  - Visualization [Lamere2009]
  - ...
Deep learning in MIR

• Modeling **high-level abstractions** in data by using **layered-architectures**
  – many based on neural-networks

• **Learning of features** for classification and detection

• Introduced to MIR around 2010
  – Genre recognition [Hamel2010]
  – Emotion-based feature extraction [Schmidt2011]
  – Rhythm genre discrimination [Pikrakis2013]
  – Drum pattern analysis [Battenberg2012]
  – Beat tracking [Krebs2013]
  – Onset detection [Schluter2013]
  – Multiple fundamental frequency estimation [Hawthorne2017]
The Compositional Hierarchical Model: Motivation
The Compositional Hierarchical Model

• An alternative deep architecture
  – Unsupervised learning of a hierarchy of parts
  – Transparency
    • Representations are explainable
  – Relativity
    • Representations are relatively encoded and reused
    • Smaller datasets needed for training
  – Compositionality
    • Parts composed of parts
    • Able to perform in discovery tasks

• Idea: complex signals can be decomposed into simpler parts
  – Parts possess various levels of granularity
  – Parts can be distributed across several layers from simple to complex
Origin of the Idea

• **Learned Hierarchy of Parts**
  - Introduced by Leonardis & Fidler for object categorization in images
  - Unsupervised learning of a hierarchy of parts
    - Small image segments on lower layers
    - Complex shapes on higher layers
    - Transparency

• **Music is hierarchical** in frequency and time
  - The nature of the model coincides well with this hierarchical structure

Source: Tabernik et al.
Our Goal

• Develop a **deep compositional model for music** processing
  – Focus on transparency, shareability and relativity of learned representations

• Develop a **general model** and test it for different tasks
  – Automated chord estimation
  – Multiple fundamental frequency estimation
  – Discovery of repeated themes and sections
  – Classification of melodies
  – Rhythm modeling
Part 2

The Compositional Hierarchical Model: Structure
Model Structure

- The model is **hierarchical** and built of layers of **parts** that **encode** the learned **concepts**
  - higher layers encode more complex concepts
- Each layer has a number of **parts**
  - parts are **compositions** of subparts
    - \( p_i^n = \left\{ p_{k_0}^{n-1}, \{ p_{k_j}^{n-1} (\mu_j, \sigma_j) \}^{j=1}_{j=1} \right\} \)
    - relations between subparts are **relative** with respect to the **central part**
- The **input** is a representation of a music signal
  - spectrogram, MIDI events, onsets ...
- The entire structure is **transparent**
Learning

• The model is built by **unsupervised learning** on a set of examples
  – Learning takes place layer-by-layer

• Learning is based on **statistical regularities** in input data
  – frequently co-occurring parts are joined into new compositions

• Learning optimizes **coverage** of the input signal vs. the number of parts

```
1: procedure SELECT(C)
2: prevCov ← 0
3: cov ← ∅
4: L_n ← ∅
5: sumInput ← |I|
6: repeat
7:   for P ∈ C do
8:     c ← 0
9:     F ← C(L_n ∪ P)
10:    c ← c + |F|
11:    cov[P] ← c/sumInput
12:  end for
13:  Chosen ← argmax(cov)_P
14:  L_n ← L_n ∪ Chosen
15:  C ← C \ Chosen
16:  if cov[Chosen] − prevCov < τ_C then
17:     break
18: end if
19:  prevCov ← cov[Chosen]
20: until prevCov > τ_P ∨ C = ∅
21: return L_n
```
Inference

• Inference calculates **activations** of parts on a given input signal
  
  \[ A = \langle A_T, A_L, A_M \rangle \]
  
  • time, location, magnitude
  
  • An activation represents the **location and form** of the learned **concept** in the input signal

• Parts on the **first layer** are activated from the corresponding **input**

• Compositions on **higher layers** are activated based on activations of their subparts:
  
  • activation time and location are propagated via central parts (indexing):
    
    \[ A_L(P^n_l) = A_L(P^{n-1}_{k_0}) \]
    \[ A_T(P^n_l) = A_T(P^{n-1}_{k_0}) \]

• Activations are **interpretable**
Inhibition

- Inhibition **reduces redundant** activations during inference
  - removes weak activations that cover the same parts of the signal as stronger ones
- **Good for**
  - Removal of redundant explanations
  - Noise filtering
  - Hypotheses refinement
Hallucination

