Computational Analysis of Sound and Music

Music Information Retrieval – Music Tagging & Similarity

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

jakob.abesser@idmt.fraunhofer.de

Music Tagging & Similarity

Outline

- **Music Tagging**
- Music Similarity

Fig-M5-6

Music Tagging & Similarity

Motivation

- Musical Instrument

Aud-M6-1 Aud-M6-2

- Musical Genre

Aud-M6-3 Aud-M6-4

Fig-M5-6

Music Tagging & Similarity

Motivation

- 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

Fig-M5-6

Music Tagging

Task

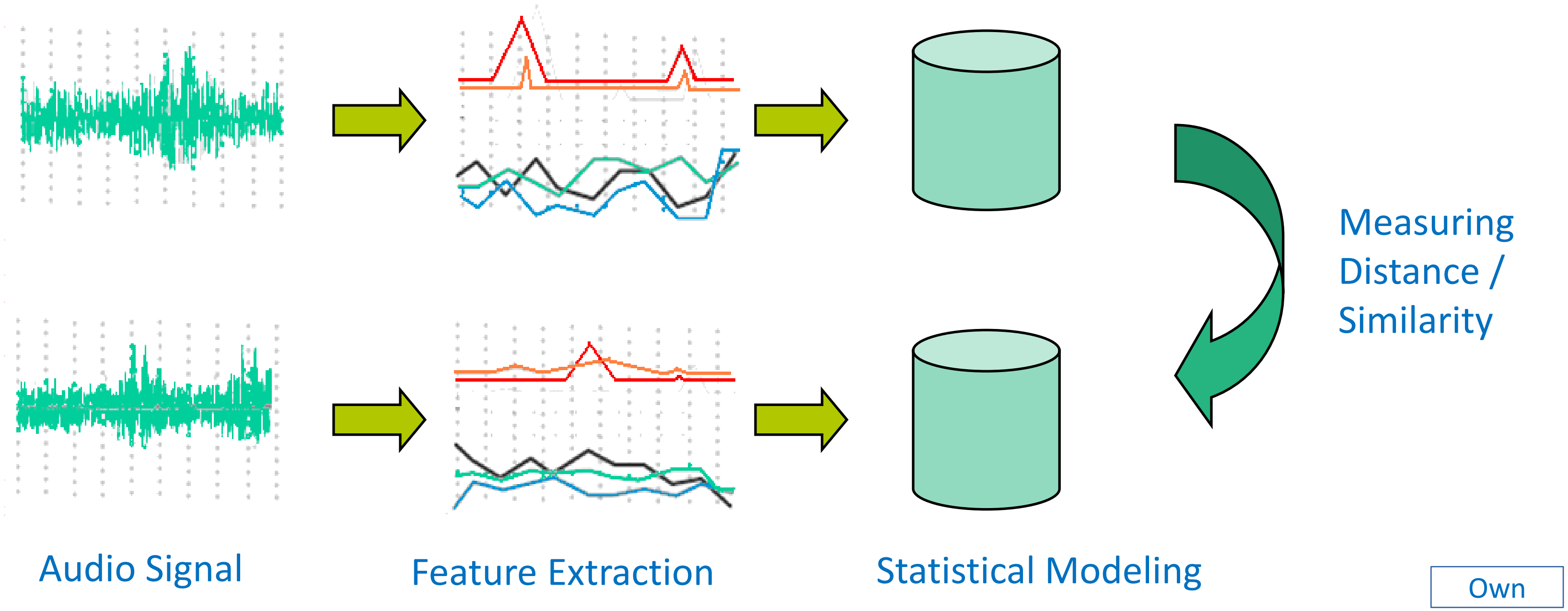
- Tags
  - Textual (objective / subjective) annotations of songs
  - Examples
    - Instruments (drums, bass, guitar, vocals …)
    - Genre (classical, electro, hip hop)
    - Mood (mellow, romantic, angry, happy)
    - Miscellaneous (noise, loud, ambient)
- Challenge
  - Music pieces change their characteristics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

