# Machine Listening for Music and Sound Analysis

# Lecture 3 – Music Information Retrieval I

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://machinelistening.github.io

# Overview

- Music Information Retrieval
- Music Tagging
- Music Similarity
- Tempo Estimation

# Music Information Retrieval
## Examples

- Musical Instrument

AUD-1 AUD-2

- Musical Genre / Tempo

AUD-3 AUD-4

# Music Information Retrieval
## Motivation

- Large music collections
- Mobile device apps / instruments

# Music Information Retrieval
## Motivation

- Large music collections
- Mobile device apps / instruments
- Music industry shifts almost completely to online products & services
- Growing market of music streaming services

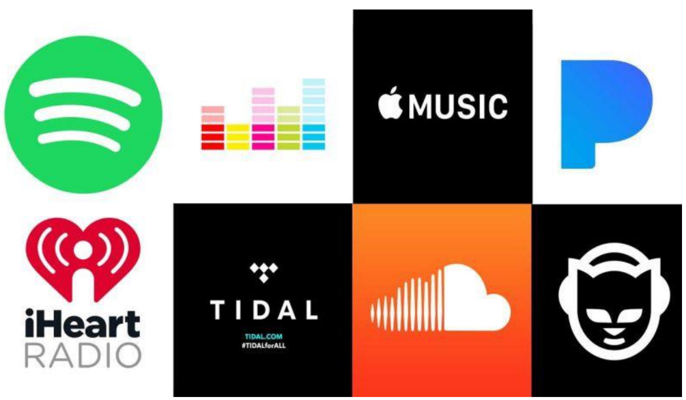



Fig. 1

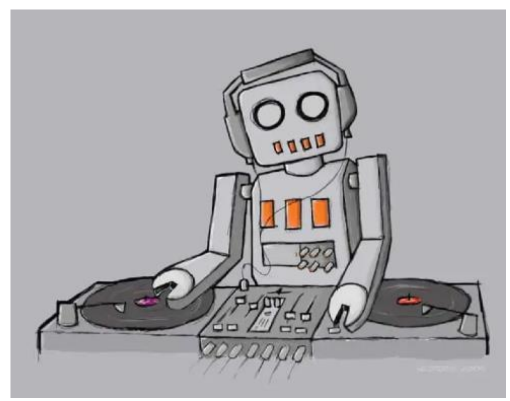


Fig. 2

# Music Information Retrieval
## Typical Research Tasks

- What’s that song again? Who’s singing that?
  - Audio identification

# Music Information Retrieval
## Typical Research Tasks

- What's that song again? Who's singing that?
  - Audio identification
- I want to learn that song on my instrument!
  - Automatic music transcription

# Music Information Retrieval
## Typical Research Tasks

- What’s that song again? Who’s singing that?
    - Audio identification
- I want to learn that song on my instrument!
    - Automatic music transcription
- What songs are similar? How to generate a playlist?
    - Audio similarity search

# Music Information Retrieval
## Typical Research Tasks

- What's that song again? Who's singing that?
  - Audio identification
- I want to learn that song on my instrument!
  - Automatic music transcription
- What songs are similar? How to generate a playlist?
  - Audio similarity search
- How to organize my music? Which genre / style?
  - Audio classification

# Music Information Retrieval
## Research Landscape

- Interdisciplinary research community
    - Musicology / Music Cognition
    - Artificial Intelligence / Signal Processing
    - Human-Computer Interaction
    - Information Retrieval, etc…

# Music Information Retrieval
## Research Landscape

- Interdisciplinary research community
    - Musicology / Music Cognition
    - Artificial Intelligence / Signal Processing
    - Human-Computer Interaction
    - Information Retrieval, etc…

- Conferences
    - ISMIR (International Society for Music Information Retrieval Conference)
    - IEEE ICASSP, DAFx, AES, ICMC, SMC

# Music Information Retrieval
## Research Landscape

- Interdisciplinary research community
  - Musicology / Music Cognition
  - Artificial Intelligence / Signal Processing
  - Human-Computer Interaction
  - Information Retrieval, etc…
- Conferences
  - ISMIR (International Society for Music Information Retrieval Conference)
  - IEEE ICASSP, DAFx, AES, ICMC, SMC
- MIREX competition (Music Information Retrieval Evaluation eXchange)

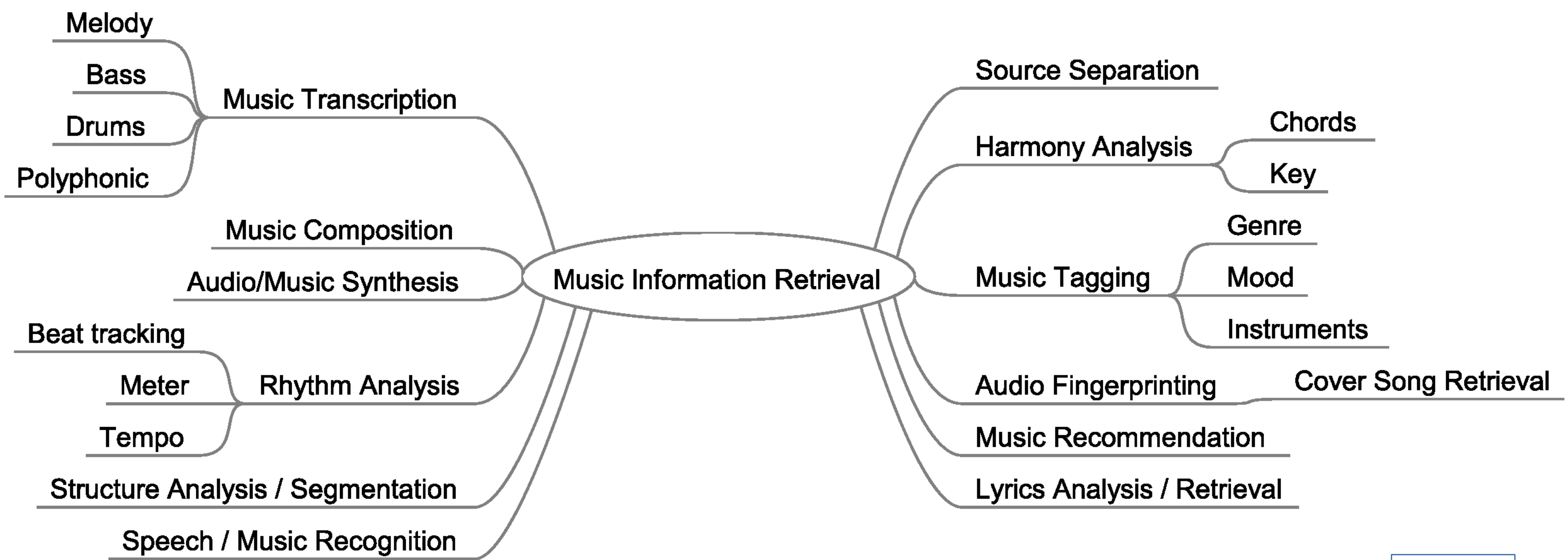
# Music Information Retrieval
## Research Landscape

- MIR @ Fraunhofer IDMT
  - Semantic music technologies (SMT) group
    - Staff + PhD / master / bachelor students + interns

# Music Information Retrieval
## Research Landscape

- MIR @ Fraunhofer IDMT
    - Semantic music technologies (SMT) group
        - Staff + PhD / master / bachelor students + interns
- National / international research groups
    - International Audio Laboratories Erlangen, Germany
    - Centre for Digital Music, Queen Mary University, London, UK
    - Universitat Pompeu Fabra, Barcelona, Spain
    - Institute for music/acoustic research and coordination (IRCAM), Paris, France
    - USA, China, Taiwan, Japan, Korea, etc.

# Music Information Retrieval
## Research Task Taxonomy




Own

# Music Information Retrieval
## Case Studies

- MIR 1 lecture
    - Music tagging / music similarity → general tasks
    - Tempo estimation → rhythm

# Music Information Retrieval
## Case Studies

- MIR 1 lecture
  - Music tagging / music similarity → general tasks
  - Tempo estimation → rhythm
- MIR 2 lecture
  - Pitch detection → pitch / tonality
  - Source separation & instrument recognition → timbre

# Music Information Retrieval
## Case Studies

- MIR 1 lecture
    - Music tagging / music similarity → general tasks
    - Tempo estimation → rhythm
- MIR 2 lecture
    - Pitch detection → pitch / tonality
    - Source separation & instrument recognition → timbre

- Teaching Concept

Task Introduction → Traditional Approach → Novel Approaches

Own

# Music Tagging
## Introduction

- Tags
  - Textual (objective / subjective) annotations of songs

# Music Tagging
## Introduction

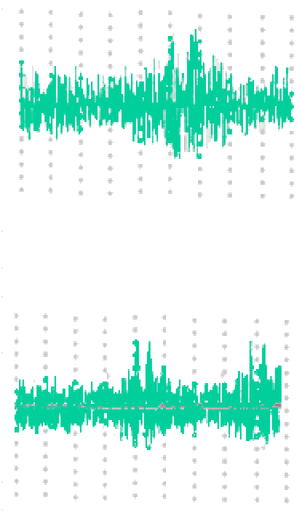
- Tags
  - Textual (objective / subjective) annotations of songs
  - Examples
    - Instruments (drums, bass, guitar, vocals …)
    - Genre (classical, electro, hip hop)
    - Mood (mellow, romantic, angry, happy)
    - Miscellaneous (noise, loud, ambient)

# Music Tagging
## Introduction

- Tags
  - Textual (objective / subjective) annotations of songs
  - Examples
    - Instruments (drums, bass, guitar, vocals …)
    - Genre (classical, electro, hip hop)
    - Mood (mellow, romantic, angry, happy)
    - Miscellaneous (noise, loud, ambient)
- Challenge
  - Music pieces change their characteristics over time
    - E.g.: trumpet plays only in the chorus (jazz)