- Hallucination activates parts in presence of **incomplete** input
- Provides the **most probable explanation** of input based on available information
- Good for:
  - Interpretation of missing information
  - Context-dependent perception
Part 3

The Compositional Hierarchical Model for Time-Frequency Representations
CHM: Time-Frequency Representations

- **Input:** audio data (e.g. CQT)
  - Time, frequency, magnitude

- **Compositions**
  - \( \mu, \sigma \) represent **frequency distances** (in bins)
  - Relatively encoded **harmonic structures** within each frame
  - Increased size over layers

- **Activations**
  - **Harmonic occurrences** in input

- **Aim**
  - Learn pitch-related compositions that occur within a piece or music corpus
Automated chord estimation

- **Goal:** *identify chords* in audio
  - CHM should produce parts that relatively encode *pitches, intervals and chords*
- **Unsupervised model training on different collections**
- **Lessons learned**
  - Harmonic structures are dominant, consequently on higher layers CHM does not produce many intervals/chords without modifications
  - CHM can efficiently model pitch

**Evaluation: CHM as feature generator**

- Learn two compositional layers
  - parts represent harmonic series
- Add an octave-invariant layer
  - *features* similar to chroma vectors
- For comparison to other approaches, use CHM’s output as input to a Hidden Markov model
- Evaluate on *The Beatles Dataset* (C. Harte)

<table>
<thead>
<tr>
<th>Model</th>
<th>Cl. acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHM</td>
<td>~ 69</td>
</tr>
<tr>
<td>Frame-based HMM [Papadopoulos2007]</td>
<td>~ 65-70</td>
</tr>
<tr>
<td>State-of-the-art in ~2013</td>
<td>80+</td>
</tr>
<tr>
<td>McFee 2017</td>
<td>85+*</td>
</tr>
</tbody>
</table>

* Significantly larger number of classes, different DB (Beatles included)

Published in Proc. Of ISMIR 2014 – *Compositional hierarchical model for music information retrieval*
Multiple Fundamental Frequency Estimation

• Goal: identify **pitches** in audio
  – CHM encodes a robust frequency-invariant concept of pitch

• Learn three compositional layers
  – part **activations** can be transparently mapped to **pitches**

• We evaluated the influence of **different training datasets** on the generated models
  – hierarchies generated from single piano notes, rock music etc. were explored
  – differences in hierarchies were small, all learned different ways to represent pitch

• Further experiments were performed on a **small dataset** of 88 piano key samples
Results: MFFE

• Evaluate if CHM can be used as a robust and transparent classifier
  – the same trained model was applied to different datasets and compared to other approaches

• CHM features:
  – Robustness (others approaches often overfit and don’t perform so well in noisy/real-world situations)
  – Low computational (is real time) & memory footprint (can be used in mobile devices ...)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CHM</th>
<th>DNMF</th>
<th>Klapuri</th>
<th>Benetos [14]</th>
<th>Benetos [56]</th>
<th>Onsets &amp; frames 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPS MIDI</td>
<td>52.6</td>
<td>61.6</td>
<td>56.0</td>
<td>56.7</td>
<td>~60</td>
<td>~78</td>
</tr>
<tr>
<td>MAPS D</td>
<td>51.8</td>
<td>57.1</td>
<td>52.5</td>
<td>50.1</td>
<td>~60</td>
<td></td>
</tr>
<tr>
<td>Su &amp; Yang</td>
<td>48.9</td>
<td>32.6</td>
<td>48.0</td>
<td>40.3</td>
<td>55.6</td>
<td></td>
</tr>
<tr>
<td>Folk song</td>
<td>49.3</td>
<td>35.0</td>
<td>31.8</td>
<td>27.5</td>
<td>16.2</td>
<td></td>
</tr>
<tr>
<td>Running time (s)</td>
<td>6.2</td>
<td>5.7*</td>
<td>19.4</td>
<td>188.1</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>RAM Usage (MB)</td>
<td>63.8</td>
<td>120.0</td>
<td>43.2</td>
<td>1914.2</td>
<td>716.5</td>
<td></td>
</tr>
</tbody>
</table>

The table shows F1 scores of different approaches on different datasets
Part 4

The Compositional Hierarchical Model for Symbolic Representations
CHM: Symbolic Representations

• Input: **symbolic** data (e.g. MIDI)
  – onset time, pitch, magnitude

• Compositions
  – $\mu, \sigma$ represent **pitch distances** (e.g. in semitones)
  – Relatively encoded **melodic patterns**, increased length over layers