Fig-M5-6

Music Tagging

Traditional Approach

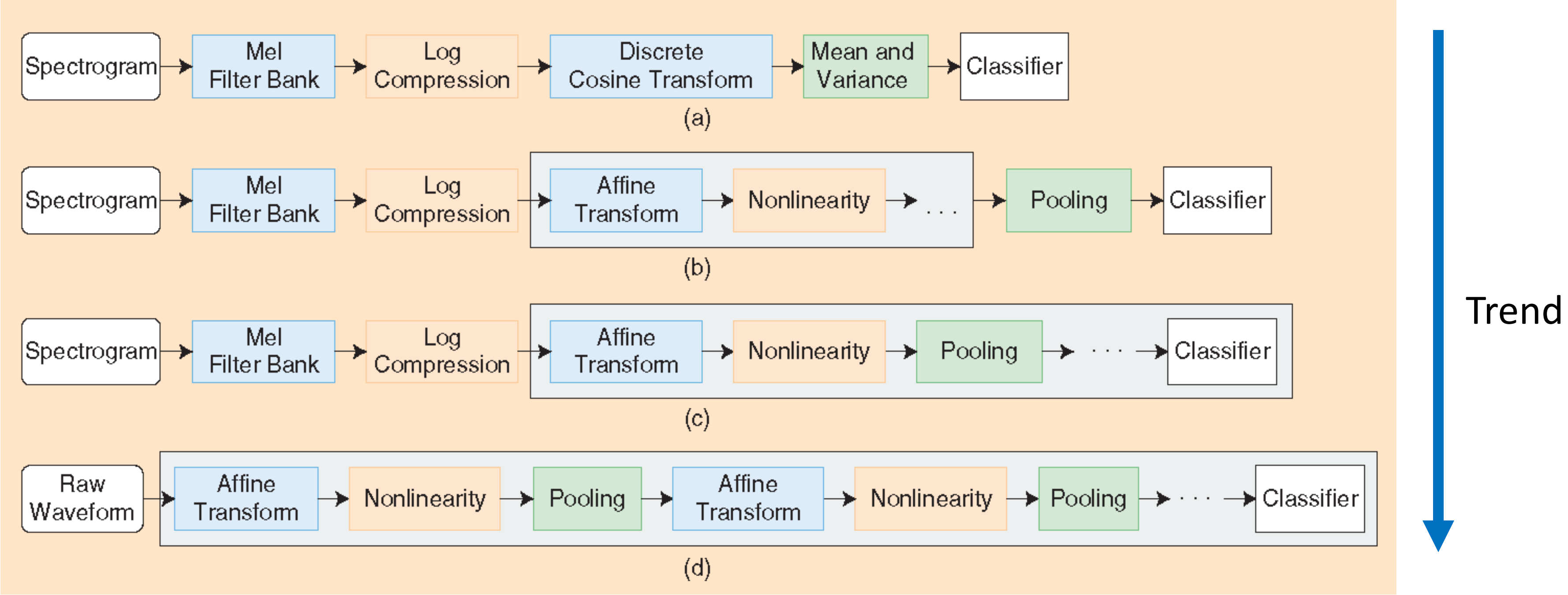
- Audio feature engineering & music domain knowledge
- Standard classification methods (GMM, SVM, kNN)



Own

Music Tagging

DL-based Approach



(a) Feature engineering (MFCC)

(b) Low-level feature

(c) Joint feature learning & classification (CNN)

(d) End-to-end learning

Fig-M6-1

Music Tagging

DL-based Approach

- 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)

