github.com/alexanderlerch/2025-AI4Opt-Lec

music information retrieval

AI4Opt

alexander lerch

Georgia Tech Center for Music Technology
College of Design

about

about me

Georgia Tech Center for Music Technology
College of Design

- **education**
  - Electrical Engineering (Technical University Berlin)
  - Tonmeister (music production, University of Arts Berlin)
- **professional**
  - Associate Professor at the School of Music, Georgia Institute of Technology
  - 2000-2013: CEO at zplane.development
- **background**
  - audio algorithm design (20+ years)
  - commercial music software development (10+ years)
  - entrepreneurship (10+ years)

www.linkedin.com/in/lerch

introduction

content in audio signals

examples for audio signal content

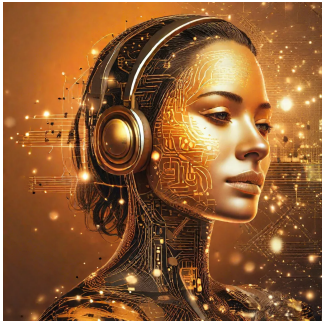
- **speech**
  - text information
  - speaker
  - recording environment
  - ...
- **music**
  - melody
  - harmony
  - structure
  - instruments
  - mood
  - genre
  - ...

matlab source: plotWaveform.m

introduction

audio content analysis — research fields

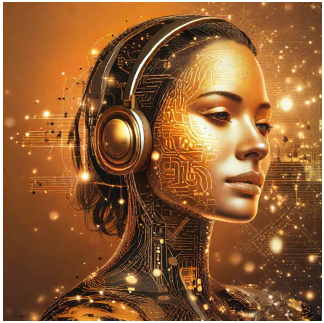
- **speech** analysis
  - speech recognition
  - speech emotion recognition, ...
- **urban sound** analysis
  - noise pollution monitoring
  - audio surveillance, ...
- **industrial sound** analysis
  - monitoring the state of mechanical devices
  - monitoring the health of livestock, ...
- **musical audio** analysis
  - music transcription
  - music classification, ...



introduction

audio content analysis — research fields

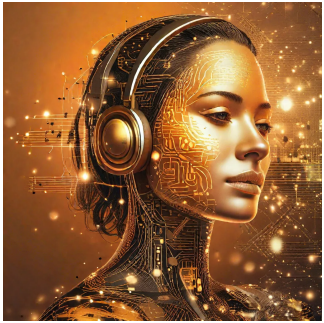
- **speech** analysis
  - speech recognition
  - speech emotion recognition, ...
- **urban sound** analysis
  - noise pollution monitoring
  - audio surveillance, ...
- **industrial sound** analysis
  - monitoring the state of mechanical devices
  - monitoring the health of livestock, ...
- **musical audio** analysis
  - music transcription
  - music classification, ...



introduction

audio content analysis — research fields

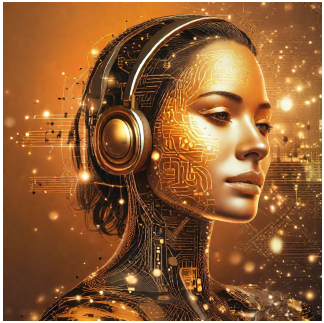
- **speech** analysis
  - speech recognition
  - speech emotion recognition, ...
- **urban sound** analysis
  - noise pollution monitoring
  - audio surveillance, ...
- **industrial sound** analysis
  - monitoring the state of mechanical devices
  - monitoring the health of livestock, ...
- **musical audio** analysis
  - music transcription
  - music classification, ...



introduction

audio content analysis — research fields

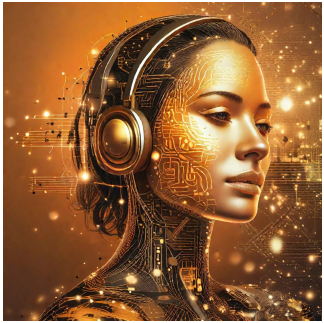
- **speech** analysis
  - speech recognition
  - speech emotion recognition, ...
- **urban sound** analysis
  - noise pollution monitoring
  - audio surveillance, ...
- **industrial sound** analysis
  - monitoring the state of mechanical devices
  - monitoring the health of livestock, ...
- **musical audio** analysis
  - music transcription
  - music classification, ...