# Music Tagging
## Traditional Approach

- Audio feature engineering & music domain knowledge

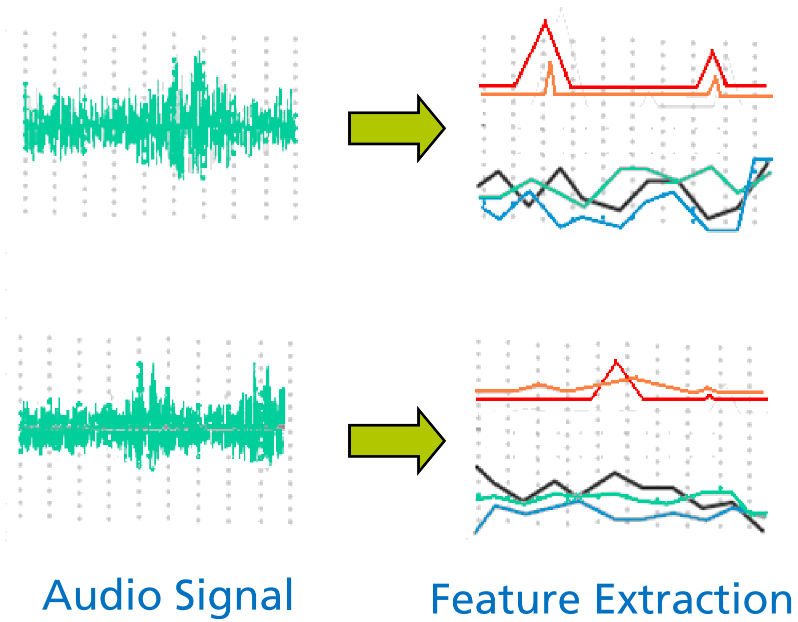


Audio Signal

Own

# Music Tagging
## Traditional Approach

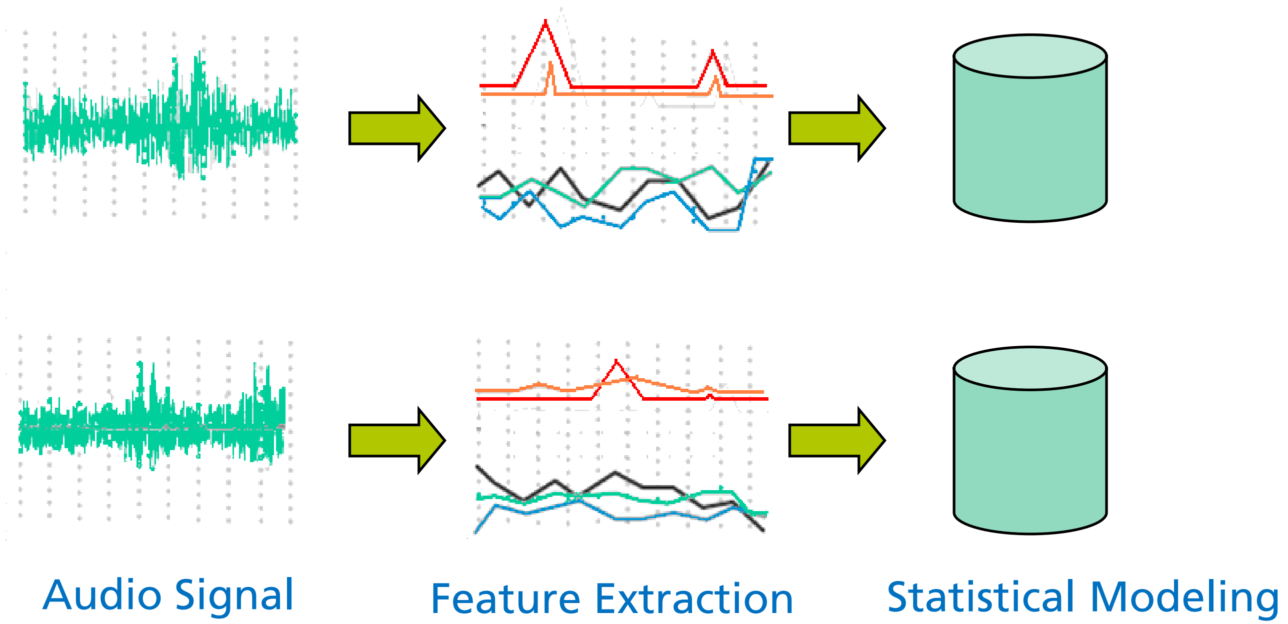
- Audio feature engineering & music domain knowledge




Own

# Music Tagging
## Traditional Approach

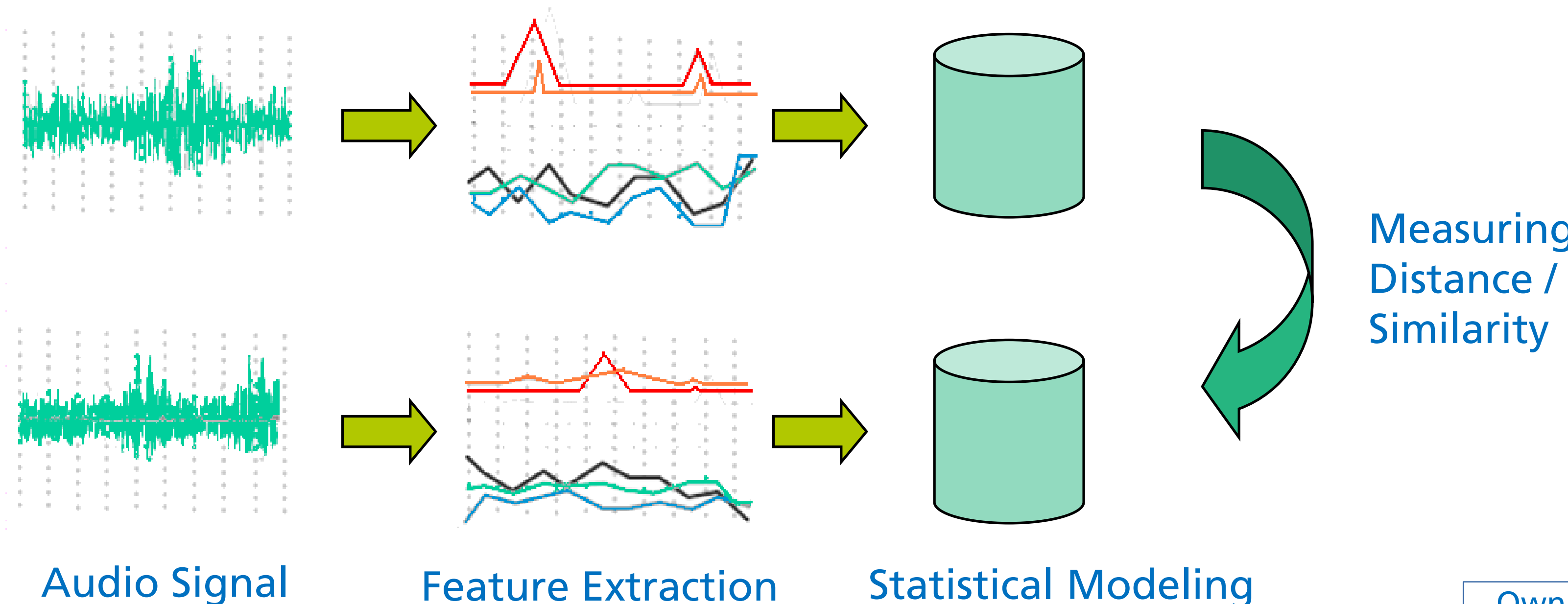
- Audio feature engineering & music domain knowledge




Own

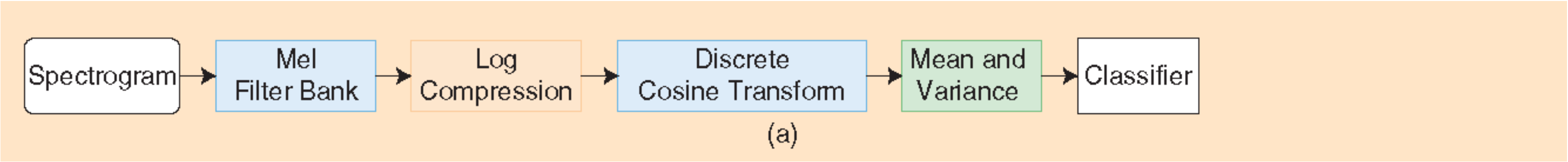
# Music Tagging
## Traditional Approach

- Audio feature engineering & music domain knowledge
- Standard classification methods (GMM, SVM, kNN)




Own

# Music Tagging
## Novel Approaches




Trend

(a) Feature engineering (MFCC)

Fig. 3

# Music Tagging
## Novel Approaches

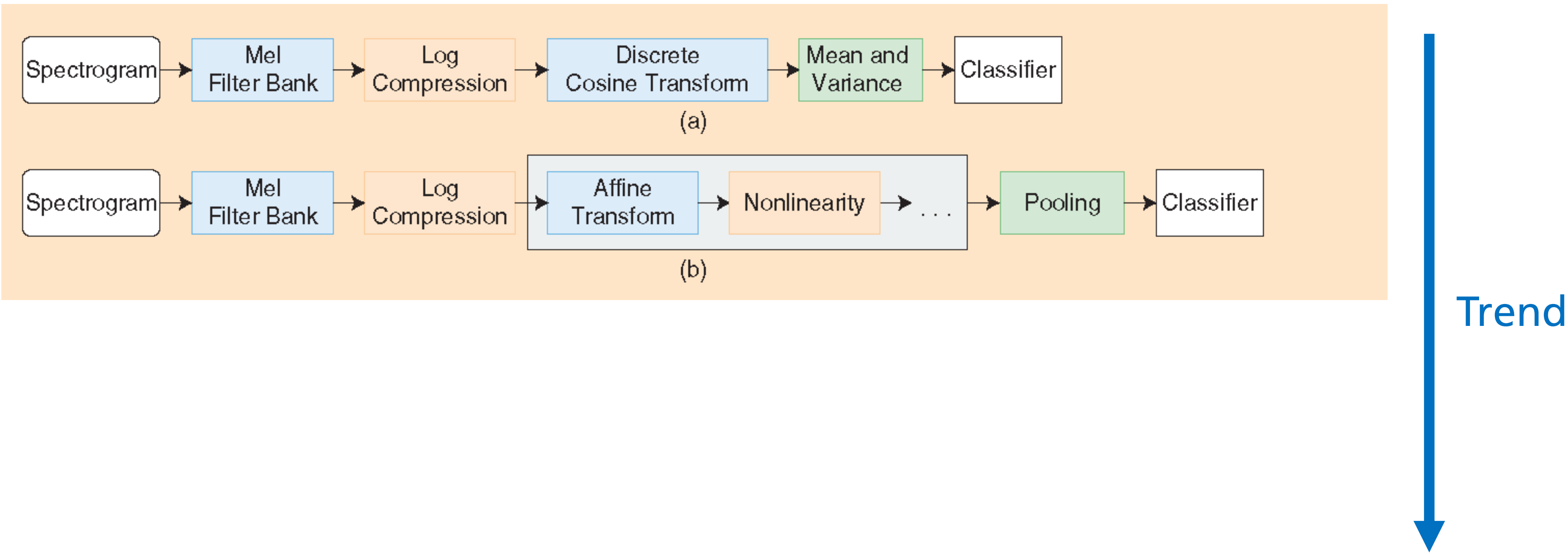
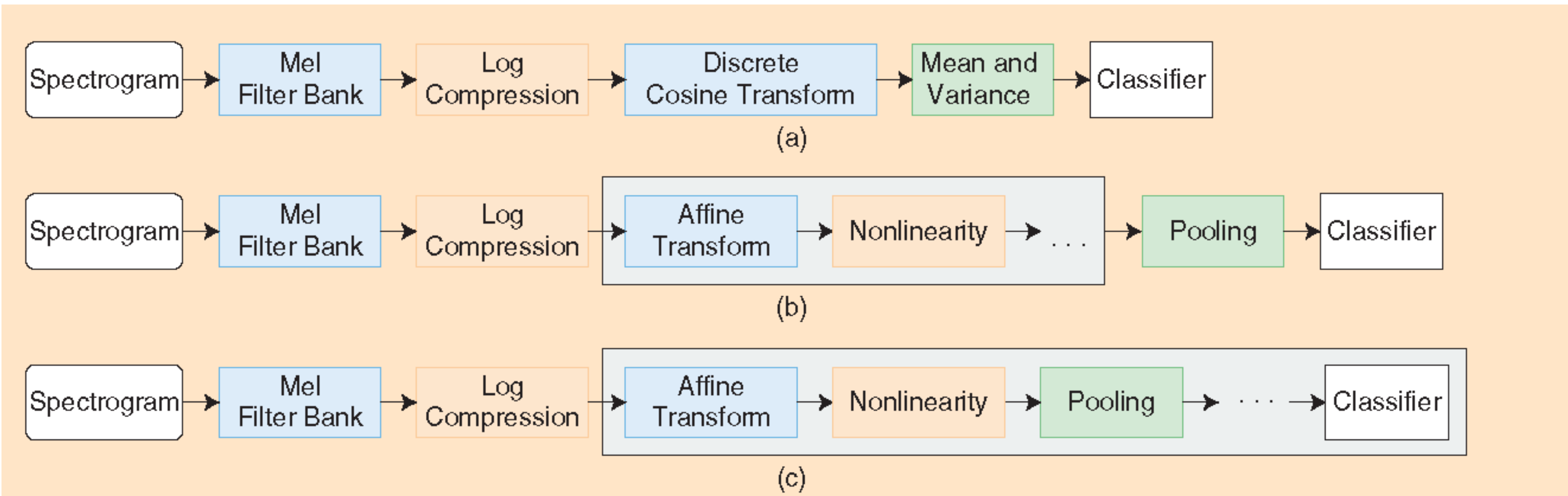