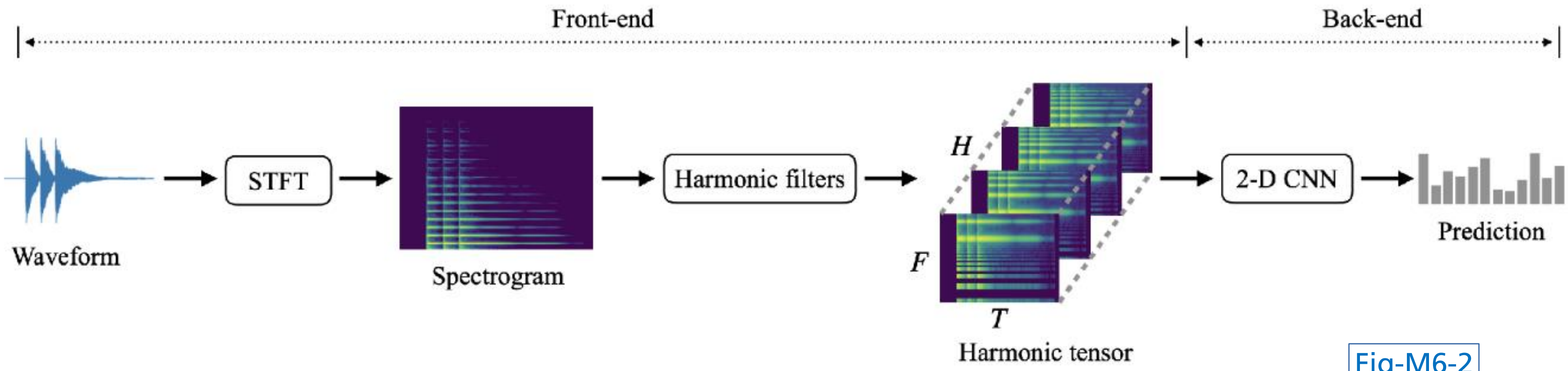


Fig-M6-2

Music Tagging

DL-based Approach

- 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

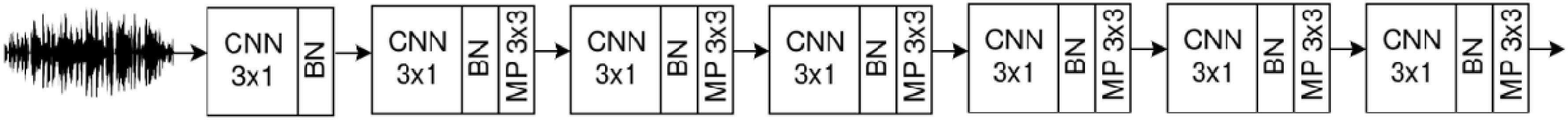


Fig-M6-3

Music Tagging

DL-based Approach

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

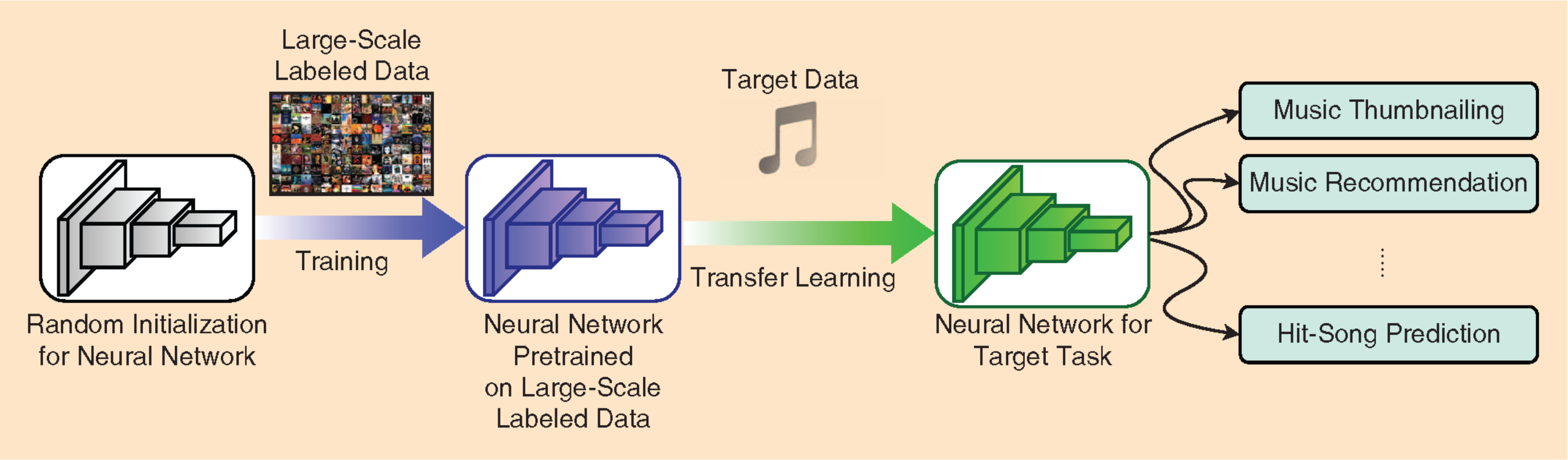


Fig-M6-4

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

Music Similarity

Task

- 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

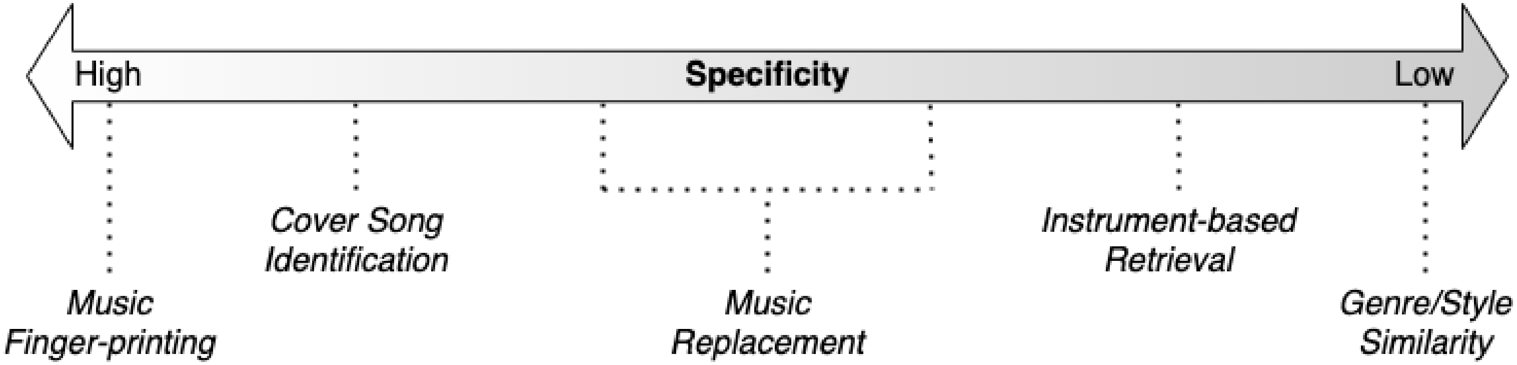


Fig-M6-5

Music Similarity

Task

- 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

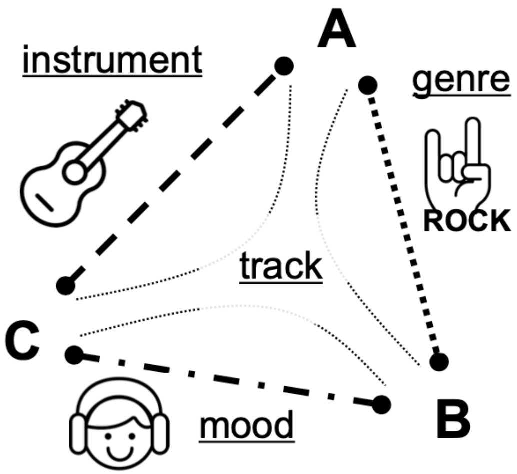
