Univerza v Ljubljani
Fakulteta za računalništvo
in informatiko

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

Dissertation
21.9.2018

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

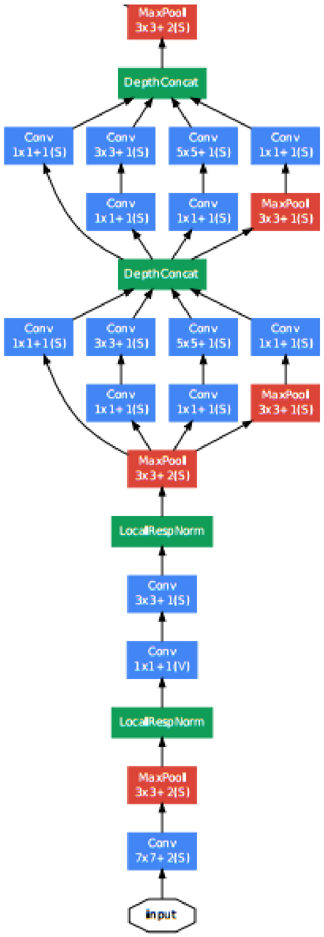
- Musicology [Lerdahl1983, McDermott2008] (rules)
- Psychology [Gelfand2004, Tirovolas2011] (perception and cognition)
- Neuroscience [Amitay2006, Peretz2003, Werner2012] (mechanisms)
- Computer Science - signal processing and **music information retrieval** (analysis, understanding, retrieval)

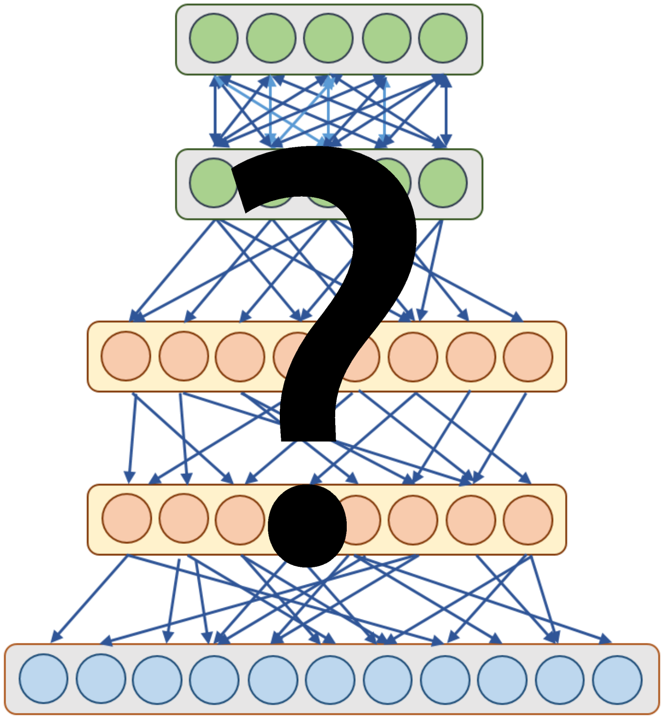
Music information retrieval

- **Interdisciplinary science** of retrieving information from music
- Relatively young field (1970's / late 1990's) [Orio2006]
- **Popular problems** [Downie2008, Downie2010]:
  - Music Recommendation [Eck2007, Song2012, Tkalčič2017]
  - Pattern recognition [Meredith2002, Conklin2010, Ren2017]
  - Extraction of high-level features:
    - Chord estimation [Bello2005, Papadopoulos2007, Deng2016, Korzeniowski2016, McFee2017]
    - Multi-pitch estimation [Klapuri2004,Marolt2004, Emiya2010, Bittner2017, Hawthorne2017]
    - Melody extraction [Ryynanen2008, Salamon2014]
    - Rhythm and beat tracking [Schmidt2013, Pikrakis2013, Bock2015]
    - Genre classification [Tzanetakis2002, Dixon2007, Salamon2012]
    - 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]