introduction

audio content analysis — research fields

- **speech** analysis
  - speech recognition
  - speech emotion recognition, ...
- **urban sound** analysis
  - noise pollution monitoring
  - audio surveillance, ...
- **industrial sound** analysis
  - monitoring the state of mechanical devices
  - monitoring the health of livestock, ...
- **musical audio** analysis
  - music transcription
  - music classification, ...



introduction

musical audio vs. other audio

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

musical audio vs. other audio

Georgia Tech Center for Music Technology
College of Design

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

musical audio vs. other audio

Georgia Tech Center for Music Technology
College of Design

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

musical audio vs. other audio

Georgia Tech Center for Music Technology
College of Design

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

musical audio vs. other audio

Georgia Tech Center for Music Technology
College of Design

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

musical audio vs. other audio

music . . .

- is a **wide band** signal(*unlike many other audio signals*)
- comprises both **tonal and noise** components(*like most audio signals*)
- combines **multiple sound sources**(*unlike speech, like urban sound*)
- is a **poly-timbral** mixture(*nlike industrial sound*)
- sources are **harmonically related and synchronous** (*unlike other multi-source signals*)
- has a highly structured **sequential** language that is **abstract** (*unlike speech*)

introduction

use cases

Georgia Tech Center for Music Technology College of Design

- **music browsing and music discovery**
  - search & retrieval, recommendation, similarity, interfaces (e.g., QBH)
- **music consumption**
  - creative music listening
- **music production**
  - adaptive parametrization, enhancements of creative process
- **music education**
  - musically intelligent software tutoring
- **generative music**
  - interactive soundtracks (games, video)

introduction

use cases

- **music browsing and music discovery**
  - search & retrieval, recommendation, similarity, interfaces (e.g., QBH)
- **music consumption**
  - creative music listening
- **music production**
  - adaptive parametrization, enhancements of creative process
- **music education**
  - musically intelligent software tutoring
- **generative music**
  - interactive soundtracks (games, video)

introduction

use cases

- **music browsing and music discovery**
  - search & retrieval, recommendation, similarity, interfaces (e.g., QBH)
- **music consumption**
  - creative music listening
- **music production**
  - adaptive parametrization, enhancements of creative process
- **music education**
  - musically intelligent software tutoring
- **generative music**
  - interactive soundtracks (games, video)

introduction

use cases

- **music browsing and music discovery**
  - search & retrieval, recommendation, similarity, interfaces (e.g., QBH)
- **music consumption**
  - creative music listening
- **music production**
  - adaptive parametrization, enhancements of creative process
- **music education**
  - musically intelligent software tutoring
- **generative music**
  - interactive soundtracks (games, video)

introduction

use cases

- **music browsing and music discovery**
  - search & retrieval, recommendation, similarity, interfaces (e.g., QBH)
- **music consumption**
  - creative music listening
- **music production**
  - adaptive parametrization, enhancements of creative process
- **music education**
  - musically intelligent software tutoring
- **generative music**
  - interactive soundtracks (games, video)

audio analysis

audio classification — tasks

- audio classification: one of the earliest and seminal tasks in Music Information Retrieval (MIR)
- includes, e.g.,
  - music/speech classification
  - genre classification
  - musical instrument recognition
  - mood recognition
  - music auto-tagging
  - artist classification
  - ...
- non-music related
  - speaker detection
  - audio event detection
  - ...

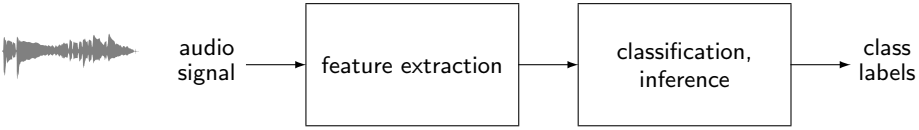
audio analysis

audio classification — tasks

- audio classification: one of the earliest and seminal tasks in Music Information Retrieval (MIR)
- includes, e.g.,
  - music/speech classification
  - genre classification
  - musical instrument recognition
  - mood recognition
  - music auto-tagging
  - artist classification
  - ...
- non-music related
  - speaker detection
  - audio event detection
  - ...

audio analysis

audio classification — traditional



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

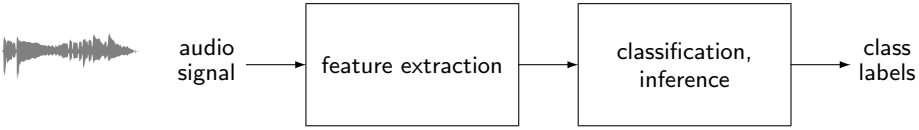
classification

- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf.