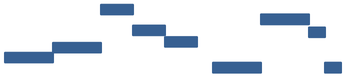


Fig-M6-6

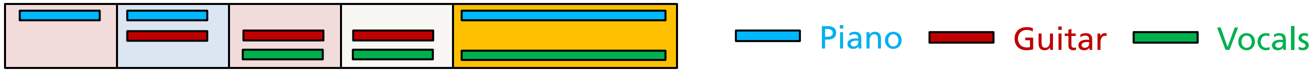
Music Similarity

Traditional Approach

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



  - Timbral similarity (instrumentation)

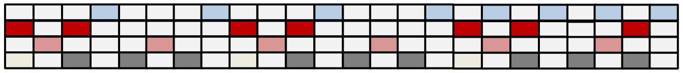


  - Structural / harmonic similarity (segments, chords)

Am	Em	Am	G	F

Fig-M6-6

  - Rhythmic similarity (patterns)



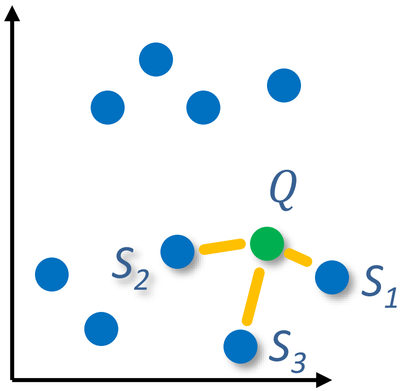
Own

Music Similarity

Traditional Approach

- 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 "→" Ranked list of most similar instances S
- Distance measures
  - Euclidean distance, Cosine distance, etc.