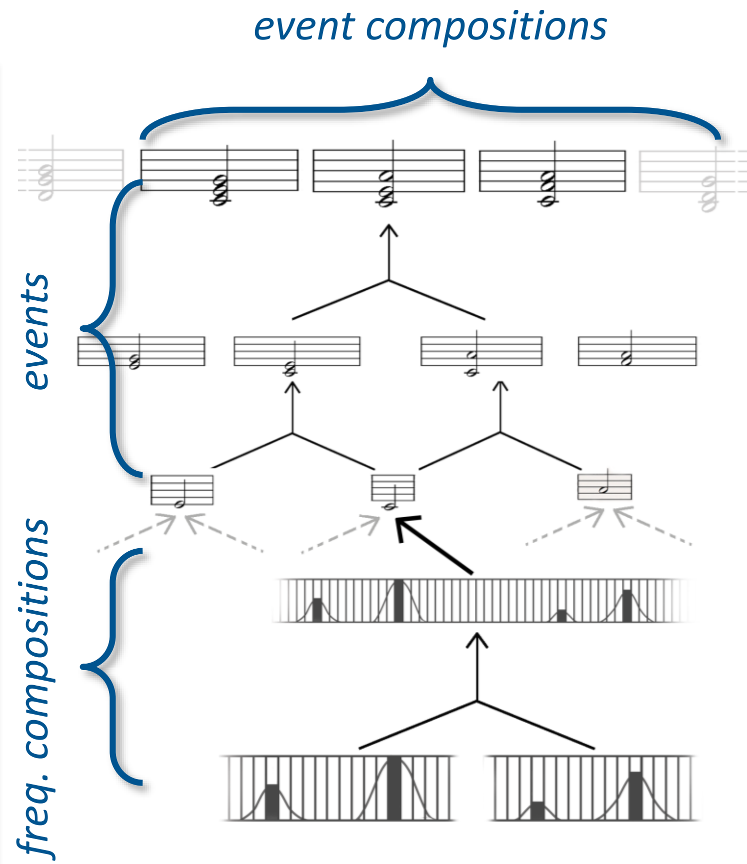


Part 1

The Compositional Hierarchical Model: Motivation

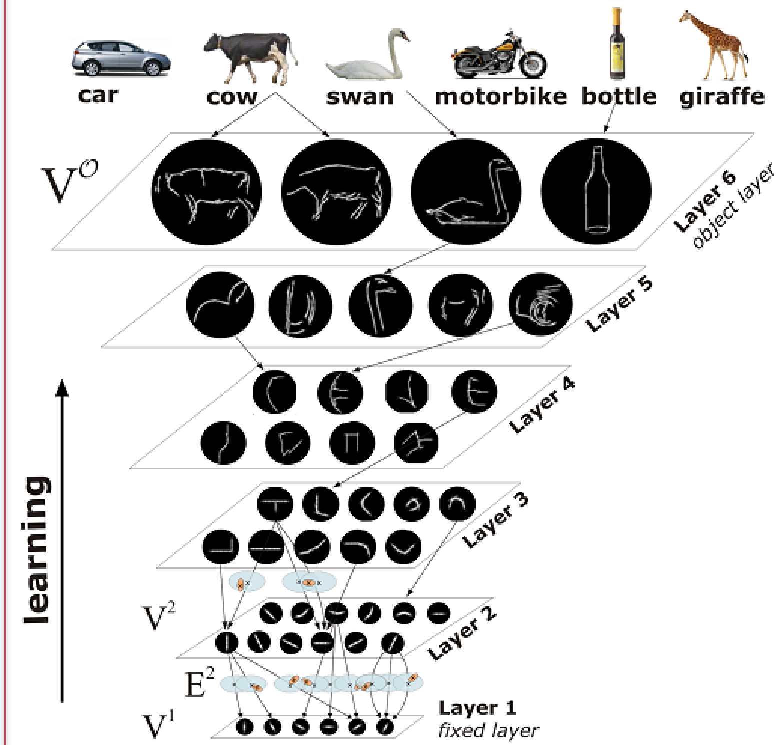
The Compositional Hierarchical Model

- 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

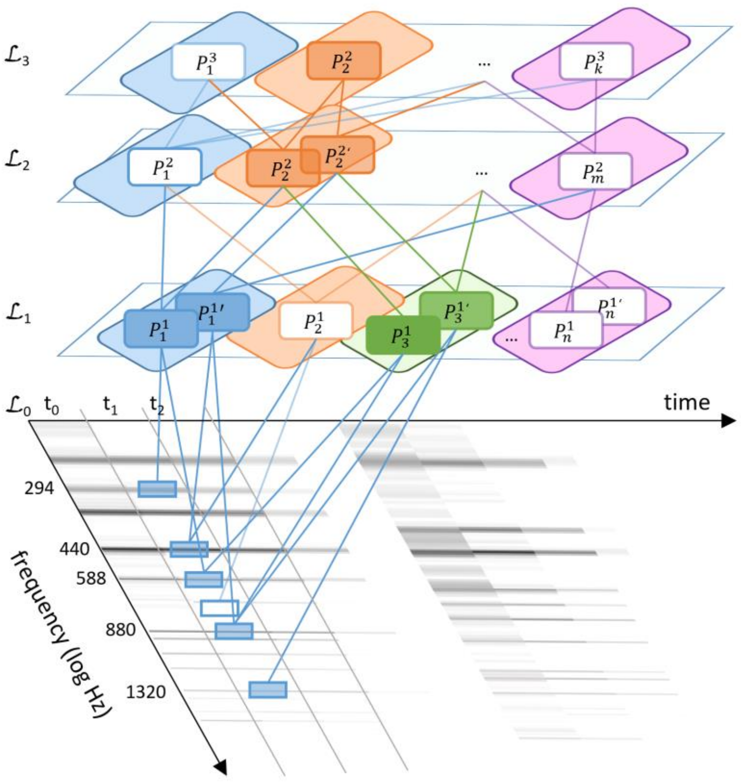
- **Learned Hierarchy of Parts**
- Introduced by Leonardis & Fidler for object categorization in images
- Unsupervised learning of a hierarchy of parts
  - Small image segments on lower layers
  - Complex shapes on higher layers
  - Transparency
- **Music is hierarchical** in frequency and time
  - The nature of the model coincides well this hierarchical structure