Fig. 3

(a) Feature engineering (MFCC)

(b) Low-level feature

# Music Tagging
## Novel Approaches




Trend

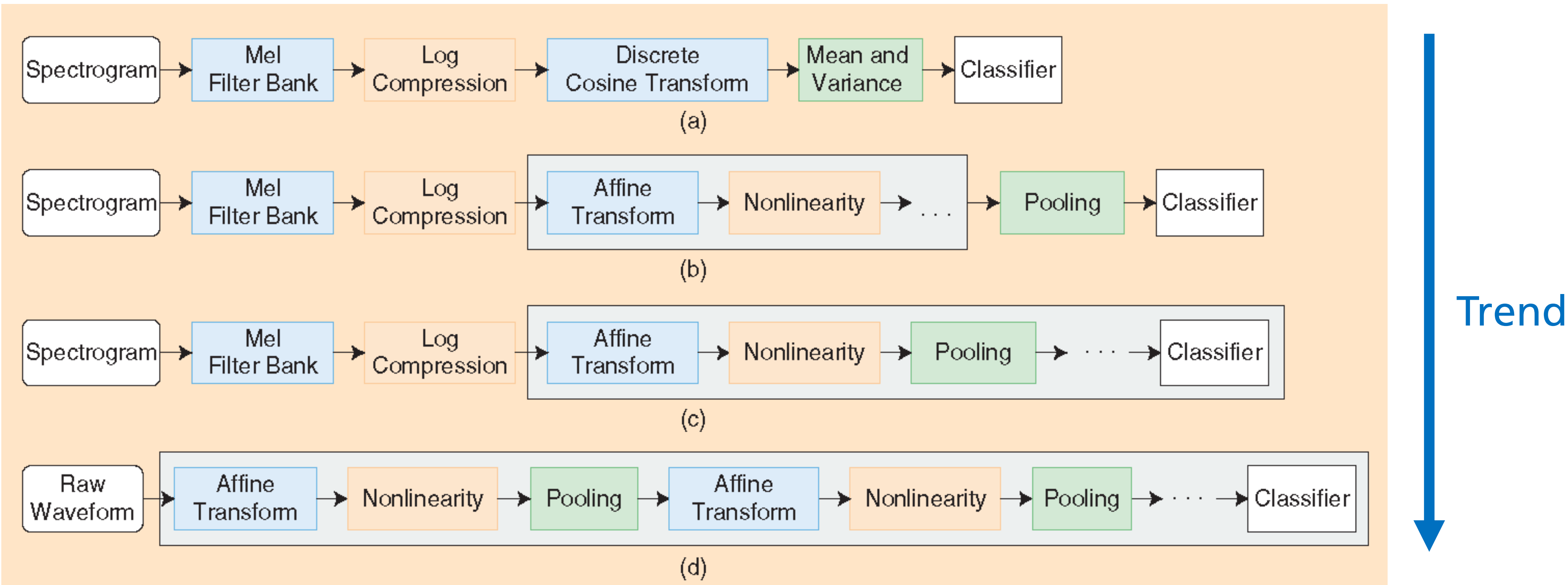
(a) Feature engineering (MFCC)

(b) Low-level feature

(c) Joint feature learning & classification (CNN)

Fig. 3

# Music Tagging
## Novel Approaches




(a) Feature engineering (MFCC)

(b) Low-level feature

(c) Joint feature learning & classification (CNN)

(d) End-to-end learning

Fig. 3

# Music Tagging
## Novel Approaches

- Joint representation learning & classification using CNNs
  - Input: spectrograms (2D) or audio samples (1D end-to-end)

# Music Tagging
## Novel Approaches

- Joint representation learning & classification using CNNs
  - Input: spectrograms (2D) or audio samples (1D end-to-end)

- Integrate musical knowledge in network design (e.g., filter shapes)

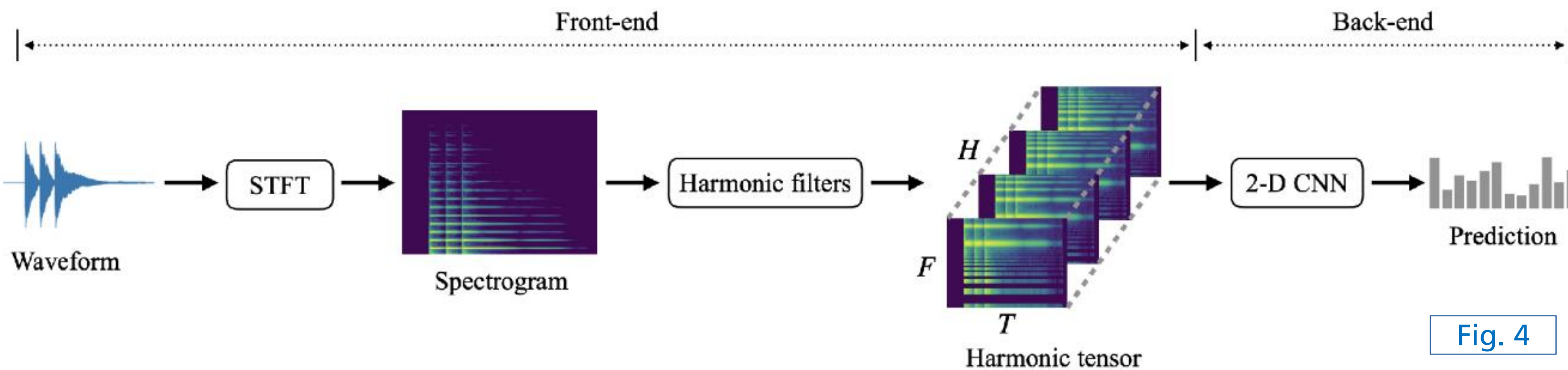



Fig. 4

# Music Tagging
## Novel Approaches

- End-to-end learning
    - Model input is low-level representation (audio waveform)
    - No pre-processing / assumptions required

# Music Tagging
## Novel Approaches

- End-to-end learning
  - Model input is low-level representation (audio waveform)
  - No pre-processing / assumptions required
  - Not restricted to spectral magnitudes → can model phase!
  - Requires large amounts of training data

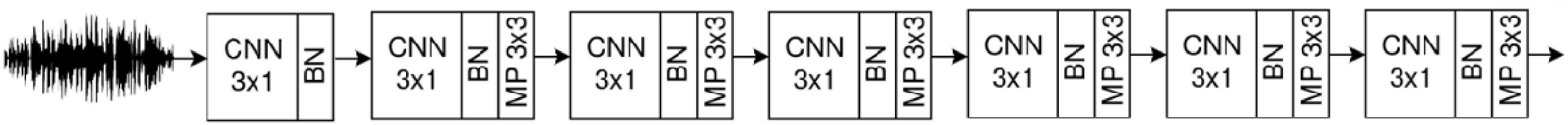



Fig. 6

# Music Tagging
## Novel Approaches

- Transfer Learning
  - Pre-train model on source task (lot of data available)

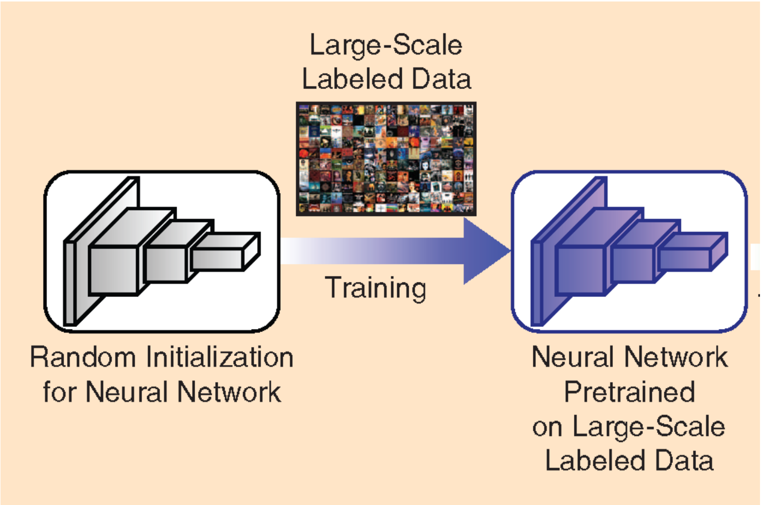



Fig. 5

# Music Tagging
## Novel Approaches

- Transfer Learning
  - Pre-train model on source task (lot of data available)
  - Fine-tune model on target task (only little data available)