audio analysis

audio classification — traditional



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

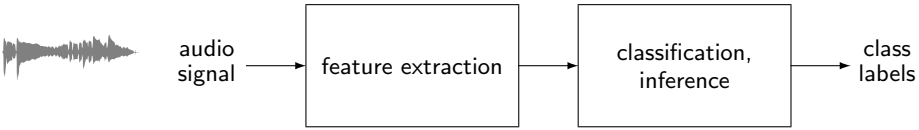
classification

- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf.

audio analysis

audio classification — traditional



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

classification

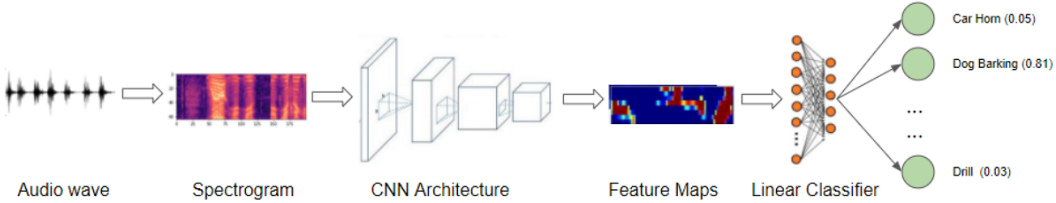
- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf.

introduction

neural network based approaches

- no custom-designed features anymore
- learn features from basic inputs (like spectrograms)



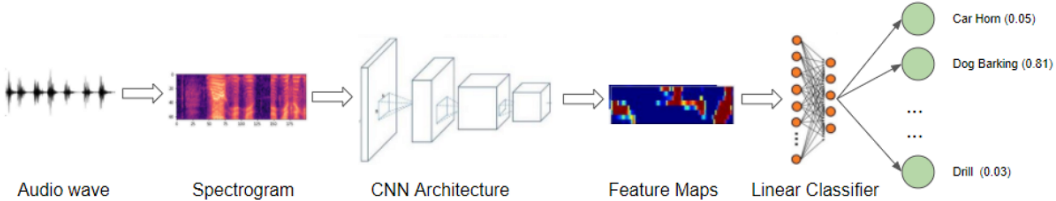
- less required expert-knowledge, more complex systems
- less expert-tweaking, more rigorous experimental requirement
- much **higher data requirements**

Fig.: towardsdatascience.com

introduction

neural network based approaches

- no custom-designed features anymore
- learn features from basic inputs (like spectrograms)



- less required expert-knowledge, more complex systems
- less expert-tweaking, more rigorous experimental requirement
- much **higher data requirements**

Fig.: towardsdatascience.com

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - *subjectivity* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *diversity & representativeness*
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

$\Rightarrow$ model **success largely depends on training data**

- **general challenges** concerning data
  - *subjectivity* of annotations
  - *noisiness* (bad quality, bad annotations, ...)
  - *imbalance & bias* (distribution is skewed, biased)
  - *diversity & representativeness*
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - *subjectivity* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *diversity & representativeness*
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - *subjectivity* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *diversity & representativeness*
  - **amount**

https://imgs.xkcd.com/comics/machine_learning.png

data
insufficient data

- **music data** itself is not scarce (although there might be copyright issues...)
- **consumer annotations** are more difficult to collect, but there are some large collections
- **detailed musical annotations** are hard to come by, because
  - time consuming & tedious annotation process
  - experts needed for annotations

data
insufficient data

- **music data** itself is not scarce (although there might be copyright issues...)
- **consumer annotations** are more difficult to collect, but there are some large collections
- **detailed musical annotations** are hard to come by, because
  - time consuming & tedious annotation process
  - experts needed for annotations

data

insufficient data

- **music data** itself is not scarce (although there might be copyright issues...)
- **consumer annotations** are more difficult to collect, but there are some large collections
- **detailed musical annotations** are hard to come by, because
  - time consuming & tedious annotation process
  - experts needed for annotations

data

previous work on insufficient data

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning
  - semi- and self-supervised approaches
  - ...

data

previous work on insufficient data

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning[2]
  - semi- and self-supervised approaches
  - ...

[2] S. Gururani *et al.*, "An Attention Mechanism for Music Instrument Recognition," in *ISMIR*, Delft, 2019.

data

previous work on insufficient data

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning
  - semi- and self-supervised approaches[2][3]
  - ...