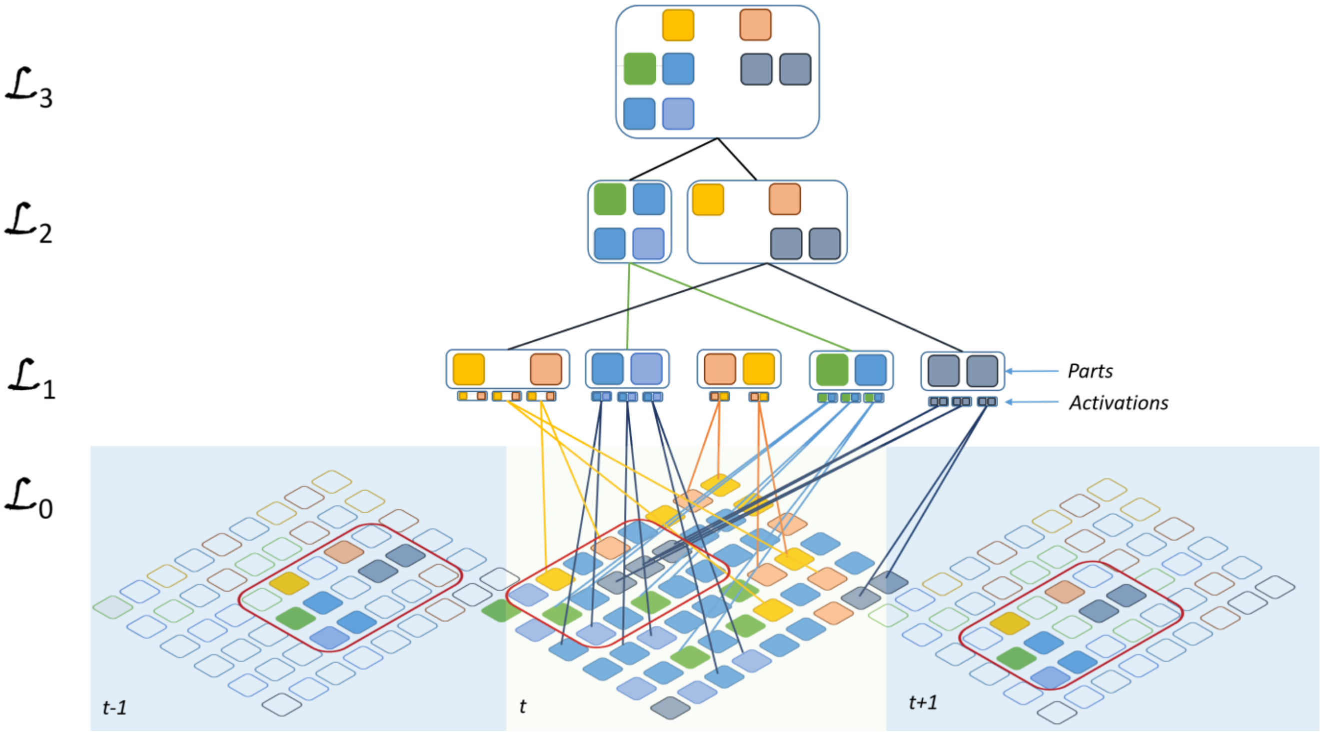


Source: Tabernik et al.

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



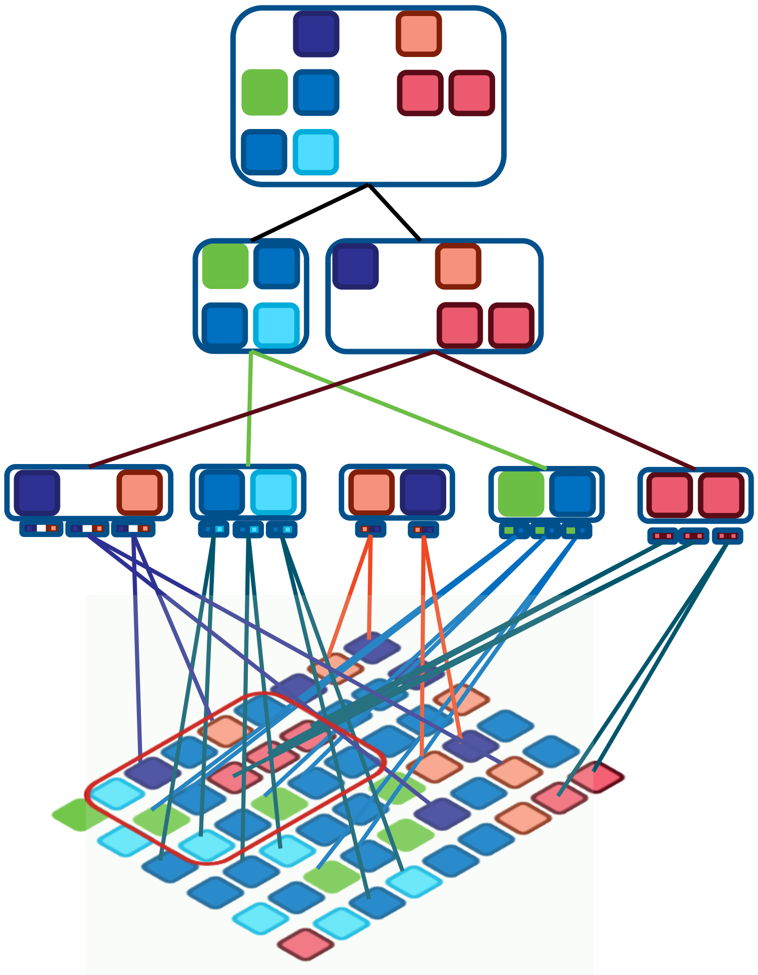


Part 2

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
  - $P_i^n = \left\{ \mathrm{P}_{k_0}^{n-1}, \left\{ P_{k_j}^{n-1}\left(\mu_j, \sigma_j\right) \right\}_{j=1}^{K-1} \right\}$
  - 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**



