Selected Topics in Deep Learning for Audio, Speech, and Music Processing

Introduction to Music Processing

Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

26.04.2021
Group Meinard Müller

- Frank Zalkow
- Sebastian Rosenzweig
- Michael Krause
- Yigitcan Özer
- Peter Meier (extern)

- Christian Dittmar
- Christof Weiß
- Stefan Balke
- Jonathan Driedger
- Thomas Prätzlich
- …
Music
Music Information Retrieval (MIR)
Music Information Retrieval (MIR)

- Signal Processing
- Machine Learning
- Information Retrieval
- Library Sciences
- User Interfaces
- Musicology
- Music
Music Information Retrieval (MIR)

Sheet Music (Image)

CD / MP3 (Audio)

MusicXML (Text)

Dance / Motion (Mocap)

MIDI

Singing / Voice (Audio)

Music Film (Video)

Music Literature (Text)
Piano Roll Representation
Player Piano (1900)
Piano Roll Representation (MIDI)

J.S. Bach, C-Major Fuge

(Well Tempered Piano, BWV 846)
Piano Roll Representation (MIDI)

Query:

Goal: Find all occurrences of the query
Piano Roll Representation (MIDI)

Query:

Goal: Find all occurrences of the query

Matches:
Music Retrieval

Retrieval tasks:

Audio identification
- Bernstein (1962)
- Beethoven, Symphony No. 5

Audio matching
- Beethoven, Symphony No. 5:
  - Bernstein (1962)
  - Karajan (1982)
  - Gould (1992)

Version identification
- Beethoven, Symphony No. 9
- Beethoven, Symphony No. 3
- Haydn Symphony No. 94

Category-based music retrieval
Music Retrieval

Retrieval tasks:

Audio identification
Audio matching
Version identification
Category-based music retrieval

Specificity

High specificity
Low specificity

Granularity

Fragment-based retrieval
Document-based retrieval

Modalities
Music Retrieval

- Remix / remaster retrieval
- Cover song detection
- Version identification
- Plagiarism detection
- Copyright monitoring
- Audio fingerprinting
- Audio identification
- Variation / motif discovery
- Musical quotations discovery
- Audio matching
- Key / mode discovery
- Year / epoch discovery
- Loudness-based retrieval
- Tag / metadata inference
- Mood classification
- Genre / style similarity
- Recommendation
- Instrument-based retrieval

Granularity

Specificity

- high
- low
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould
Application: Interpretation Switcher
Music Synchronization: Audio-Audio

Task

**Given:** Two different audio recordings (two versions) of the same underlying piece of music.

**Goal:** Find for each position in one audio recording the *musically* corresponding position in the other audio recording.
Music Synchronization: Audio-Audio

Traditional Engineering Approach:

1.) Feature extraction
   - Robust to variations (e.g., instrumentation, timbre, dynamics)
   - Discriminative (e.g., capturing harmonic, melodic, tonal aspects)
   - Chroma features

2.) Temporal alignment
   - Capturing local and global tempo variations
   - Trade-off: Robustness vs. accuracy
   - Efficiency
   - Dynamic time warping (DTW)
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould

Time–chroma representations
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould

Time–chroma representations
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Time–chroma representations

Karajan

Gould
Music Synchronization: Audio-Audio

Beethoven’s Fifth

Karajan

Gould

Time–chroma representations
Music Synchronization: Audio-Audio

[Diagram showing synchronization between Karajan and Gould]
Music Synchronization: Audio-Audio

Cost matrix
Music Synchronization: Audio-Audio

Cost matrix
Music Synchronization: Audio-Audio

Cost-minimizing warping path
Music Synchronization: Audio-Audio

Optimal alignment (cost-minimizing warping path)
Music Synchronization: Audio-Audio

Deep Learning Approaches:

- Learn audio features from data
  - Should be able to achieve high alignment accuracy
  - Should be robust to performance variations
  - Musical relevance?