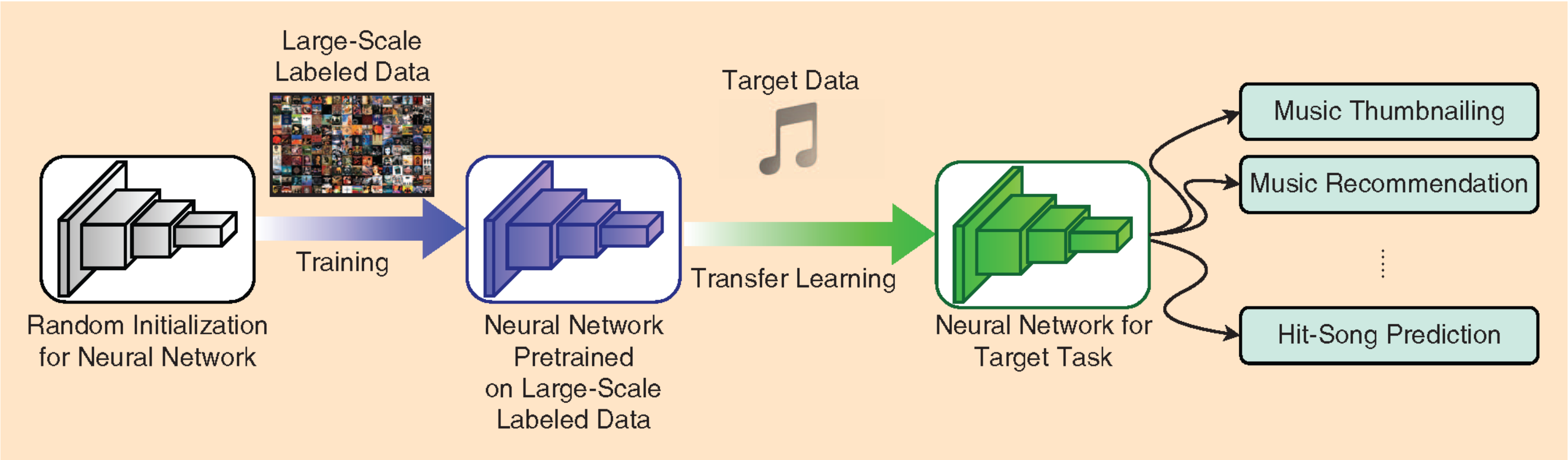



Fig. 5

- Source model (CNN) → Target model (embeddings + shallow classifier)

# Music Similarity
## Introduction

- Music → inherently multi-dimensional

# Music Similarity
## Introduction

- Music → inherently multi-dimensional
  - Example: similarity between three tracks A, B, and C

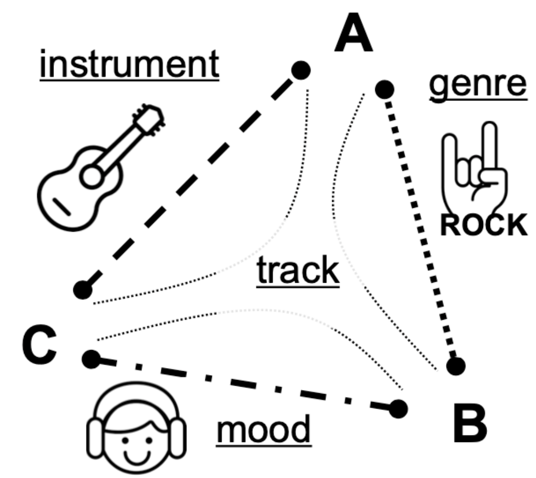



Fig. 7

# Music Similarity
## Introduction

- Music → inherently multi-dimensional
  - Example: similarity between three tracks A, B, and C

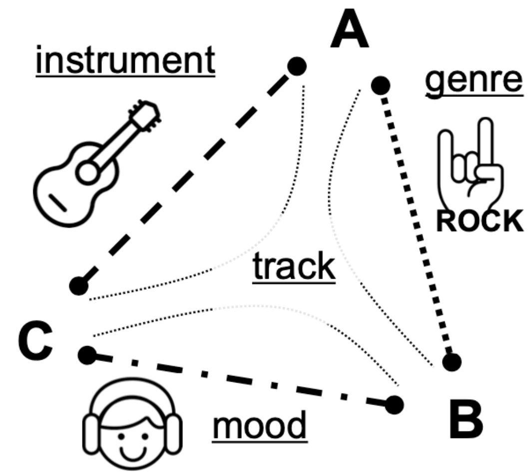



Fig. 7

- Challenge
  - Large music databases
  - Incomplete / missing metadata

# Music Similarity
## Introduction

- Music → inherently multi-dimensional
  - Example: similarity between three tracks A, B, and C
- Challenge
  - Large music databases
  - Incomplete / missing metadata
- Query by example → general retrieval approach
  - Retrieval most similar song $S$ given a query song $Q$

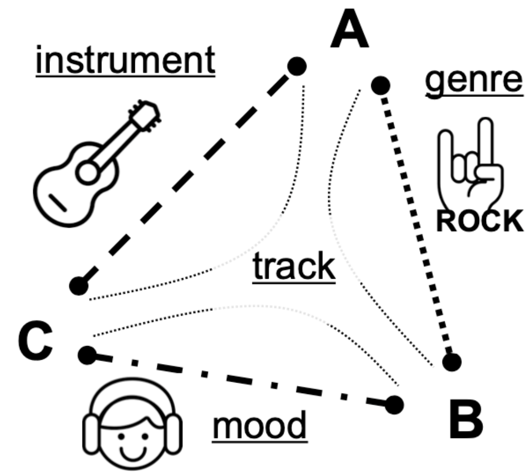



Fig. 7

# Music Similarity
## Introduction

- Retrieval tasks
    - Music fingerprinting (retrieve title, artist, e.g., Shazam app)

Fig. 8

# Music Similarity
## Introduction

- Retrieval tasks
  - Music fingerprinting (retrieve title, artist, e.g., Shazam app)
  - Cover song identification (similar text, chord progressions …)

Fig. 8

# Music Similarity
## Introduction

- Retrieval tasks
    - Music fingerprinting (retrieve title, artist, e.g., Shazam app)
    - Cover song identification (similar text, chord progressions …)
    - Music replacement (similar style, instrumentation)

Fig. 8

# Music Similarity
## Introduction

- Retrieval tasks
  - Music fingerprinting (retrieve title, artist, e.g., Shazam app)
  - Cover song identification (similar text, chord progressions …)
  - Music replacement (similar style, instrumentation)

- Specificity of different tasks

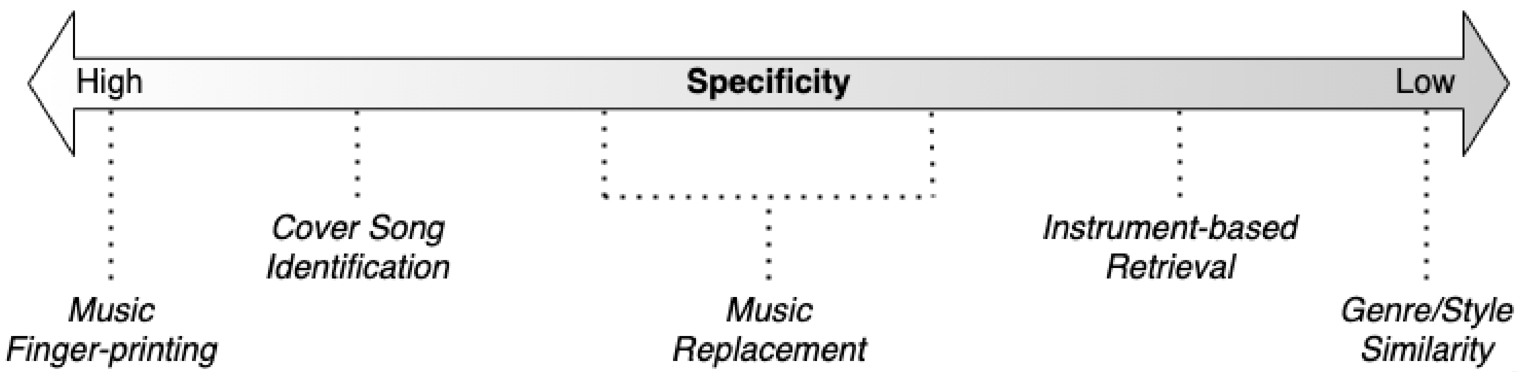



Fig. 8

# Music Similarity
## Traditional Approaches

- Different dimensions of music similarity

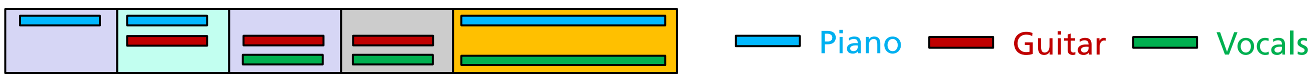
# Music Similarity
## Traditional Approaches

- Different dimensions of music similarity
  - Melodic similarity (pitch contours)

# Music Similarity
## Traditional Approaches

- Different dimensions of music similarity
  - Melodic similarity (pitch contours)

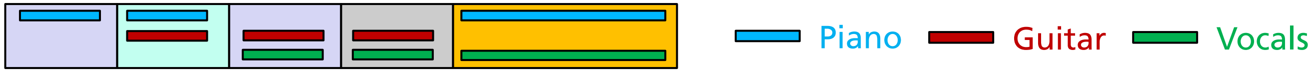
  - Timbral similarity (instrumentation)



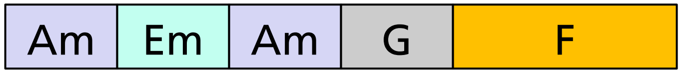
# Music Similarity
## Traditional Approaches

- Different dimensions of music similarity
  - Melodic similarity (pitch contours)

  - Timbral similarity (instrumentation)

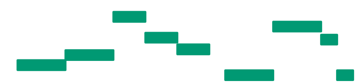



  - Structural / harmonic similarity (segments, chords)

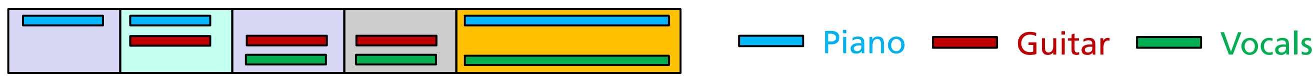


# Music Similarity
## Traditional Approaches

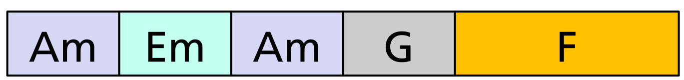
- Different dimensions of music similarity
  - Melodic similarity (pitch contours)



  - Timbral similarity (instrumentation)




  - Structural / harmonic similarity (segments, chords)




  - Rhythmic similarity (patterns)

Own

# Music Similarity
## Novel Approaches

- Metric learning
    - Model (abstract) notion of similarity between data instances

# Music Similarity
## Novel Approaches

- Metric learning
  - Model (abstract) notion of similarity between data instances
  - Pair-wise distance between feature representations