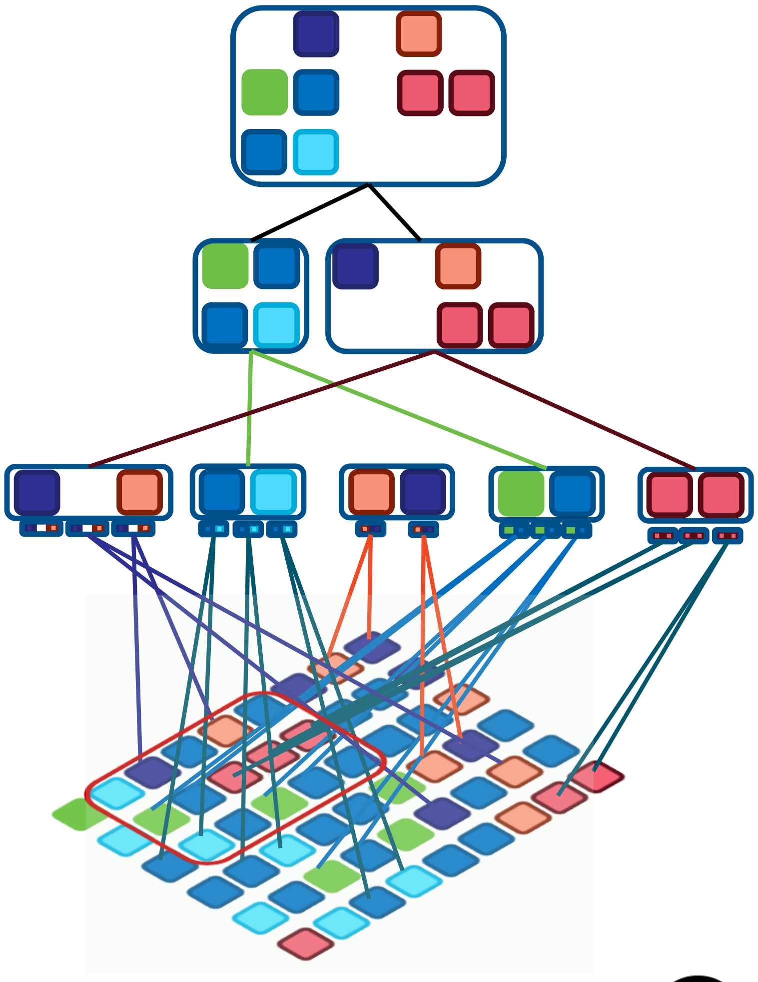
Learning

- 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

procedure SELECT($\mathcal{C}$)
$prevCov \leftarrow 0$
$cov \leftarrow \emptyset$
$\mathcal{L}_n \leftarrow \emptyset$
$sumInput \leftarrow |\mathcal{I}|$
repeat
for $P \in \mathcal{C}$ do
$c \leftarrow 0$
$\mathcal{F} \leftarrow C(\mathcal{L}_n \cup P)$
$c \leftarrow c + |\mathcal{F}|$
$cov[P] \leftarrow c / sumInput$
end for
$Chosen \leftarrow \underset{P}{\operatorname{argmax}}(cov)$
$\mathcal{L}_n \leftarrow \mathcal{L}_n \cup Chosen$
$\mathcal{C} \leftarrow \mathcal{C} \setminus Chosen$
if $cov[Chosen] - prevCov < \tau_C$ then
break
end if
$prevCov \leftarrow cov[Chosen]$
until $prevCov > \tau_P \vee \mathcal{C} = \emptyset$
return $\mathcal{L}_n$

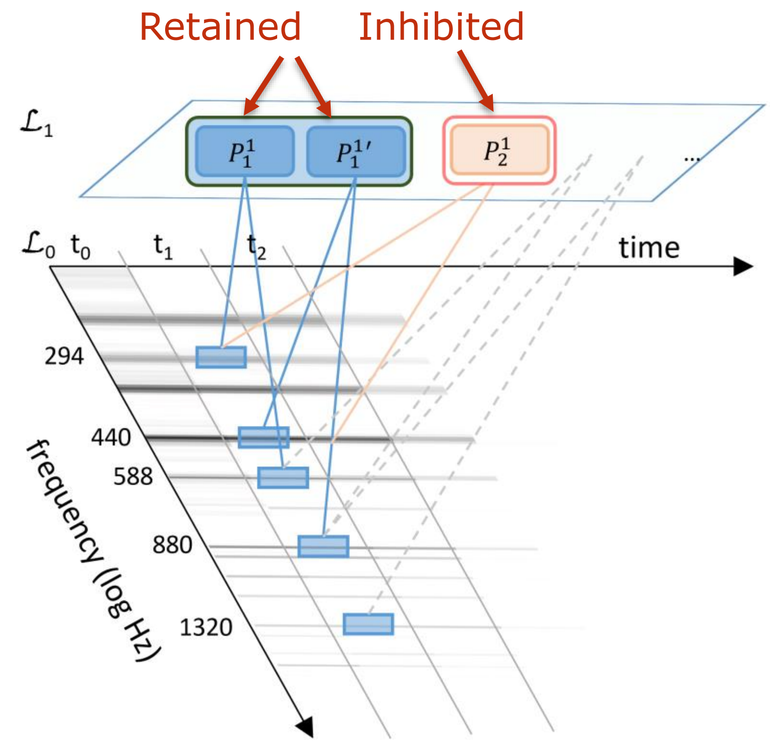
Inference

- 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):
    - $A_L(P_i^n) = A_L(P_{k_0}^{n-1})$
    - $A_T(P_i^n) = A_T(P_{k_0}^{n-1})$
- Activations are **interpretable**