[2] C.-W. Wu and A. Lerch, "Automatic drum transcription using the student-teacher learning paradigm with unlabeled music data," in *ISMIR*, Suzhou, 2017.

[3] S. Gururani and A. Lerch, "Semi-Supervised Audio Classification with Partially Labeled Data," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, online: Institute of Electrical and Electronics Engineers (IEEE), 2021. [Online]. Available: https://arxiv.org/abs/2111.12761.

reprogramming
introduction

Georgia Tech Center for Music Technology
College of Design

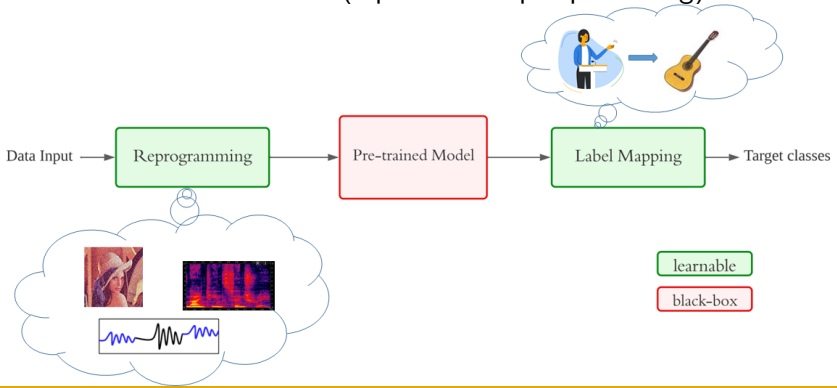
- **observation**
  - pre-trained deep models can be very powerful if trained with sufficient data, even for different tasks
- **idea**
  - re-using pre-trained models for a new task **without** re-training
- **goals**
  - keep number of training parameters minimal
  - utilize unmodified network trained on different task

reprogramming
introduction

- **observation**
  - pre-trained deep models can be very powerful if trained with sufficient data, even for different tasks
- **idea**
  - re-using pre-trained models for a new task **without** re-training
- **goals**
  - keep number of training parameters minimal
  - utilize unmodified network trained on different task

reprogramming
overview

- inspired by
  - transfer learning
  - adversarial learning
- allows for small trainable model (input and output processing)



reprogramming

experimental setup: baselines

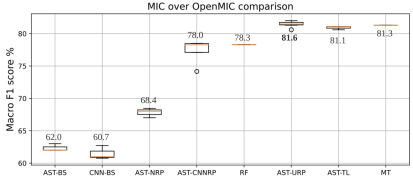
- baseline AST:
  - good performance on audio event classification[4]
- data
  - OpenMic (instrument classification):
    - 20 classes of musical instruments
    - 10 s audio snippets (20000)
- ablation study:
  - CNN only
  - U-Net only
  - CNN + AST + FC
  - U-Net + AST + FC

[4] Y. Gong *et al.*, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech*, arXiv: 2104.01778, Brno, Czechia, Jul. 2021. Acce
Apr. 17, 2022. [Online]. Available: http://arxiv.org/abs/2104.01778.

reprogramming
results: classification metrics

Georgia Tech Center for Music Technology
College of Design

method	F1 (macro)	train. param. (M)
AST + simple output mapping	62.03	0.001
CNN	60.77	0.017
U-Net	62.73	0.017
CNN + AST + FC	78.08	0.017
U-Net + AST + FC	**81.60**	0.018



- a powerful model trained on a different task cannot easily be used directly
- proper input and output processing can significantly improve performance
- *re-programming can beat the state-of-the-art* at a fraction of trainable parameters (at least factor 10)

[5] H.-H. Chen and A. Lerch, "Music Instrument Classification Reprogrammed," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*, Bergen, Norway, 2023. [Online]. Available: https://arxiv.org/abs/2211.08379.

embeddings as teachers

introduction

Georgia Tech Center for Music Technology College of Design

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

embeddings as teachers

introduction

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

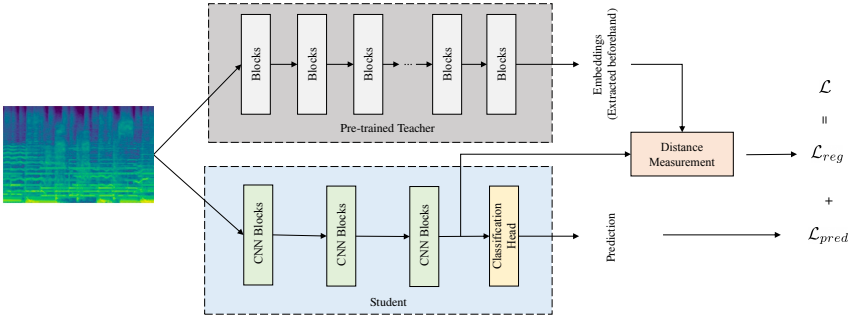
embeddings as teachers

introduction

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

embeddings as teachers

method overview



- **transfer learning**
  - use embeddings from a different task for the target task
- **knowledge distillation**
  - use a teacher to train a less complex student on the same task

embeddings as teachers

experimental setup

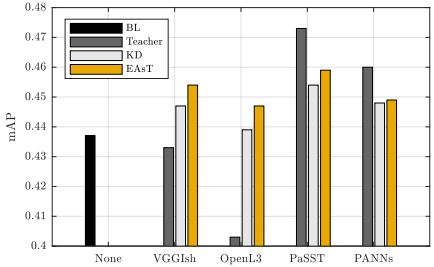
- task: auto-tagging
  - MagnaTagATune (MTAT) dataset:
    - 50 music tags
    - 30 s audio snippets ($\approx$ 21000)

- systems:
  - baseline: student without teacher
  - teacher: embedding plus logistic regression
    - VGGish
    - OpenL3
    - PaSST
    - PANNs
  - KD: student trained with soft targets from teacher
  - EasT: student regularized with teacher embeddings

embeddings as teachers

results

- student model consistently outperforms baseline
- student model consistently outperforms knowledge distillation
- student model outperforms teacher for "old" embeddings
- modern embeddings are powerful but complex



[6] Y. Ding and A. Lerch, "Audio Embeddings as Teachers for Music Classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy, 2023. DOI: 10.48550/arXiv.2306.17424. Accessed: Jul. 3, 2023. [Online]. Available: http://arxiv.org/abs/2306.17424.

systematic evaluation

evaluation targets

- **system output**
  - originality
    - plagiarism
    - diversity
    - creativity
  - audio quality
  - musical & aesthetic qualities
- **user experience**
- **other criteria**
  - explainability
  - bias
  - ethical use of data & data curation practices
  - resource use & environmental impact

systematic evaluation
methods

- **subjective** testing
  - preference test
  - Turing test
  - rating of properties
- **objective** testing
  - *reference-independent*
  - *comparison of distributions*

⇒ even fundamental, trivial properties are often not matched between training and generated data

systematic evaluation
methods

- **subjective** testing
  - preference test
  - Turing test
  - rating of properties

- **objective** testing
  - *reference-independent*
  - *comparison of distributions*

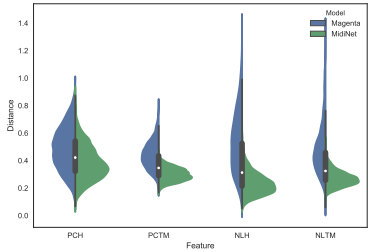
⇒ even fundamental, trivial properties are often not matched between training and generated data

systematic evaluation
methods

- **subjective** testing
  - preference test
  - Turing test
  - rating of properties

- **objective** testing
  - *reference-independent*
  - *comparison of distributions*

  ⇒ even fundamental, trivial properties are often not matched between training and generated data

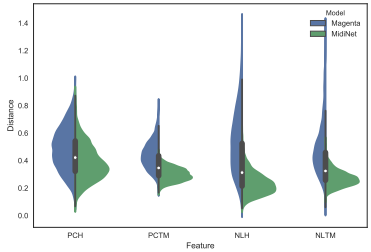


systematic evaluation
methods

- **subjective** testing
  - preference test
  - Turing test
  - rating of properties

- **objective** testing
  - *reference-independent*
  - *comparison of distributions*

  $\Rightarrow$ even fundamental, trivial properties are often not matched between training and generated data



conclusion

conclusion

TODO: write me

thank you!

Georgia Tech Center for Music Technology
College of Design

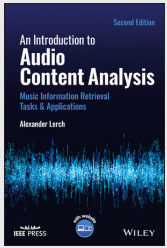
links

alexander lerch: www.linkedin.com/in/lerch

mail: alexander.lerch@gatech.edu

book: www.AudioContentAnalysis.org

music informatics group: musicinformatics.gatech.edu



github.com/alexanderlerch