- Alignment problem
  - Pre-aligned data for training
  - Part of loss function $\rightarrow$ differentiability?

Lecture 9: Connectionist Temporal Classification (CTC) Loss with Applications to Theme-Based Music Retrieval
Music Synchronization: Image-Audio

![Image](image.png)

**Image**

**Audio**
Music Synchronization: **Image-Audio**
Application: Score Viewer
How to make the data comparable?
How to make the data comparable?

Image Processing: Optical Music Recognition
How to make the data comparable?

Image Processing: Optical Music Recognition

Audio Processing: Fourier Analysis
How to make the data comparable?

Image Processing: Optical Music Recognition

Audio Processing: Fourier Analysis
Music Synchronization: Image-Audio

Deep Learning Approach:

- Cross-modal embedding
- Requires corresponding snippets of audio and sheet music for training
- Triplet Loss function
  \[ \max(0, d(x^a, y^p) - d(x^a, y^n) + \alpha) \]
- Problem very hard
  - Performance variations
  - Layout variations

Music Synchronization: Image-Audio

Deep Learning Approach: Soft Attention Mechanism

Lecture 7: Attention in Sound Source Localization and Speaker Extraction
## Music Processing

<table>
<thead>
<tr>
<th>Coarse/Relative Level</th>
<th>Fine/Absolute Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do different versions or instances have in common?</td>
<td>What are the characteristics of a specific version or instance?</td>
</tr>
<tr>
<td>Provide coarse description: What makes up a piece of music?</td>
<td>Capture nuances and subtleties: What makes music come alive?</td>
</tr>
<tr>
<td>Identify despite of differences</td>
<td>Identify the differences</td>
</tr>
<tr>
<td>Example tasks:</td>
<td>Example tasks:</td>
</tr>
<tr>
<td><em>Music Retrieval</em></td>
<td><em>Music Transcription</em></td>
</tr>
<tr>
<td><em>Genre Classification</em></td>
<td><em>Performance Analysis</em></td>
</tr>
<tr>
<td><em>Global Tempo Estimation</em></td>
<td><em>Local Tempo Estimation</em></td>
</tr>
</tbody>
</table>
Tempo Estimation and Beat Tracking

Basic task: “Tapping the foot when listening to music”
Tempo Estimation and Beat Tracking

Basic task: “Tapping the foot when listening to music”

Example: Queen – Another One Bites The Dust
Tempo Estimation and Beat Tracking

Basic task: “Tapping the foot when listening to music”

Example: Queen – Another One Bites The Dust
Tempo Estimation and Beat Tracking

Example: Chopin – Mazurka Op. 68-3

Pulse level: Quarter note

Tempo: ???
Tempo Estimation and Beat Tracking

Example: Chopin – Mazurka Op. 68-3

Pulse level: Quarter note

Tempo: 50-200 BPM

Tempo curve
Tempo Estimation and Beat Tracking

Tasks

- Onset detection
- Beat tracking
- Tempo estimation
Tempo Estimation and Beat Tracking

Tasks
- Onset detection
- Beat tracking
- Tempo estimation
Tempo Estimation and Beat Tracking

Tasks

- Onset detection
- Beat tracking
- Tempo estimation
Tempo Estimation and Beat Tracking

Tasks

- Onset detection
- Beat tracking
- Tempo estimation

Tempo := 60 / period

Beats per minute (BPM)
Onset Detection (Spectral Flux)
Onset Detection (Spectral Flux)

Steps:

1. Spectrogram
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation

Spectral difference

Novelty curve
Onset Detection (Spectral Flux)

Steps:

1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation

Novelty function
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation
5. Normalization

Novelty function
Subtraction of local average
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation
5. Normalization

Normalized novelty function
Onset Detection (Spectral Flux)

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation
5. Normalization

Normalized novelty function
Peak positions indicate beat candidates
Onset Detection (Spectral Flux)