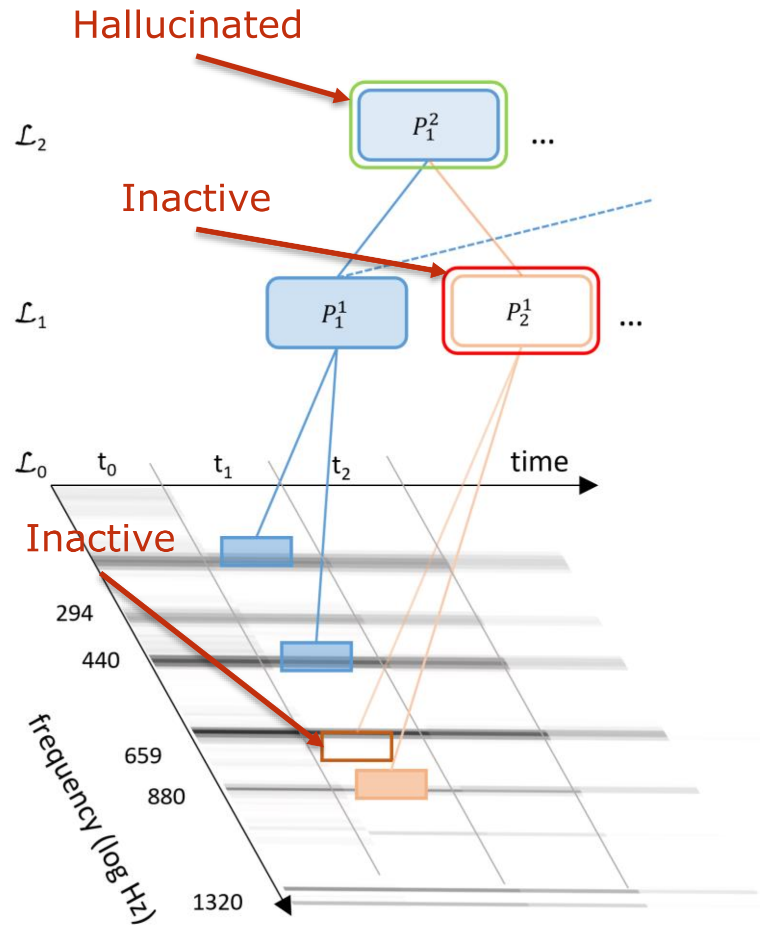
Inhibition

- 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

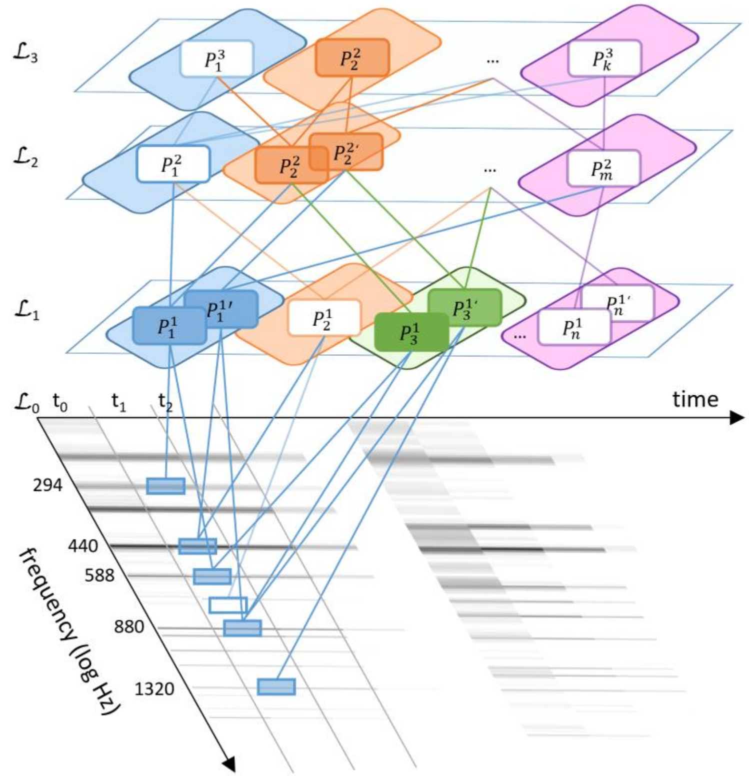


Part 3

The Compositional Hierarchical Model for Time-Frequency Representations

CHM: Time-Frequency Representations

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

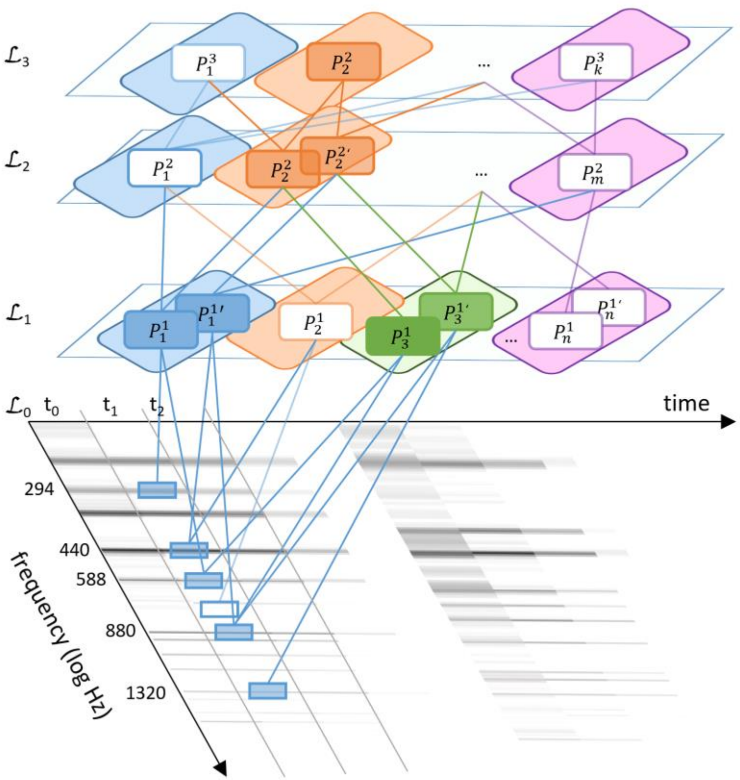
Model	Cl. acc. (%) (The Beatles)
CHM	~ 69
Frame-based HMM [Papadopoulos2007]	~ 65-70
State-of-the-art in ~2013	80+
McFee 2017	85+*