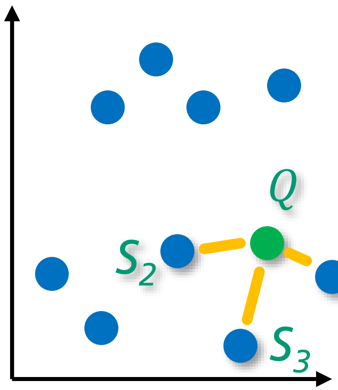
# Music Similarity
## Novel Approaches

- Metric learning
  - Model (abstract) notion of similarity between data instances
  - Pair-wise distance between feature representations
- Training
  - Proximity between similar instances
  - Distance between dissimilar instances

# Music Similarity
## Novel Approaches

- Metric learning
  - Model (abstract) notion of similarity between data instances
  - Pair-wise distance between feature representations
- Training
  - Proximity between similar instances
  - Distance between dissimilar instances
- Query $Q \rightarrow$ Ranked list of most similar instances $S$

Feature Space

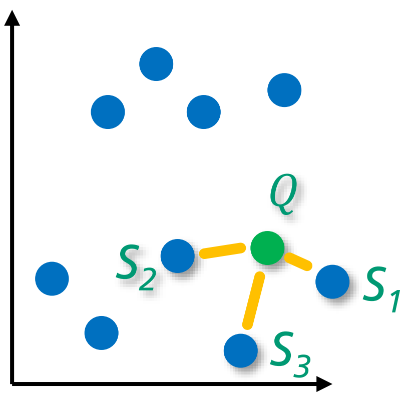



Own

# Music Similarity
## Novel Approaches

- Metric learning
  - Model (abstract) notion of similarity between data instances
  - Pair-wise distance between feature representations
- Training
  - Proximity between similar instances
  - Distance between dissimilar instances
- Query $Q \rightarrow$ Ranked list of most similar instances $S$
- Distance measures
  - Euclidean distance, Cosine distance, etc.

Feature Space




Own

# Music Similarity
## Novel Approaches

- Disentanglement learning
    - Goal → separate underlying semantic concepts (e.g., genre, instrument, mood)
        - learnt jointly
        - remain separable in the embedding space

# Music Similarity
## Novel Approaches

- Disentanglement learning
  - Goal → separate underlying semantic concepts (e.g., genre, instrument, mood)
    - learnt jointly
    - remain separable in the embedding space

- Improves
  - Music tagging (classification)
  - Music recommendation (similarity)

# Music Similarity
## Novel Approaches

- Triplet-based Training
  - Conditional Similarity Networks (CSN) [Lee, 2020]

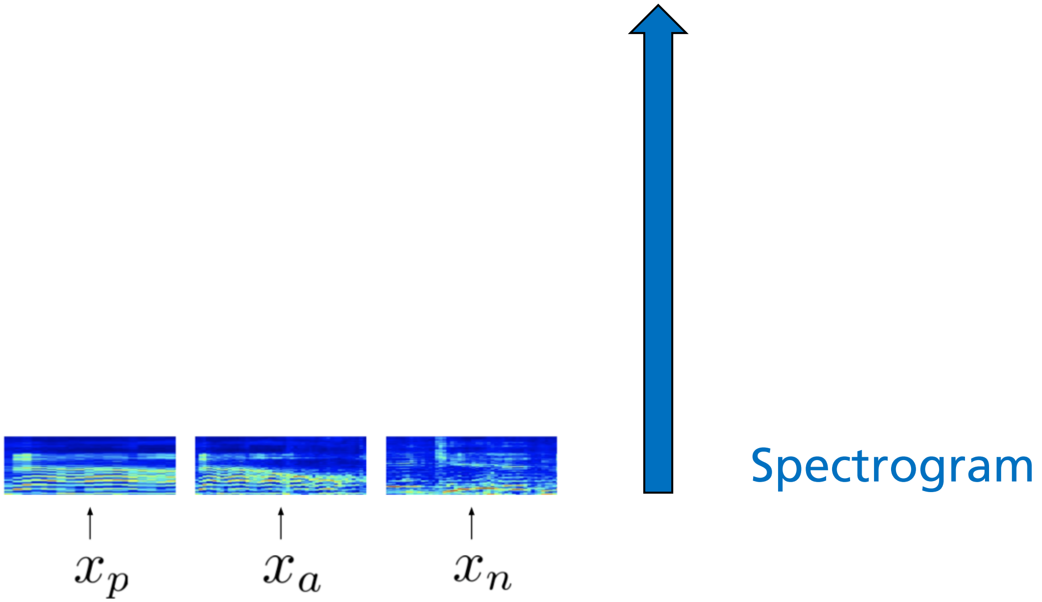



Fig. 10

# Music Similarity
## Novel Approaches

- Triplet-based Training
  - Conditional Similarity Networks (CSN) [Lee, 2020]

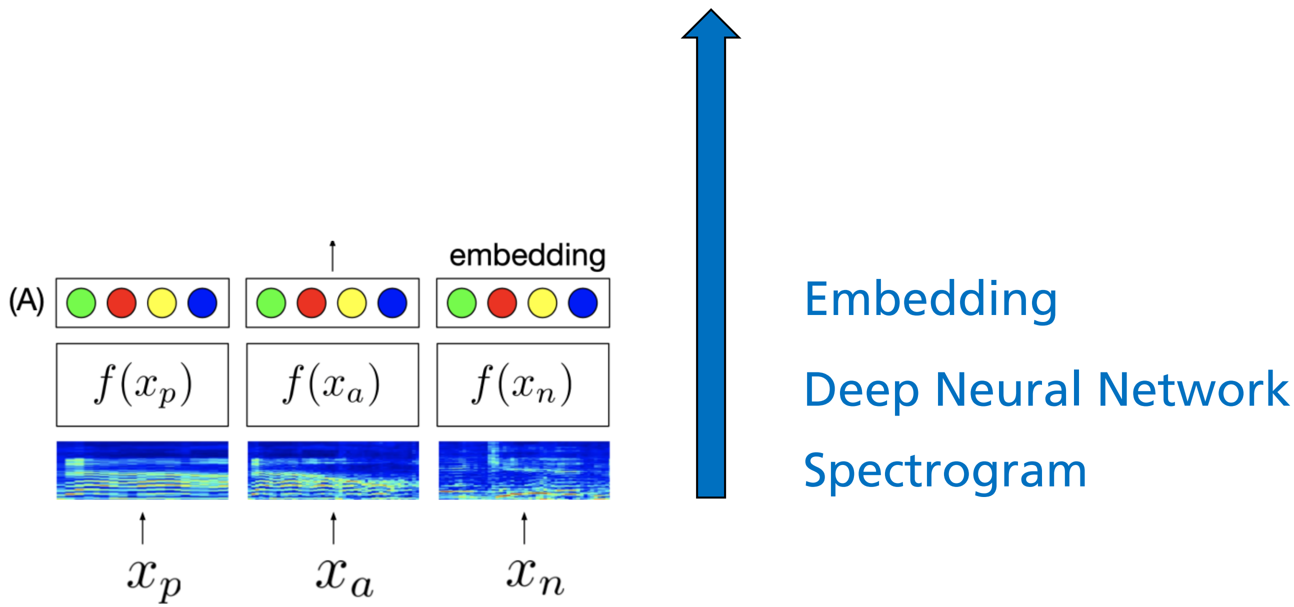
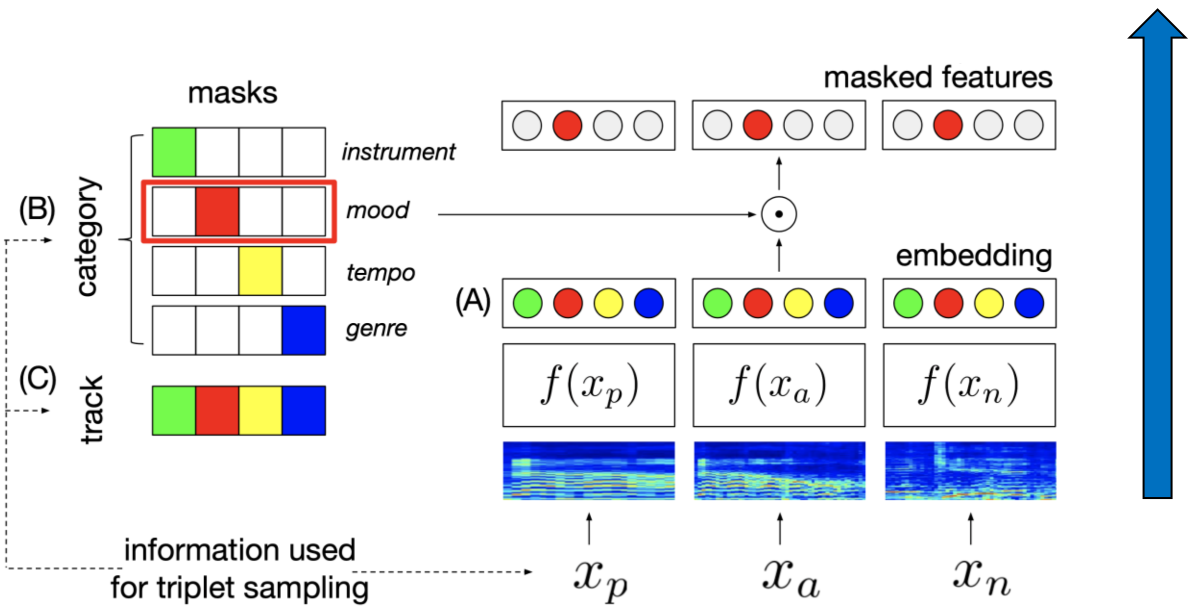



Fig. 10

# Music Similarity
## Novel Approaches

- Triplet-based Training
  - Conditional Similarity Networks (CSN) [Lee, 2020]