• Activations
  – **pattern occurrences** in input

• Aim
  – Learn and analyze melodic patterns that occur within a piece or music corpus
A practical example
Evaluation

- **MIREX intra-opus** pattern discovery task:
  - find melodic patterns in individual works
  - good for comparison to other approaches
- Model with **6 layers** trained on pieces
  - patterns from layers 4-6 exported
- **Measures**: compare discovered to annotated patterns
  - $F_{\text{1est}}$: to what extent an algorithm can discover one pattern occurrence (time shifted, transposed)
  - $F_{\text{1occ}}$: to what extent it can find all occurrences
  - $TLF_1$: balanced three layer F1 score
- Good results
  - make use of model **transparency**
  - no musicological know-how used
  - improved pattern selection algorithm developed: SymCHM Merge

<table>
<thead>
<tr>
<th>Alg</th>
<th>$F_{\text{1est}}$</th>
<th>$F_{\text{1occ}}$</th>
<th>$TLF_1$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SymCHM</td>
<td>42.32</td>
<td>67.24</td>
<td>37.78</td>
<td>5.12</td>
</tr>
<tr>
<td>NF1</td>
<td>50.21</td>
<td>40.8</td>
<td>33.29</td>
<td>2.35</td>
</tr>
<tr>
<td>OL1</td>
<td>49.76</td>
<td>74.5</td>
<td>42.75</td>
<td>12.36</td>
</tr>
<tr>
<td>VM2</td>
<td>62.73</td>
<td>51.54</td>
<td>46.19</td>
<td>6.19</td>
</tr>
<tr>
<td>NF1'13</td>
<td>43.87</td>
<td>34.19</td>
<td>30.41</td>
<td>1.18</td>
</tr>
<tr>
<td>DM10'13</td>
<td>54.78</td>
<td>56.94</td>
<td>43.26</td>
<td>3.25</td>
</tr>
</tbody>
</table>

**MIREX 2015 evaluation**

Published in MDPI Applied Sciences 2017 – SymCHM—An Unsupervised Approach for Pattern Discovery in Symbolic Music with a Compositional Hierarchical Model
Tune family identification

• Goal: classify melodies into classes of related melodies
  – tune families

• SymCHM as a feature extractor for classification
  – Single model for a set of songs
  – Activations of model parts -> feature vectors

• Datasets:
  – OSNP - Slovenian folk songs - Ethnomusicological institute
    • compare also to human classification
  – MTC-ANN – Dutch folk songs – Meertens institute

<table>
<thead>
<tr>
<th></th>
<th>SymCHM</th>
<th>Ann. 1</th>
<th>Ann. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSNP</td>
<td>0.34</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>MTC-ANN</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tune family classification F1 scores

Published in Proc. of FMA 2018 – Modeling song similarity with unsupervised learning
Part 5

The Compositional Hierarchical Model for Rhythm Modeling
Rhythm Modeling - Goals

- Input: event onset times & magnitudes
- Basic unit: distance of two events
- Extend **part definition**: two \((\sigma, \mu)\) parameters
  - \(\sigma_1, \mu_1\) - relative scale
  - \(\sigma_2, \mu_2\) - relative offset

- **Activation**
  - Location, scale, magnitude

- **Goals:**
  - Learn tempo independent **rhythmic patterns**
  - Rhythm genre identification
  - Robustness tempo/beat variations in live music
Analysis

- Extract **patterns** from the Ballroom dataset
  - compare patterns of different genres
- Extract patterns from **live** audio
- The model can
  - Differentiate between music genres
  - Differentiate between different meters within a song
  - Adjust to uneven tempo
Conclusion

• The **scientific contributions** as envisioned in the proposal were met:
  – The Compositional hierarchical model was developed and applied to different MIR tasks (ISMIR 2014)
  – The model was extended for time-dependent music processing (Plos ONE 2017)
  – Model was applied to classification and discovery tasks (MDPI Applied sciences 2017)
• Work currently **in progress:**
  – Tune family classification (FMA 2018)
  – Rhythm modeling (TBP)
  – Melodic prediction (TBP)
Publications

http://musiclab.si


This dissertation is a result of doctoral research, in part financed by the European Union, European Social Fund and the Republic of Slovenia, Ministry for Education, Science and Sport in the framework of the Operational programme for human resources development for the period 2007 – 2013.