* 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



Published in Plos ONE 2017 – *Robust Real-Time Music Transcription with a Compositional Hierarchical Model*

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 ...)

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

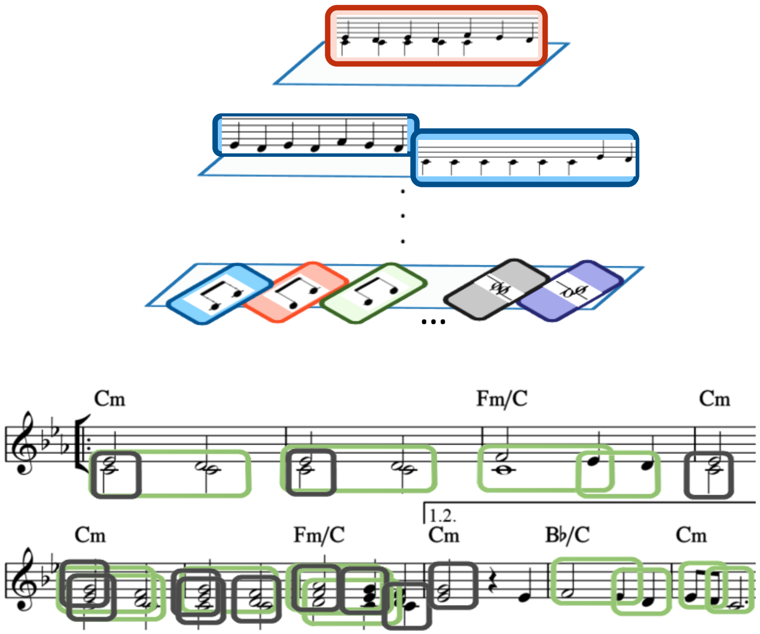
The table shows F1 scores of different approaches on different datasets

Part 4

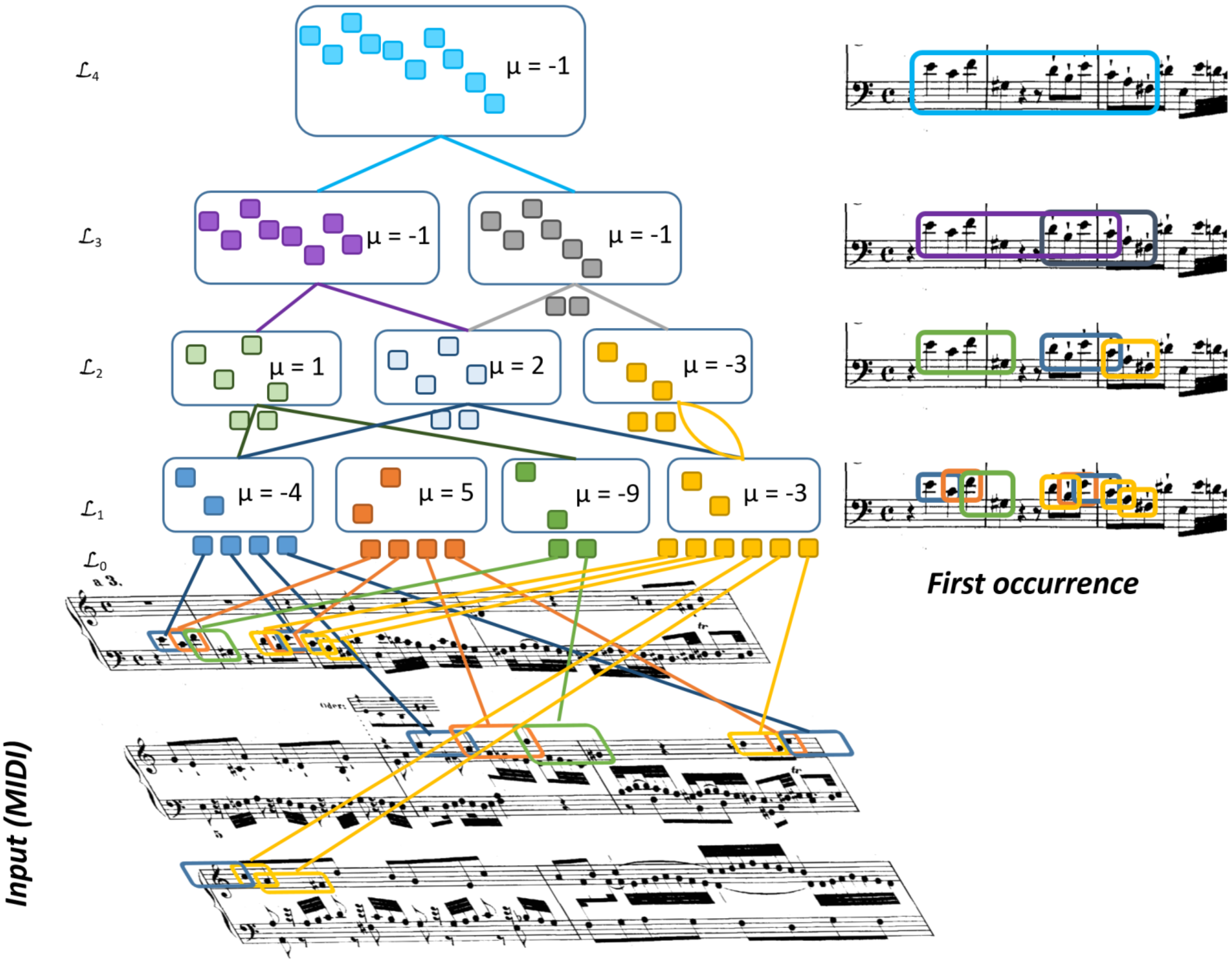
The Compositional Hierarchical Model for Symbolic Representations