Deep Learning Approaches:
1. Input representation
2. Sigmoid activation
3. Convolution & rectified linear unit (ReLU)
4. Pooling
5. Convolution & ReLU

Steps:
1. Spectrogram
2. Logarithmic compression
3. Differentiation & half wave rectification
4. Accumulation
5. Normalization

Normalized novelty function
Peak positions indicate beat candidates
Local Pulse and Tempo Tracking

Normalized novelty function

Time (seconds)
Local Pulse and Tempo Tracking

Fourier temogram (STFT of novelty function)

Normalized novelty function
Local Pulse and Tempo Tracking

Fourier temogram (STFT of novelty function)

Optimizing local periodicity kernel
Local Pulse and Tempo Tracking

**Fourier temogram** (STFT of novelty function)

![Fourier temogram](image)

**Optimizing local periodicity kernel**

![Optimizing local periodicity kernel](image)
Local Pulse and Tempo Tracking

Fourier temogram (STFT of novelty function)

Optimizing local periodicity kernel
Local Pulse and Tempo Tracking

**Fourier temogram** (STFT of novelty function)

Accumulation of kernels
Local Pulse and Tempo Tracking

Fourier temogram (STFT of novelty function)

Halfwave rectification
Local Pulse and Tempo Tracking

Novelty Curve

Predominant Local Pulse (PLP)
Local Pulse and Tempo Tracking

Deep Learning Approaches:

- **End-to-end approach**
  - Input: Short audio snippets
  - Output: Tempo value

- **DL architecture inspired by traditional engineering**
  - Layers and activation functions
  - Shape of convolutional kernels

Automatic Music Transcription

Task: Convert a music recording into sheet music
Automatic Music Transcription

Task: Convert a music recording into sheet music (or another symbolic music representation)
Automatic Music Transcription

Task: Convert a music recording into sheet music (or another symbolic music representation)

Multitask learning for estimating
- pitches,
- note onsets & offsets,
- beat & measure positions,
- musical voices & instrumentation,
- pedalling, dynamics, …
Why is Music Processing Challenging?

Example: Chopin, Mazurka Op. 63 No. 3
Why is Music Processing Challenging?

Example: Chopin, Mazurka Op. 63 No. 3

- Waveform
Why is Music Processing Challenging?

Example: Chopin, Mazurka Op. 63 No. 3

- Waveform / Spectrogram
Why is Music Processing Challenging?

**Example:** Chopin, Mazurka Op. 63 No. 3

- Waveform / Spectrogram
- Performance
  - Tempo
  - Dynamics
  - Note deviations
  - Sustain pedal
Why is Music Processing Challenging?

Example: Chopin, Mazurka Op. 63 No. 3

- **Waveform / Spectrogram**
- **Performance**
  - Tempo
  - Dynamics
  - Note deviations
  - Sustain pedal
- **Polyphony**
  
  - Main Melody
  - Additional melody line
  - Accompaniment
Source Separation

- Decomposition of audio stream into different sound sources
- Central task in digital signal processing
- “Cocktail party effect”
Source Separation

- Decomposition of audio stream into different sound sources
- Central task in digital signal processing
- “Cocktail party effect”
- Several input signals
- Sources are assumed to be statistically independent
Source Separation (Music)

- Main melody, accompaniment, drum track
- Instrumental voices
- Individual note events
- Only mono or stereo
- Sources are often highly dependent
Singing Voice Extraction

Original Recording

Singing voice

Accompaniment
Singing Voice Extraction

**Traditional engineering approach:**

Singing Voice Extraction

Reference voices:

Engineering approach:

Deep learning approach:

Deep learning has lead to breakthrough

Lecture 5: Music Source Separation

Score-Informed Audio Decomposition

Exploit musical score to support decomposition process

NMF (Nonnegative Matrix Factorization)

\[ M \geq 0 \quad \approx \quad K \geq 0 \quad \approx \quad M \geq 0 \]
NMF (Nonnegative Matrix Factorization)