Embedding

Deep Neural Network

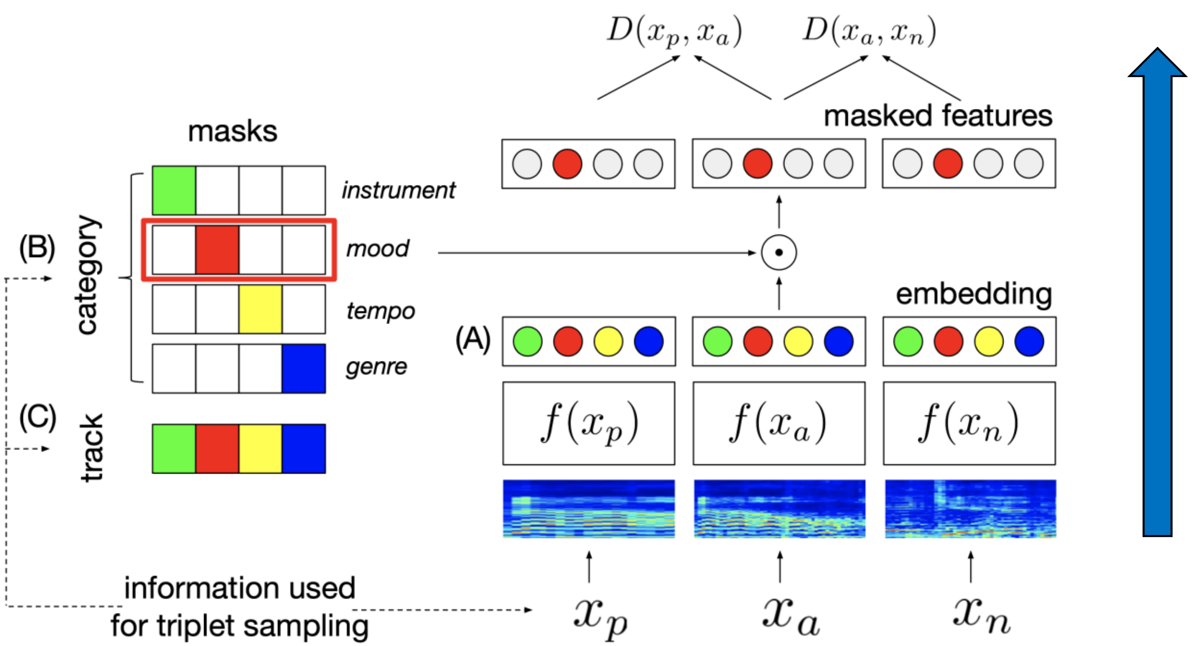
Spectrogram

Applying binary masks to embeddings

Fig. 10

# Music Similarity
## Novel Approaches

- Triplet-based Training
  - Conditional Similarity Networks (CSN) [Lee, 2020]




Pair-wise distances

Embedding

Deep Neural Network

Spectrogram

Applying binary masks to embeddings

Fig. 10

# Tempo Detection
## Introduction

- Tempo [beats / minute]
  - Frequency with which humans tap along the beat

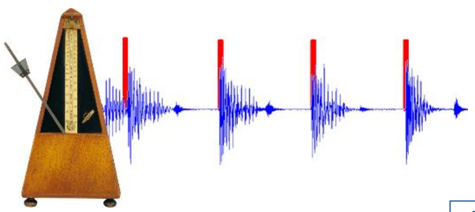


Fig. 11

# Tempo Detection
## Introduction

- Tempo [beats / minute]
  - Frequency with which humans tap along the beat

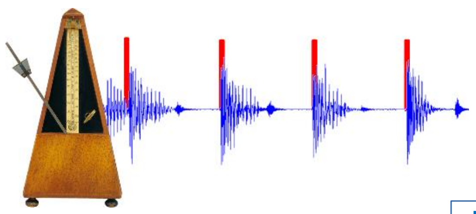


Fig. 11

- Beat tracking
  - Estimating precise beat positions

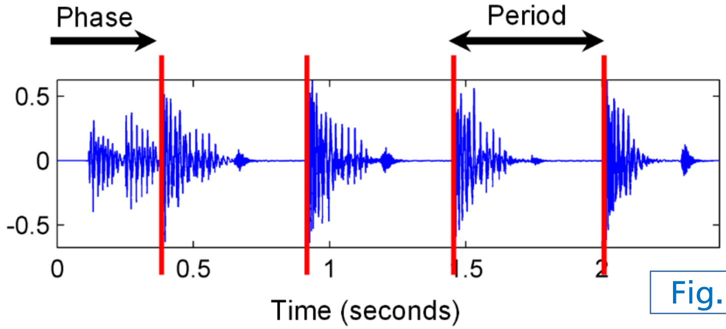



Fig. 12

# Tempo Detection
## Introduction

- Tempo [beats / minute]
  - Frequency with which humans tap along the beat

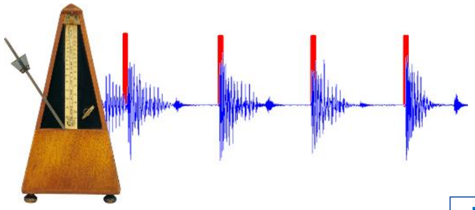


Fig. 11

- Beat tracking
  - Estimating precise beat positions

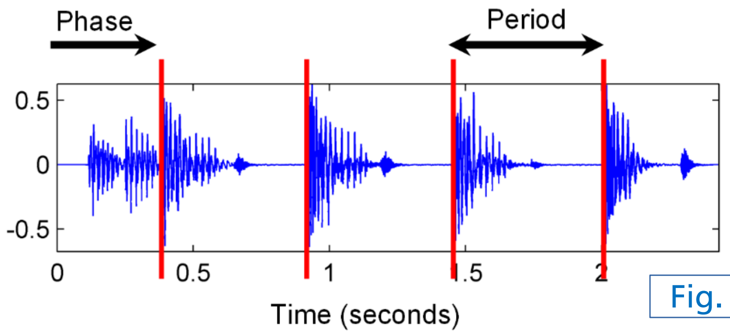
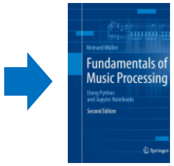


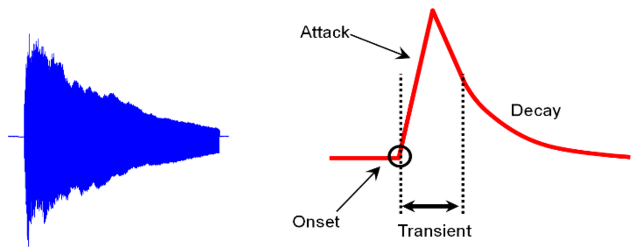

Fig. 12




FMP Notebooks

# Tempo Detection
## Introduction

- Note onsets → note beginning times




Audio samples → Note envelope

Fig. 13

# Tempo Detection
## Introduction

- Note onsets → note beginning times
  - Clearly defined for plucked string and percussion instruments
  - Ambiguous for wind & brass instruments

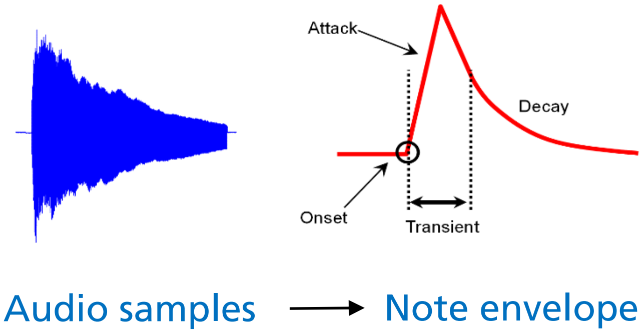



Fig. 13

# Tempo Detection
## Introduction

- Note onsets → note beginning times
    - Clearly defined for plucked string and percussion instruments
    - Ambiguous for wind & brass instruments
- Onset detection
    - Onset detection function
    - Peak picking

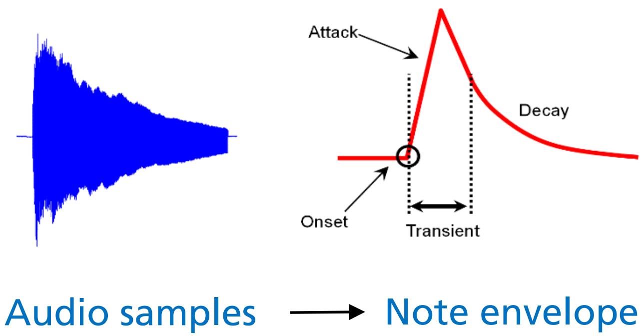
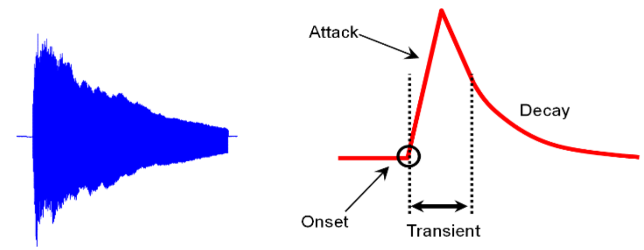



Fig. 13

# Tempo Detection

## Introduction

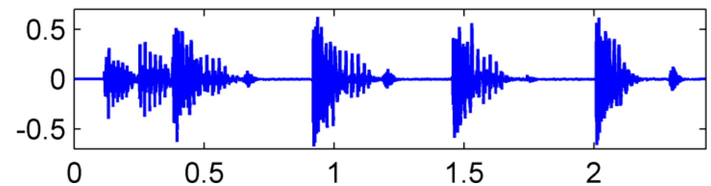
- Note onsets → note beginning times
  - Clearly defined for plucked string and percussion instruments
  - Ambiguous for wind & brass instruments
- Onset detection
  - Onset detection function
  - Peak picking




Audio samples → Note envelope

Fig. 13

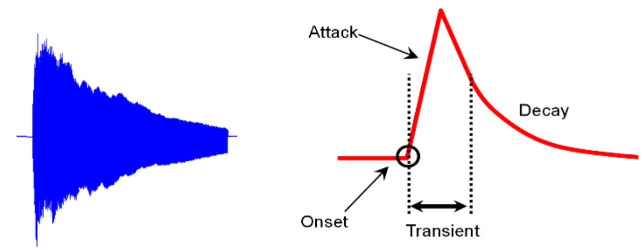
Audio samples



# Tempo Detection

## Introduction

- Note onsets → note beginning times
  - Clearly defined for plucked string and percussion instruments
  - Ambiguous for wind & brass instruments
- Onset detection
  - Onset detection function
  - Peak picking




Audio samples → Note envelope

Fig. 13

Audio samples

↓

Onset detection function & peaks

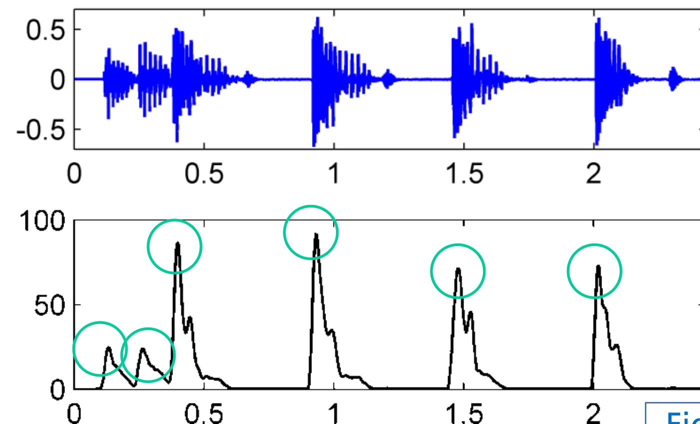



Fig. 14

# Tempo Detection
## Traditional Methods

- Predominant local pulse (PLP)

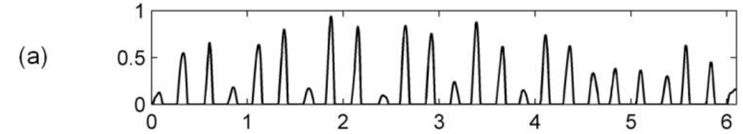



Fig. 15

# Tempo Detection
## Traditional Methods

- Predominant local pulse (PLP)
  - Correlation with local (windowed) periodic patterns

(a)

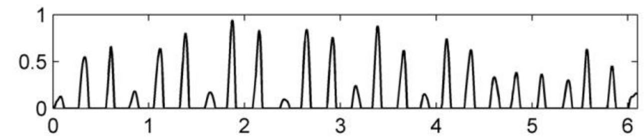



Fig. 15

# Tempo Detection
## Traditional Methods

- Predominant local pulse (PLP)
  - Correlation with local (windowed) periodic patterns
- Tempogram [Grosche & Müller, 2011]
  - Local likelihood of different tempo candidates

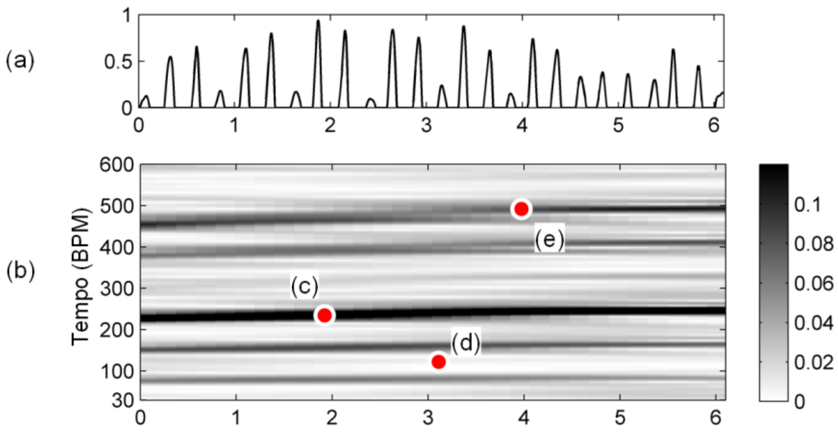
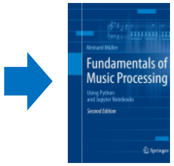



Fig. 15

# Tempo Detection
## Traditional Methods

- Predominant local pulse (PLP)
  - Correlation with local (windowed) periodic patterns
- Tempogram [Grosche & Müller, 2011]
  - Local likelihood of different tempo candidates
  - Allows to follow tempo changes (e.g., classical music)




FMP Notebooks

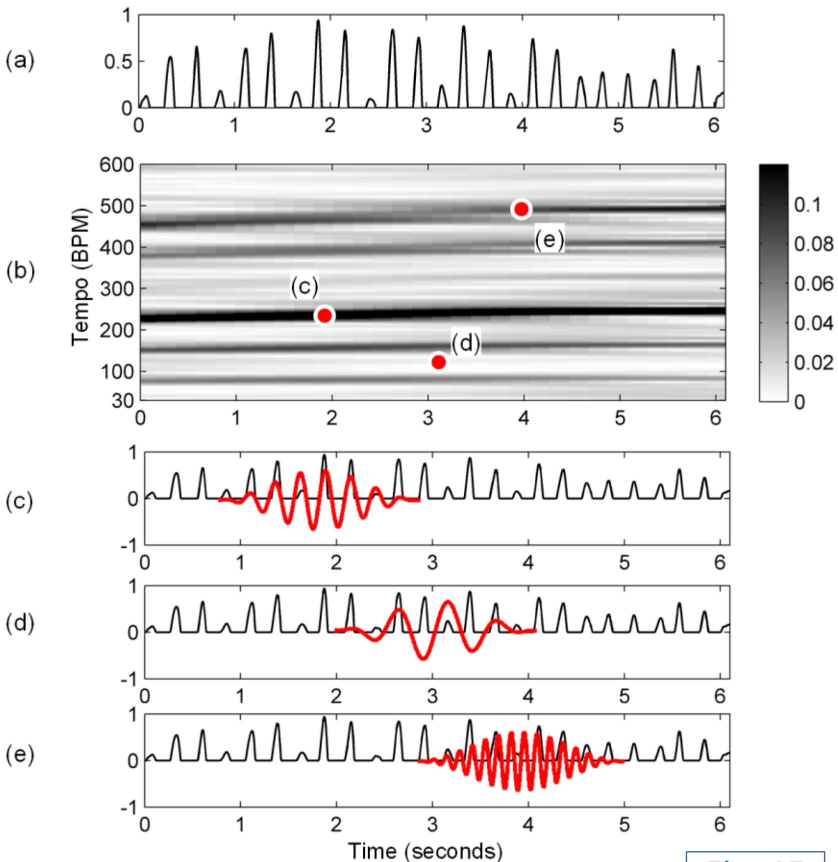



Fig. 15

# Tempo Detection
## Novel Methods

- Approach [Böck et al., 2015]

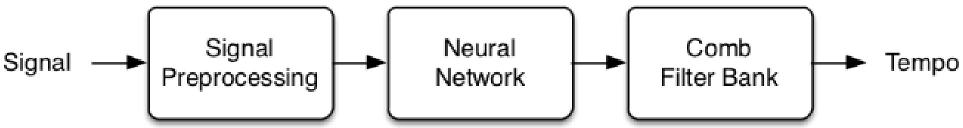
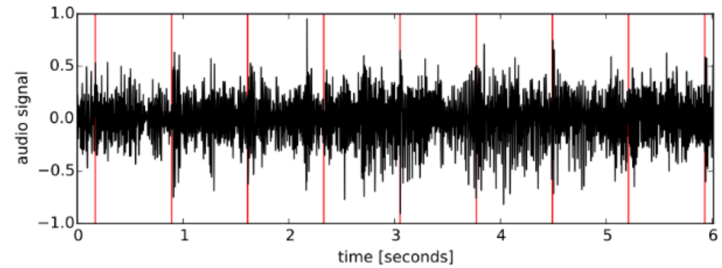



Fig. 16




(a) Input audio signal

Fig. 17

# Tempo Detection
## Novel Methods

- Approach [Böck et al., 2015]

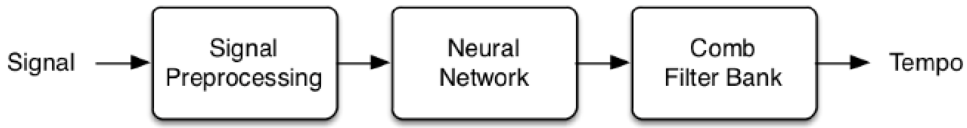
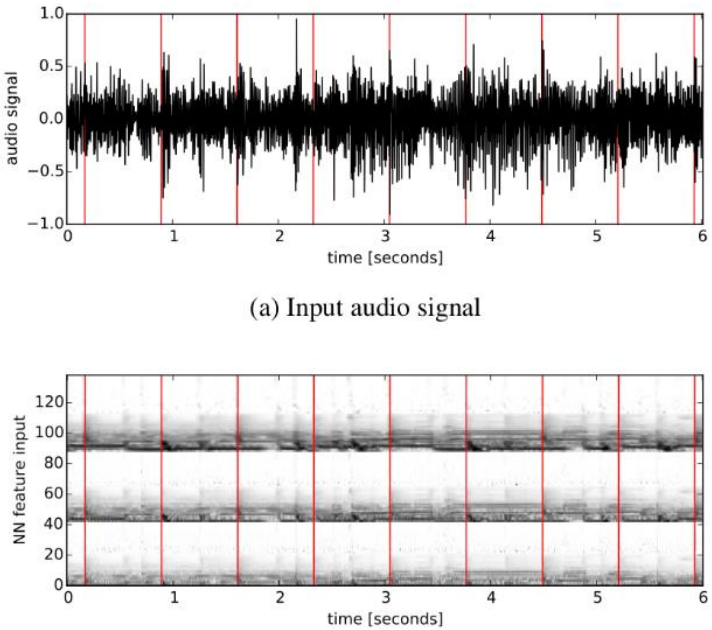



Fig. 16




(a) Input audio signal

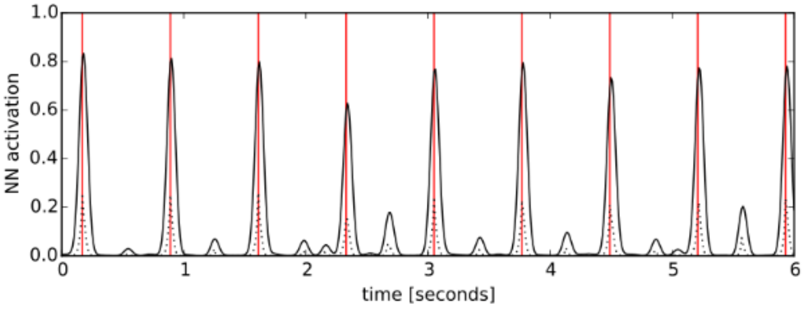
(b) Input to the neural network

Fig. 17

- Signal representation
  - Stacking of 3 STFT magnitude spectrograms (N=1024, 2048, 4096)
  - Log-amplitude & log-frequency

# Tempo Detection
## Novel Methods

- Neural Network
    - Recurrent (bi-directional LSTM) layer
    - Outputs beat activation function

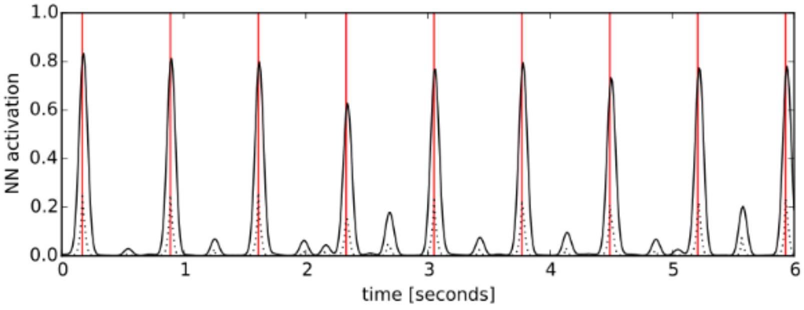



(c) Neural network output (beat activation function)

Fig. 18

# Tempo Detection
## Novel Methods

- Neural Network
  - Recurrent (bi-directional LSTM) layer
  - Outputs beat activation function
- Comb filter bank
  - Multiple comb filters → detect periodicities

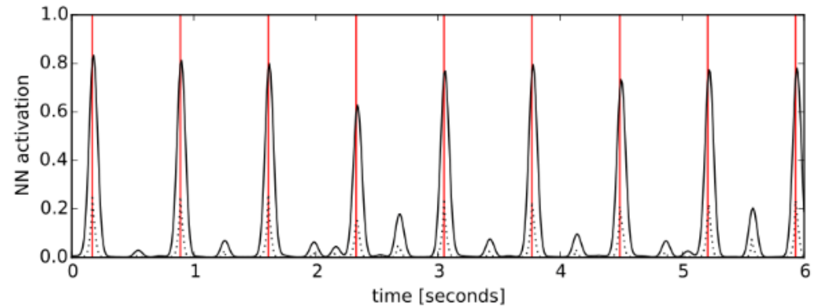



(c) Neural network output (beat activation function)

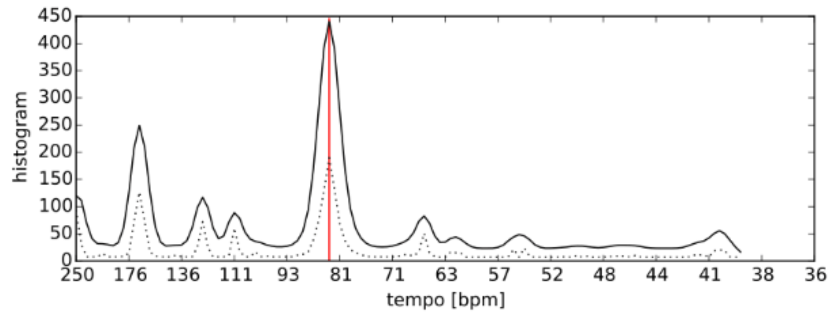
Fig. 18

# Tempo Detection
## Novel Methods

- Neural Network
  - Recurrent (bi-directional LSTM) layer
  - Outputs beat activation function
- Comb filter bank
  - Multiple comb filters → detect periodicities
- Estimate tempo from histogram maximum




(c) Neural network output (beat activation function)




(f) Weighted histogram with summed maxima

Fig. 18

# Tempo Detection
## Novel Methods

- Approach [Schreiber & Müller, 2018]
  - Sample rate ~ 11 kHz, 40-band mel spectrogram
- Main contributions
  - End-to-end tempo without intermediate novelty function

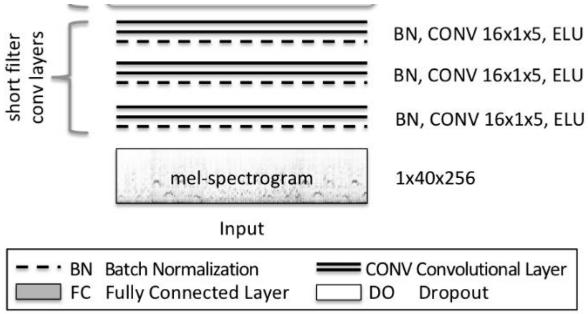



Fig. 19

# Tempo Detection
## Novel Methods

- Approach [Schreiber & Müller, 2018]
  - Sample rate ~ 11 kHz, 40-band mel spectrogram
- Main contributions
  - End-to-end tempo without intermediate novelty function
  - 4 multi-filter modules → compress along frequency & find periodicities

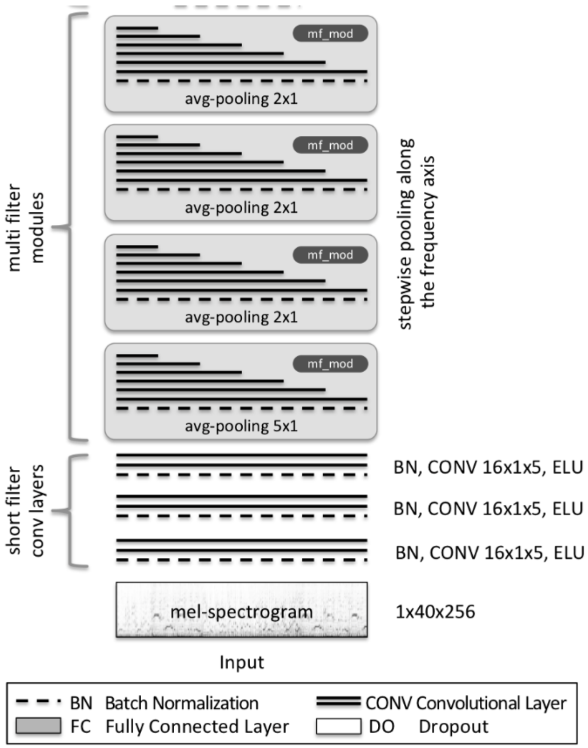



Fig. 19

# Tempo Detection
## Novel Methods

- Approach [Schreiber & Müller, 2018]
    - Sample rate ~ 11 kHz, 40-band mel spectrogram
- Main contributions
    - End-to-end tempo without intermediate novelty function
    - 4 multi-filter modules → compress along frequency & find periodicities
    - Dense layers → tempo classification
        - 256 classes: 30 – 285 bpm

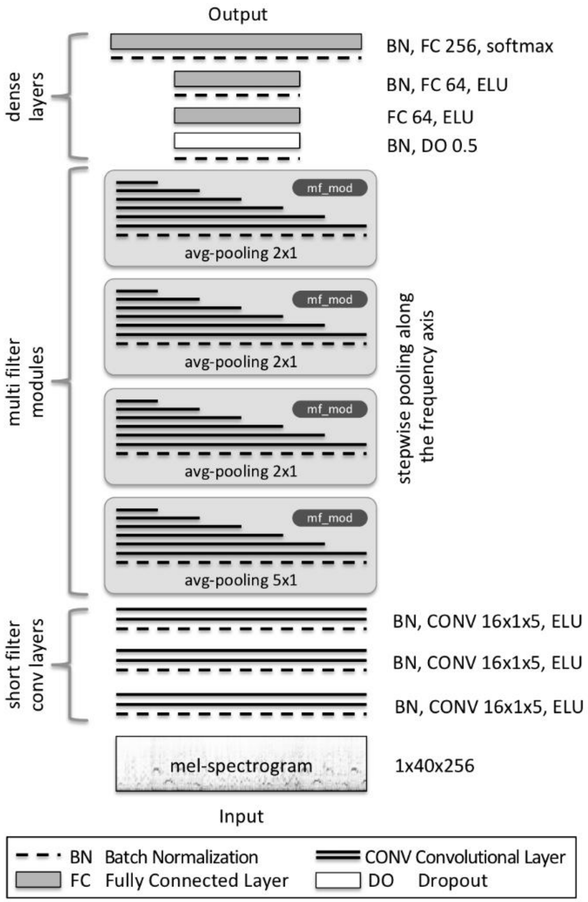



Fig. 19

# Summary

- Music Information Retrieval
- Music Tagging
- Music Similarity
- Tempo Estimation

- Main trends
    - Adapt (data-driven) deep learning methods to music domain
    - Incorporate music domain knowledge

# References


Böck, S., Krebs, F., & Widmer, G. (2015). Accurate tempo estimation based on recurrent neural networks and resonating comb filters. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 625–631.

Grosche, P., & Müller, M. (2011). Extracting Predominant Local Pulse Information From Music Recordings. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(6), 1688–1701.

Lee, J., Bryan, N. J., Salamon, J., Jin, Z., & Nam, J. (2020). Disentangled Multidimensional Metric Learning for Music Similarity. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 6–10. Barcelona, Spain.

Lee, J., Bryan, N. J., Salamon, J., Jin, Z., & Nam, J. (2020). Metric learning vs classification for disentangled music representation learning. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 439–445. Montréal, Canada.

Müller, M. (2021). *Fundamentals of Music Processing - Using Python and Jupyter Notebooks* (2nd ed.). Springer.

Nam, J., Choi, K., Lee, J., Chou, S. Y., & Yang, Y. H. (2019). Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach. *IEEE Signal Processing Magazine*, *36*(1), 41–51.

Pons, J., Nieto, O., Prockup, M., Schmidt, E., Ehrmann, A., & Serra, X. (2018). End-to-End Learning for Music Audio Tagging at Scale. *Proceedings of the International Society for Music Information Retrieval (ISMIR)2*, 637–644. Paris, France.

# References

Ribecky, S. (2021). *Disentanglement Representation Learning for Music Annotation and Music Similarity*. Master Thesis. Technische Universität Ilmenau.

Schreiber, H., & Müller, M. (2018). A Single-Step Approach to Musical Tempo Estimation using a Convolutional Neural Network. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 98–105. Paris, France.

Won, M., Chun, S., Nieto, O., & Serra, X. (2020). Data-Driven Harmonic Filters for Audio Representation Learning. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 536–540. Barcelona, Spain.

# Images

Fig. 1: https://www.synchtank.com/wp-content/uploads/2018/06/1476277072027.jpg

Fig. 2: https://miro.medium.com/max/800/1*cC1KOdyzzt1nazak42cBdg.jpeg

Fig. 3: [Nam, 2019], p. 42, Fig. 1

Fig. 4: [Won, 2020], p. 537, Fig. 1a

Fig. 5: [Nam, 2019], p. 48, Fig. 4

Fig. 6: [Pons, 2018], p. 639, Fig. 2 (top left)

Fig. 7: [Lee, 2020, ICASSP], p. 1, Fig. 1

Fig. 8: [Ribecky, 2021], p. 26, Fig. 2.11

Fig. 10: [Lee, 2020, ICASSP], p. 2, Fig. 2

Fig. 11: [Müller, 2021], p. 309, chapter 6 (cover image)

Fig. 12: [Müller, 2021], p, 310, Fig. 6.1(b)

Fig. 13: [Müller, 2021], p. 311, Fig. 6.2

Fig. 14: [Müller, 2021], p. 313, Fig. 6.3(a)&(b)

# Images

Fig. 15: [Grosche & Müller, 2009], p. 2, Fig. 1(e-g) & p. 3, Fig. 2 (a)

Fig. 16: [Böck et al., 2015], p. 2, Fig. 1

Fig. 17: [Böck et al., 2015], p. 3, Fig. 2 (a) & (b)

Fig. 18: [Böck et al., 2015], p. 3, Fig. 2 (c) & (f)

Fig. 19: [Schreiber & Müller, 2018], p. 3, Fig. 2

# Sounds

AUD-1: Mr Smith – Black Top (2021), https://freemusicarchive.org/music/mr-smith/studio-city/black-top

AUD-2: Crowander – Humbug (2021), https://freemusicarchive.org/music/crowander/from-the-piano-solo-piano/humbug

AUD-3: Bumy Goldson: Keep Walking (2021), https://freemusicarchive.org/music/bumy-goldson/parlor/keep-walking

AUD-4: Cloudjumper: Mocking the god (2016), https://freemusicarchive.org/music/Cloudjumper/Memories_of_Snow/05_Cloudjumper_-_Mocking_the_gods

# Thank you!

- Any questions?

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

Jakob.abesser@idmt.fraunhofer.de

https://www.machinelistening.de