CHM: Symbolic Representations

- Input: **symbolic** data (e.g. MIDI)
  - onset time, pitch, magnitude
- Compositions
  - μ, σ 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_{1est}: to what extent an algorithm can discover one pattern occurrence (time shifted, transposed)
  - F_{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: SymCHMMerge

Alg	F_{1est}	F_{1occ}	TLF_1	F_1
SymCHM	42.32	67.24	37.78	5.12
NF1	50.21	40.8	33.29	2.35
OL1	49.76	74.5	42.75	12.36
VM2	62.73	51.54	46.19	6.19
NF1'13	43.87	34.19	30.41	1.18
DM10'13	54.78	56.94	43.26	3.25

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

	SymCHM	Ann. 1	Ann. 2
OSNP	0.34	0.36	0.35
MTC-ANN	0.74		

Tune family classification F1 scores

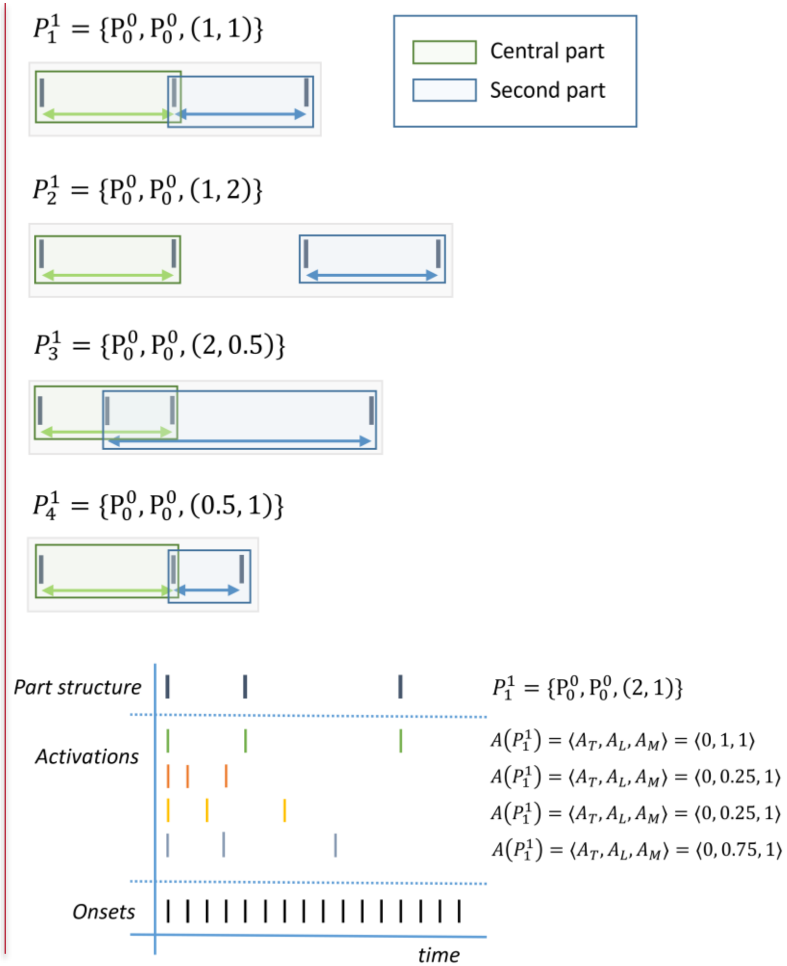
Published in Proc. of FMA 2018 – *Modeling song similarity with unsupervised learning*