**Magnitude Spectrogram** $N \times M$  $\approx$  **Templates** $M \times K$  $\approx$  **Activations** $K \times M$

**Templates:** Pitch + Timbre

**Activations:** Onset time + Duration

“*How does it sound*”

“*When does it sound*”
NMF-Decomposition

Initialized template

Random initialization
NMF-Decomposition

Random initialization  →  No semantic meaning
NMF-Decomposition

Initialized template

Initialized activations

Constrained initialization
NMF-Decomposition

- Template constraint for $p=55$
- Initialized template

- Initialized activations

- Activation constraints for $p=55$

- Constrained initialization
NMF-Decomposition

Constrained initialization → NMF as refinement
NMF-Decomposition

\[ V \approx WH = \hat{V} \]

Encoder(\(E\)) \hspace{1cm} Decoder(\(D\))


Lecture 6: Nonnegative Autoencoders with Applications to Music Audio Decomposing
Score-Informed Audio Decomposition

Exploit musical score to support decomposition process

NMF-based spectrogram decomposition
Score-Informed Audio Decomposition

Exploit musical score to support decomposition process

NMF-based spectrogram decomposition
Score-Informed Audio Decomposition

Application: Audio editing
Informed Drum-Sound Decomposition

Informed Drum-Sound Decomposition

Major challenge: Reconstructed sound events often have artifacts

Approaches:

- Resynthesize certain sound components
- Differentiable Digital Signal Processing (DDSP) combines classical DSP and deep learning
- Generative adversarial networks may help to reduce the artifacts
Audio Mosaicing

**Target** signal: Beatles–Let it be

**Source** signal: Bees

**Mosaic** signal: Let it Bee

Selected Topics in Deep Learning for Audio, Speech, and Music Processing

1. Introduction to Audio and Speech Processing
2. Introduction to Music Processing
3. Permutation Invariant Training Techniques for Speech Separation
5. Music Source Separation
6. Nonnegative Autoencoders with Applications to Music Audio Decomposing
7. Attention in Sound Source Localization and Speaker Extraction
8. Recurrent and Generative Adversarial Network Architectures for Text-to-Speech
9. Connectionist Temporal Classification (CTC) Loss with Applications to Theme-Based Music Retrieval
10. From Theory to Practise
Book: Fundamentals of Music Processing

Meinard Müller
Fundamentals of Music Processing
Audio, Analysis, Algorithms, Applications
483 p., 249 illus., hardcover
ISBN: 978-3-319-21944-8
Springer, 2015

Accompanying website:
www.music-processing.de
### Book: Fundamentals of Music Processing

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Music Processing Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Music Representations</td>
</tr>
<tr>
<td>2</td>
<td>Fourier Analysis of Signals</td>
</tr>
<tr>
<td>3</td>
<td>Music Synchronization</td>
</tr>
<tr>
<td>4</td>
<td>Music Structure Analysis</td>
</tr>
<tr>
<td>5</td>
<td>Chord Recognition</td>
</tr>
<tr>
<td>6</td>
<td>Tempo and Beat Tracking</td>
</tr>
<tr>
<td>7</td>
<td>Content-Based Audio Retrieval</td>
</tr>
<tr>
<td>8</td>
<td>Musically Informed Audio Decomposition</td>
</tr>
</tbody>
</table>

Meinard Müller  
Fundamentals of Music Processing  
Audio, Analysis, Algorithms, Applications  
483 p., 249 illus., hardcover  
ISBN: 978-3-319-21944-8  
Springer, 2015

Accompanying website:  
www.music-processing.de
Software & Audio: FMP Notebooks

FMP Notebooks
Python Notebooks for Fundamentals of Music Processing

The FMP notebooks offer a collection of educational material closely following the textbook Fundamentals of Music Processing (FMP). This is the starting website, which is opened when calling https://www.audiolabs-erlangen.de/FMP. Besides giving an overview, this website provides information on the license, the main contributors, and some links.

https://www.audiolabs-erlangen.de/FMP