Feature Space



Own

Music Similarity

Traditional Approach

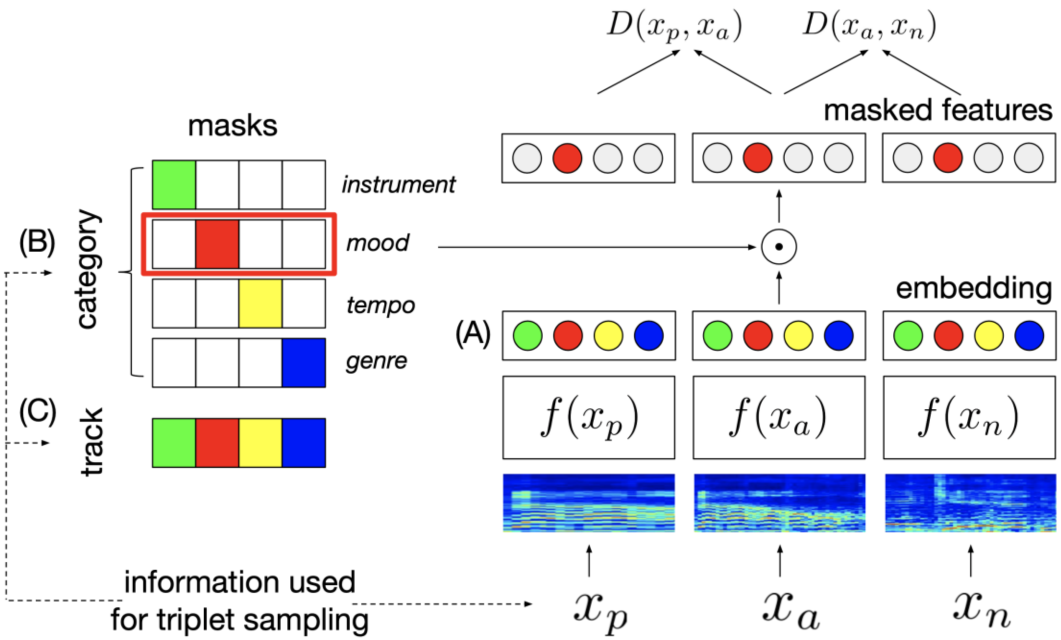
- Disentanglement learning
    - Goal → separate underlying semantic concepts (e.g., genre, instrument, mood)
        - learnt jointly
        - remain separable in the embedding space

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

Music Similarity

Traditional Approach

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



Pair-wise distances

Embedding

Deep Neural Network

Spectrogram

Fig-M6-7

Applying binary masks to embeddings

Programming session

Fig-A2-13

References

Images

Fig-M6-1: [Nam, 2019], p. 42, Fig. 1

Fig-M6-2: [Won, 2020], p. 537, Fig. 1a

Fig-M6-3: [Pons, 2018], p. 639, Fig. 2 (top left)

Fig-M6-4: [Nam, 2019], p. 48, Fig. 4

Fig-M6-5: [Ribecky, 2021], p. 26, Fig. 2.11

Fig-M6-6: [Lee, 2020, ICASSP], p. 1, Fig. 1

Fig-M6-7: [Lee, 2020, ICASSP], p. 2, Fig. 2

References

Audio

Aud-M6-1: Mr Smith – Black Top (2021), https://freemusicarchive.org/music/mr-smith/studio-city/black-top

Aud-M6-2: Crowander – Humbug (2021), https://freemusicarchive.org/music/crowander/from-the-piano-solo-piano/humbug

Aud-M6-3: Bumy Goldson: Keep Walking (2021), https://freemusicarchive.org/music/bumy-goldson/parlor/keep-walking

Aud-M6-4: Cloudjumper: Mocking the god (2016), https://freemusicarchive.org/music/Cloudjumper/Memories_of_Snow/05_Cloudjumper_-_Mocking_the_gods

References

References

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.

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.

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.

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

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.