Part 5

The Compositional Hierarchical Model for Rhythm Modeling

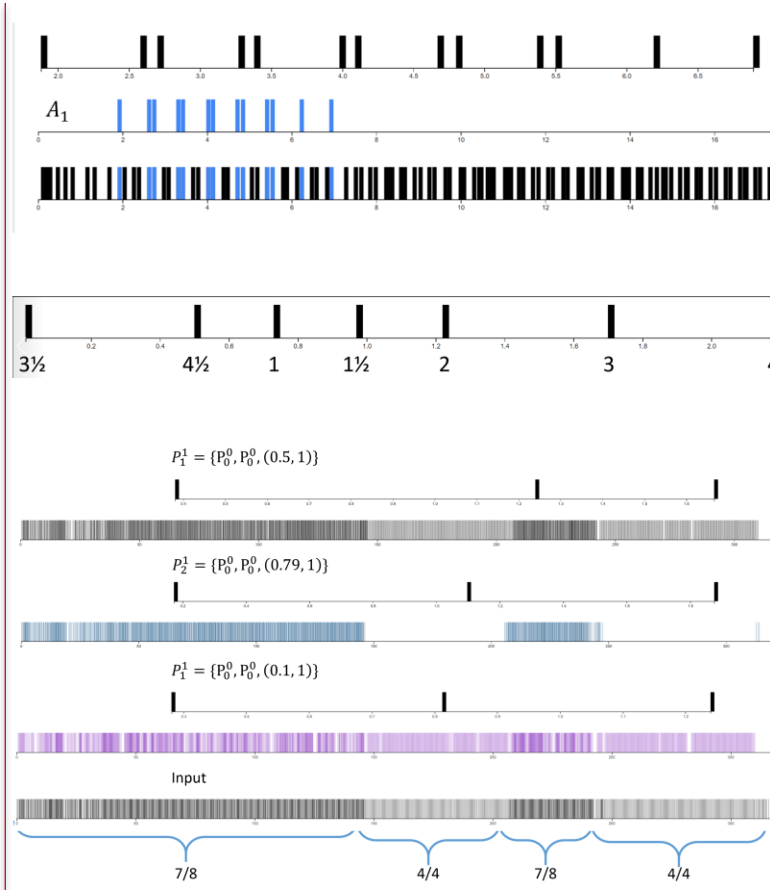
Rhythm Modeling - Goals

- Input: event onset times & magnitudes
- Basic unit: distance of two events
- Extend **part definition**: two (σ, μ) parameters
  - σ_1, μ_1 - relative scale
  - σ_2, μ_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

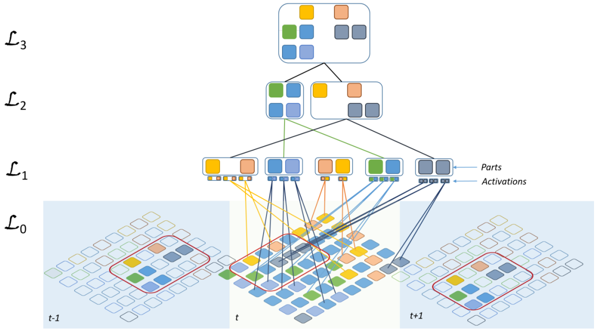


Conclusion

- 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



- M. Pesek, A. Leonardis, and M. Marolt. A compositional hierarchical model for music information retrieval. Y-H. Yang, J. H. Lee, editors, Proc. of ISMIR 2014, pages 131–136, Taipei (TW), 2014.
- M. Pesek, A. Leonardis, and M. Marolt. Robust real-time music transcription with a compositional hierarchical model. PloS one, 12(1):1–21, 2017.
- M. Pesek, A. Leonardis, and M. Marolt. SymCHM—An Unsupervised Approach for Pattern Discovery in Symbolic Music with a Compositional Hierarchical Model. Applied Sciences, 7(11):1–20, 2017.
- M. Pesek, and M. Marolt. Compositional hierarchical model for music understanding. In Proc. of CogMIR 2013, Toronto (CA), 2013.
- M. Pesek, F. Mihelic. Hidden Markov model for chord estimation using compositional hierarchical model features. In Proc. of ERK 2013, pages 145–148, Portoroz (SI), 2013. IEEE.
- M. Pesek, Guna J, A. Leonardis, and M. Marolt. Visualization of a deep architecture using the compositional hierarchical model. In Proc. of ICWUD 2013, pages 145–148, Ljubljana (SI), 2013.
- M. Pesek, and M. Marolt. Chord estimation using compositional hierarchical model. In Proc. of MML 2013, Prague (CZ), 2013.
- M. Pesek, A. Leonardis, and M. Marolt. Boosting audio chord estimation using multiple classifiers. In Proc. of IWSSIP 2014, pages 107–110, Zagreb (HR), 2014. IEEE
- M. Pesek, A. Leonardis, and M. Marolt. A preliminary evaluation of robustness to noise using the compositional hierarchical model for music information retrieval. In Zajc B., Trost A., editors, Proc. of ERK 2014, pages 104–107, Portoroz (SI), 2014. IEEE.
- M. Zerovnik, M. Pesek, and M. Marolt. Ocenjevanje osnovnih frekvenc z uporabo kompozicionalnega hierarhicnega modela. In Proc. of ERK 2014, pages 265–268, Portoroz (SI), 2014. IEEE.
- M. Pesek, A. Leonardis, and M. Marolt. Compositional hierarchical model for pattern discovery in music. Berge P., editor, Proc. of EuroMAC 2014, pages 288, Leuven (BE), 2014.
- M. Pesek, A. Leonardis, and M. Marolt. Towards pattern discovery in symbolic music representations using a compositional hierarchical model. In Proc. of ERK 2014, pages 57–60, Portoroz (SI), 2015. IEEE.
- M. Pesek, L. Zakrajsek, and M. Marolt. WEBCHM: an online tool for music analysis, transcription and annotation. In Proc. of ISMIR 2015, Malaga (ES), 2016.
- M. Pesek, A. Leonardis, and M. Marolt. Pattern discovery and music similarity with compositional hierarchical model. In Proc. of CogMIR 2016, New York (NY), 2016.
- M. Pesek, A. Leonardis, and M. Marolt. SymCHMMerge - hypothesis refinement for pattern discovery with a compositional hierarchical model. In Proc. of MML 2013, Riva del Garda (IT), 2016.
- M. Pesek, M. Žerovnik, A. Leonardis, M. Marolt. Modeling song similarity with unsupervised learning. In Proc. of FMA 2018, Thessaloniki (GR), 2018

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.

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad