Machine learning paradigms for music and audio understanding

Emmanouil Benetos
http://www.eecs.qmul.ac.uk/~emmanouilb/
http://machine-listening.eecs.qmul.ac.uk/

Conversational AI Reading Group - November 2025

Queen Mary
University of London

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

CIS centre for intelligent sensing

cis.eecs.qmul.ac.uk

Context

centre for digital music

c4dm.eecs.qmul.ac.uk

www.aim.qmul.ac.uk

CIS centre for intelligent sensing

cis.eecs.qmul.ac.uk

The Alan Turing Institute

www.turing.ac.uk

Talk outline

1. Machine listening
2. Machine listening with limited data
3. Multimodal learning for machine listening
4. Self-supervised learning for machine listening
5. Future perspectives

Machine listening

Machine listening

Machine listening

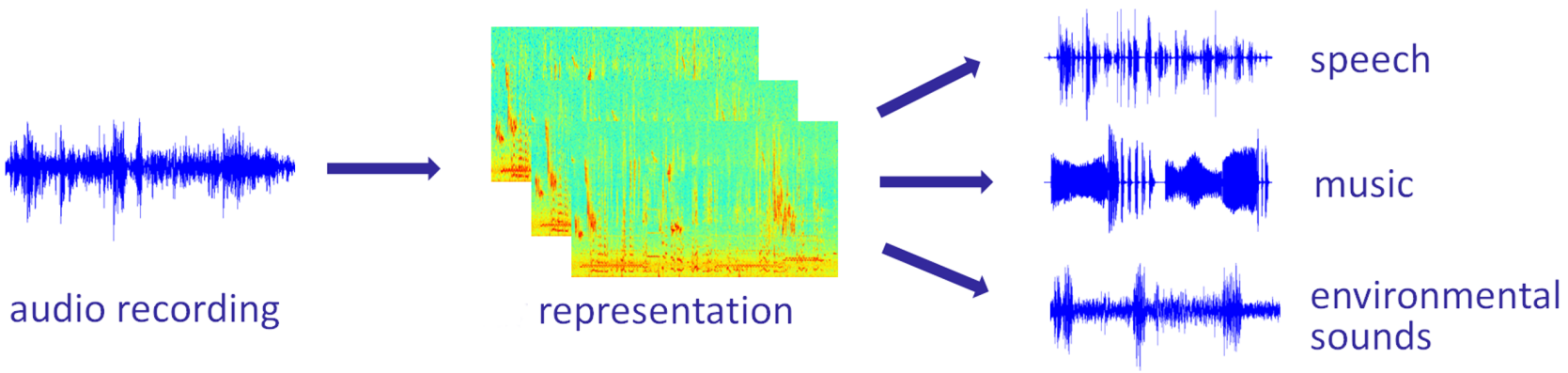
The ability of a machine to interpret and understand audio signals.

Machine listening

Machine listening

The ability of a machine to interpret and understand audio signals.

- **Sounds**: speech, music, environmental/everyday sounds
- **Disciplines**: signal processing, machine learning, acoustics, perception

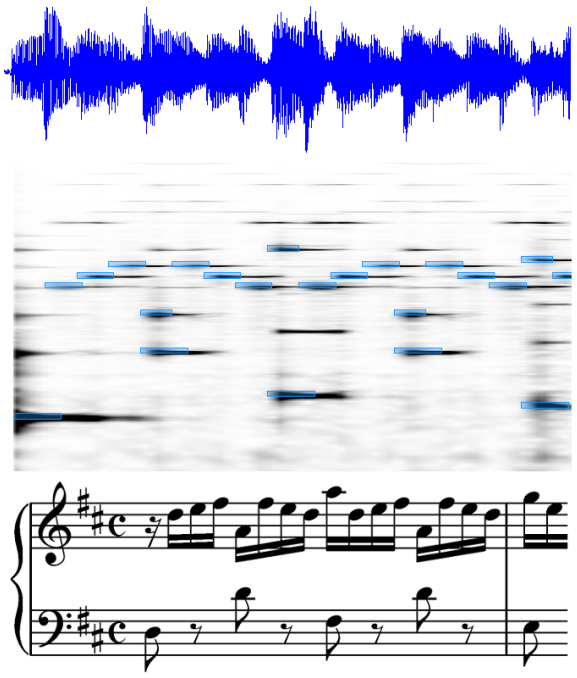


Machine listening for music

Related to the field of **Music Information Retrieval (MIR)**

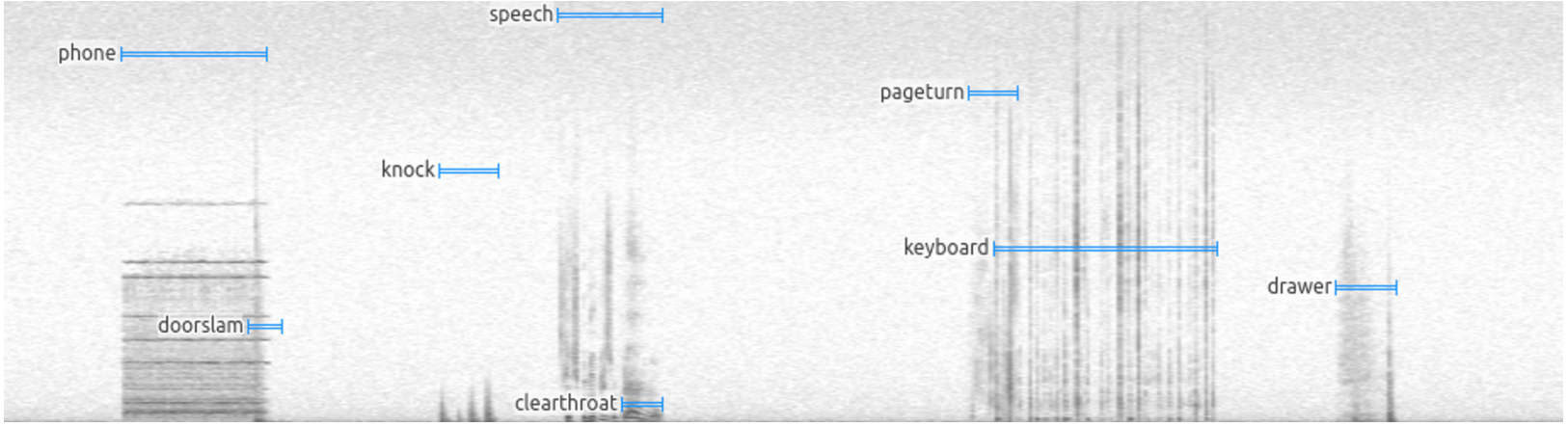
Core problems:

- Music tagging
- Music source separation
- Music transcription
- Audio identification
- & new multimodal music tasks



Challenges in machine listening

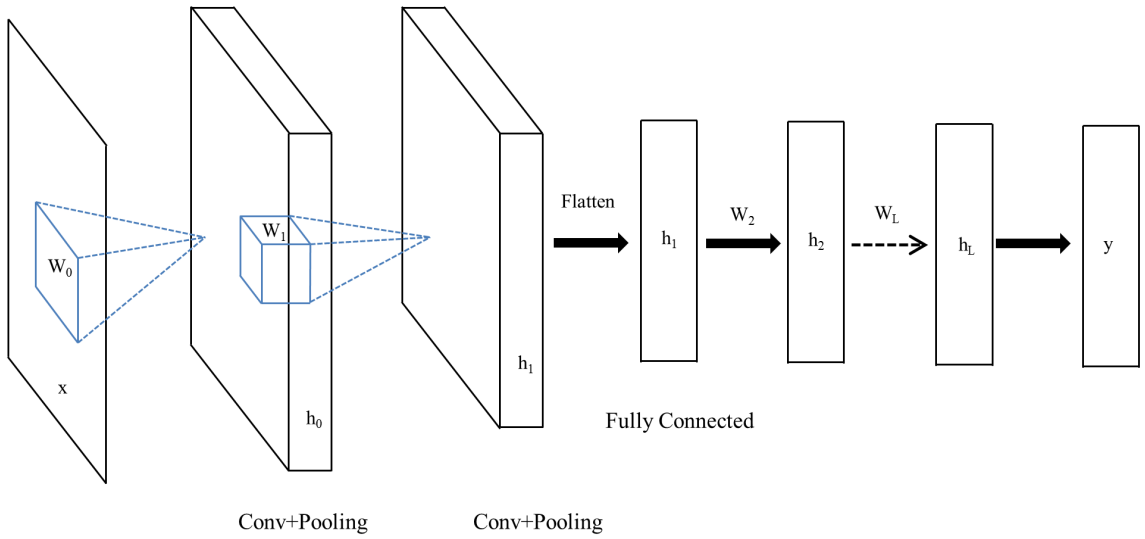
- Multiple overlapping sources
- Data scarcity
- Temporal dependencies
- Noise, distortions and effects
- Unseen domains



Supervised learning for machine listening

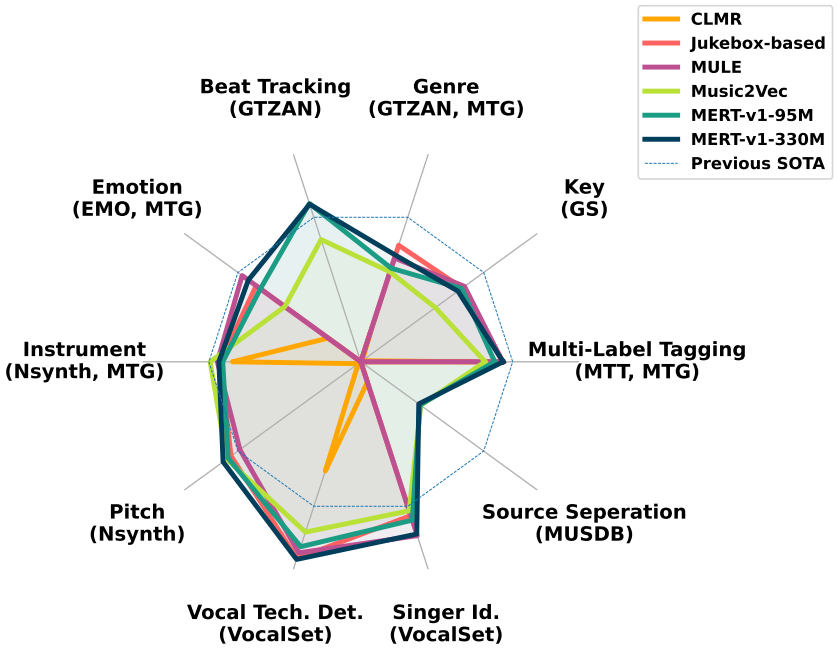
Benchmark approaches for machine listening tasks:

- Adopt a **supervised** deep learning approach
- Assume a sufficiently large, **strongly labelled** dataset
- Time-frequency representations or raw waveforms as **input**
- **Building blocks**: feedforward, convolutional & recurrent layers
- **Loss functions**: cross-entropy, MSE



MARBLE benchmark

- **MARBLE**: an MIR benchmark, defining 18 tasks across four hierarchy levels, utilising 12 public datasets.
- **Task taxonomy**: score-level, performance-level, acoustic-level, high-level description
- Evaluation on several music audio pretrained models: MusiCNN, Jukebox, CLMR, Musicnet-ULarge, MAP-Music2Vec...



R. Yuan et al, "MARBLE: music audio representation benchmark for universal evaluation", in NeurIPS, 2023.

Machine listening with limited data

Domain adaptation for sound recognition

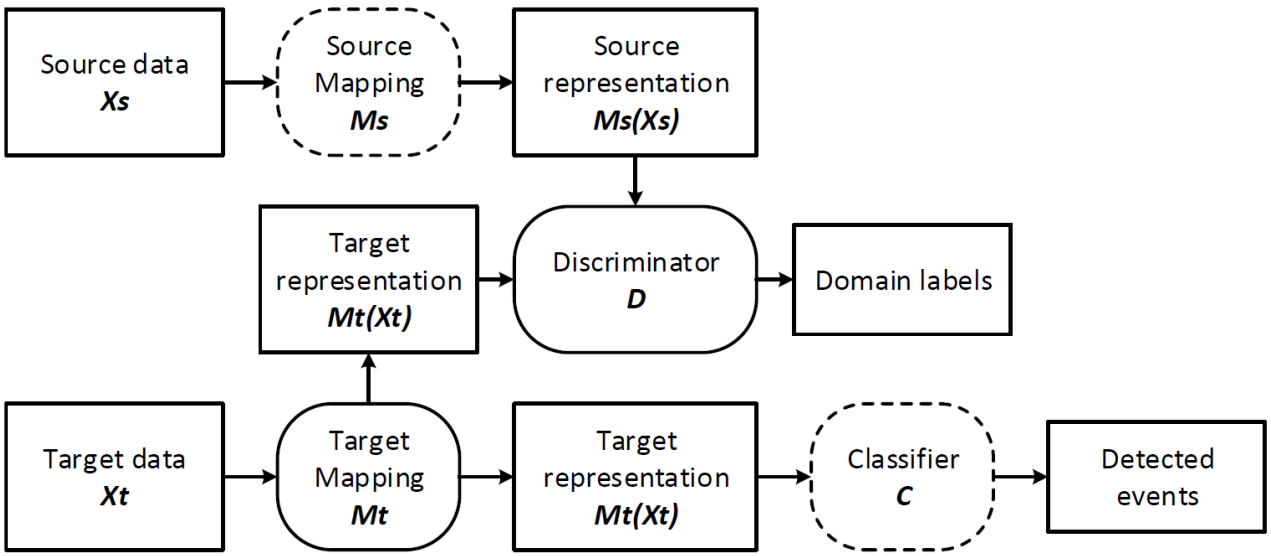
Domain adaptation

Sub-discipline of machine learning which deals with scenarios in which a model trained on a source distribution is used in the context of a different target distribution.

Domain adaptation for sound recognition

Domain adaptation

Sub-discipline of machine learning which deals with scenarios in which a model trained on a source distribution is used in the context of a different target distribution.

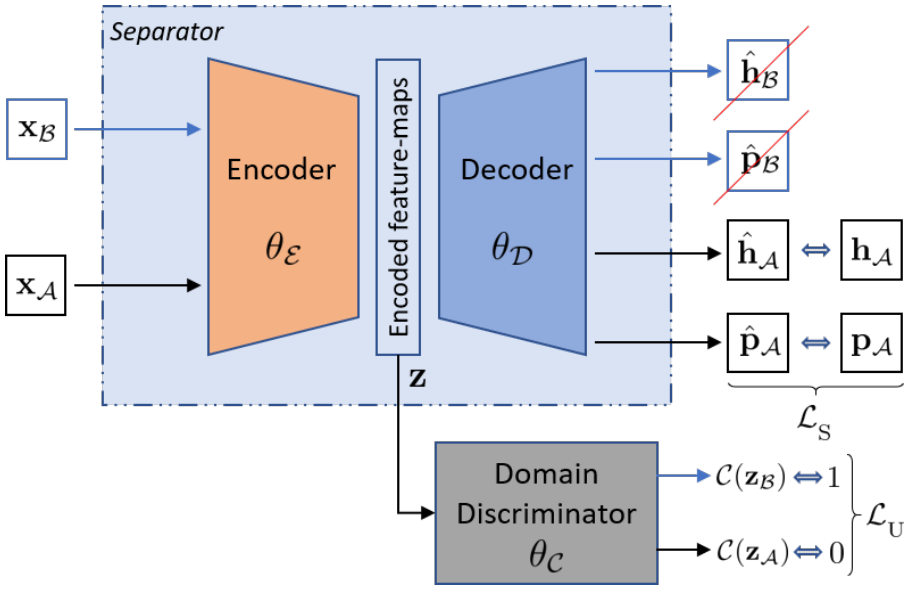


W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: a domain adaptation model for sound event detection", in Proc. IEEE ICASSP, 2020.

Domain adaptation for music source separation

Music source separation system able to adapt to unlabelled mixtures from a new domain.

Framework can be used with any architecture, number of sources, and input representation.



C. Lordelo, E. Benetos, S. Dixon, S. Ahlbäck, and P. Ohlsson, “Adversarial Unsupervised Domain Adaptation for Harmonic-Percussive Source Separation”, IEEE Signal Processing Letters, 28:81-85, 2021.

Few-shot learning for audio classification

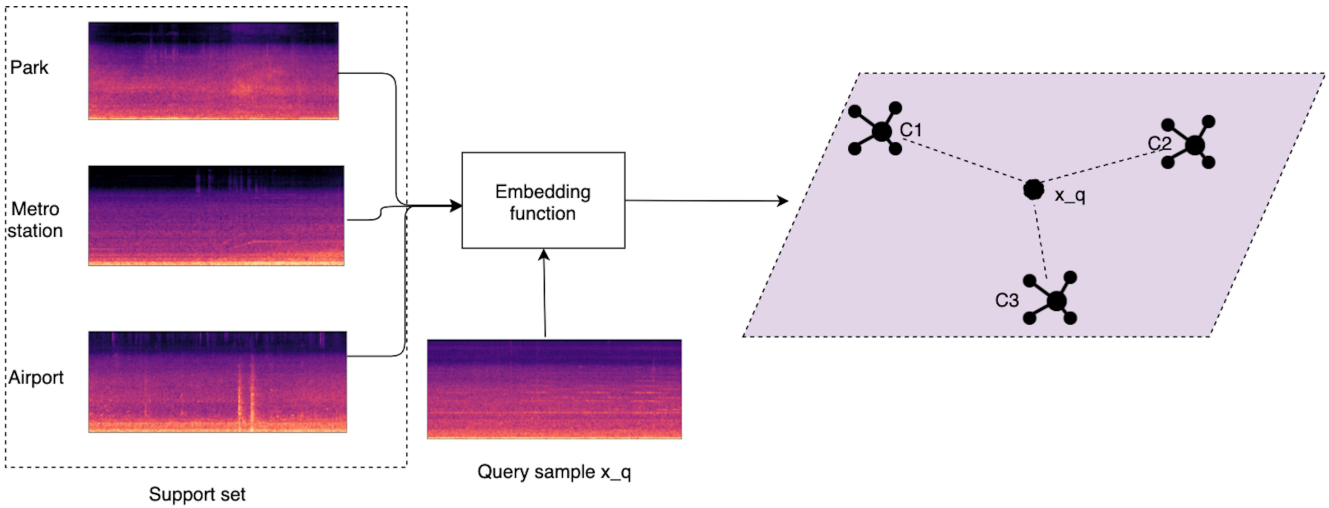
Few-shot learning

Learning from a limited number of labelled examples.

Few-shot learning for audio classification

Few-shot learning

Learning from a limited number of labelled examples.



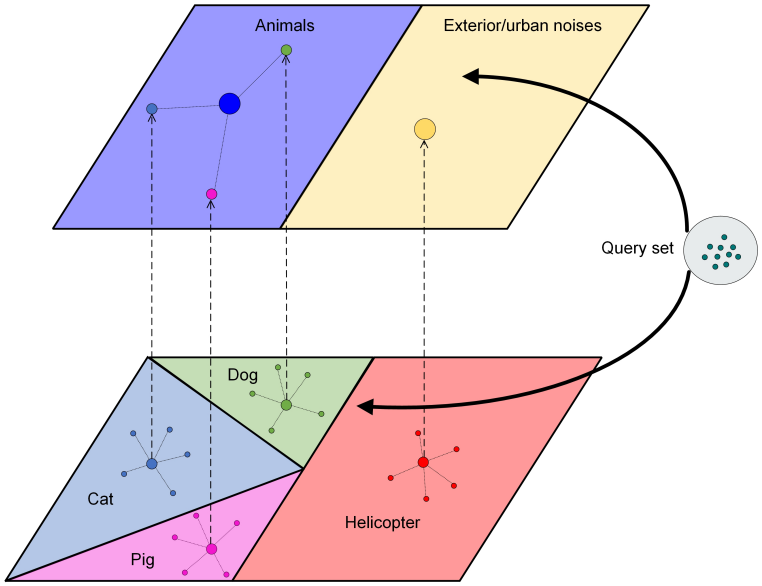
Each class prototype c_k is the mean of the embedded support points x_i belonging to its class: $c_k = \frac{1}{|S_k|} \sum_{(x_i) \in S_k} f_\phi(x_i)$

S. Singh, H. L. Bear, and E. Benetos, "Prototypical networks for domain adaptation in acoustic scene classification", in Proc. ICASSP, 2021.

Few-shot learning for sound recognition

Proposing a **hierarchical prototypical network** to leverage knowledge rooted in audio taxonomies.

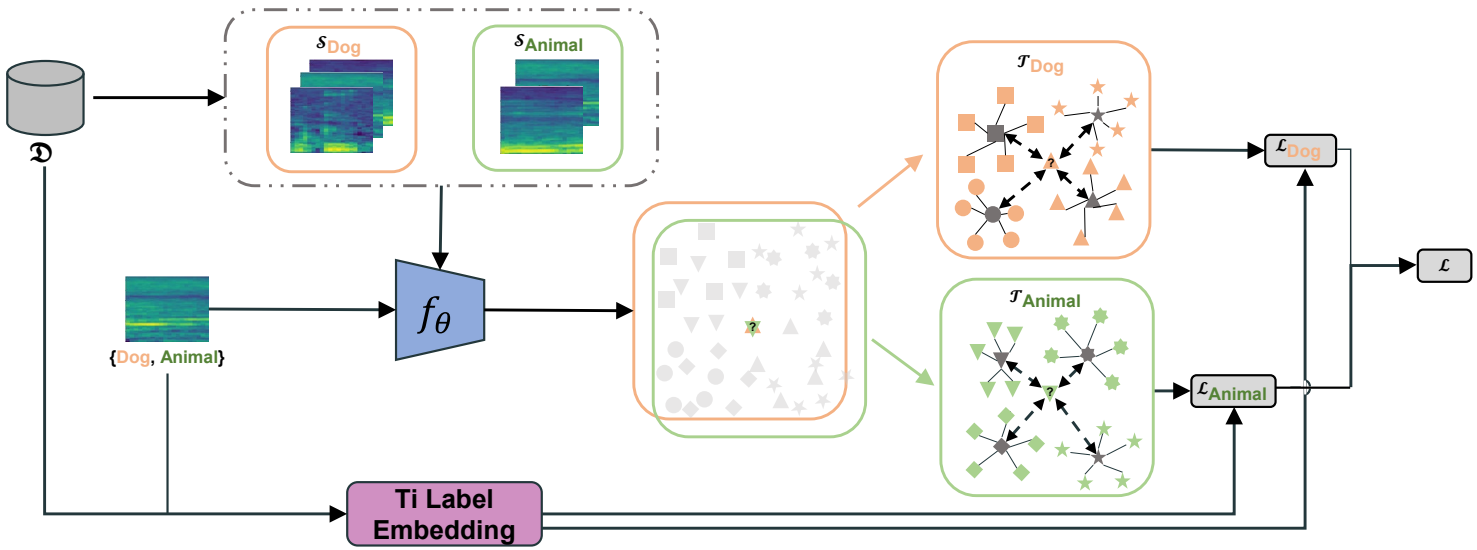
- Prototypes at the lower level: $c_k^{(0)} = \frac{1}{|S_k|} \sum_{(x_i) \in S_k} f_\phi(x_i)$
- Prototypes at a higher level h in the taxonomy: $c_j^{(h)} = \frac{1}{|C_k^{(h)}|} \sum_{c_j^{(h)} \in C_j^{(h)}} c_j^{(h-1)}$



J. Liang, H. Phan, and E. Benetos, "Leveraging label hierarchies for few-shot everyday sound recognition", in Proc. DCASE, 2022.

Few-shot learning for sound recognition

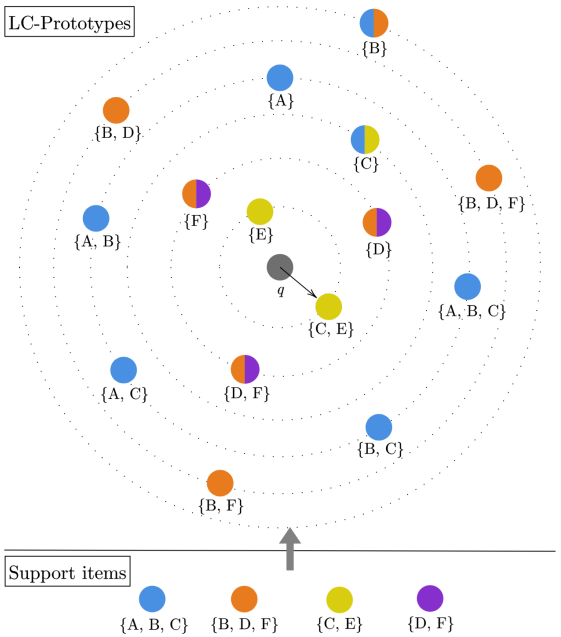
- Extending hierarchical prototypical networks to multi-label classification problems.
- Converts a multi-label classification problem to multiple single-label tasks & incorporates taxonomy knowledge in the training objective.



J. Liang, H. Phan, E. Benetos, "Learning from taxonomy: multi-label few-shot classification for everyday sound recognition", in IEEE ICASSP, 2024.

Few-shot learning for world music

- Label-Combination Prototypical Networks (LC-Protonets) for multi-label few-shot learning
- LC-Protonets generate one prototype per label combination
- Applied to automatic audio tagging across diverse music datasets covering various cultures



C. Papaioannou et al, “LC-Protonets: multi-label few-shot learning for world music audio tagging”, IEEE Open Journal of Signal Processing, 6:138-146, 2025.

Multimodal learning for machine listening

From tags to natural language for audio description

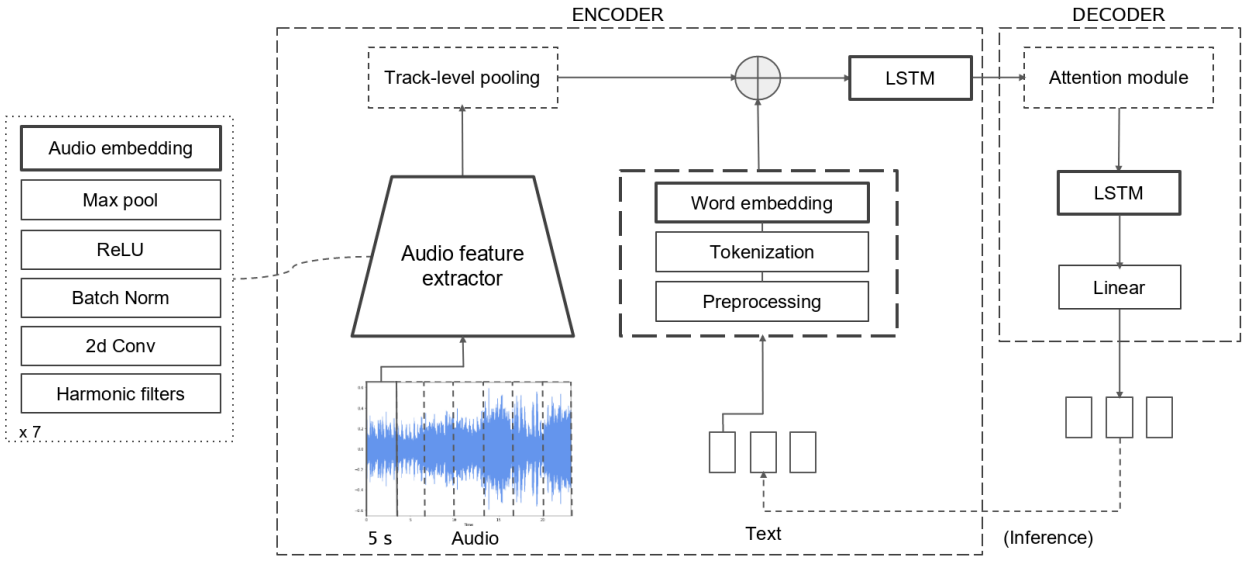
- Audio description typically tackled with classification/regression tasks (e.g. tagging)
- Captioning provides more nuanced description, uses natural language and can be extended to new concepts
- Community activity on **automated audio captioning** as part of DCASE Challenge series

~~"rock", "powerful", "emotional", "guitars"~~

"*This is a powerful rock track featuring guitars with an emotional bassline*"

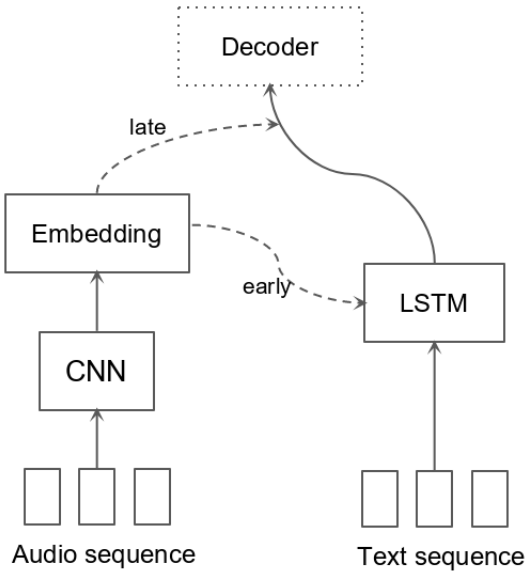
MusCaps: generating captions for music audio

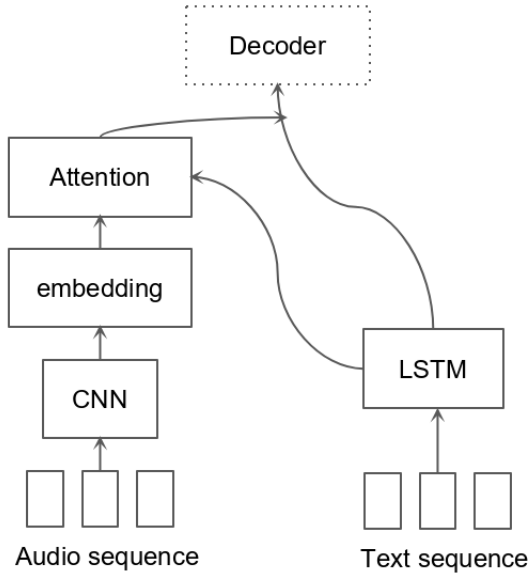
- First audio captioning model focussed on music
- Encoder-decoder network consisting of a multimodal CNN-LSTM encoder with temporal attention and an LSTM decoder



I. Manco, E. Benetos, E. Quinton and G. Fazekas, “MusCaps: generating captions for music audio”, in IJCNN, 2021.

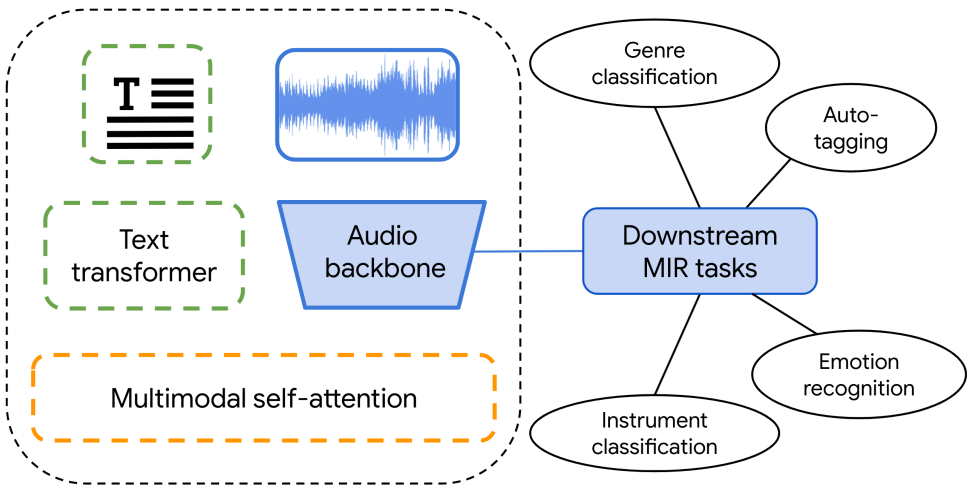
MusCaps: modality fusion





MuLaP: music and language pretraining

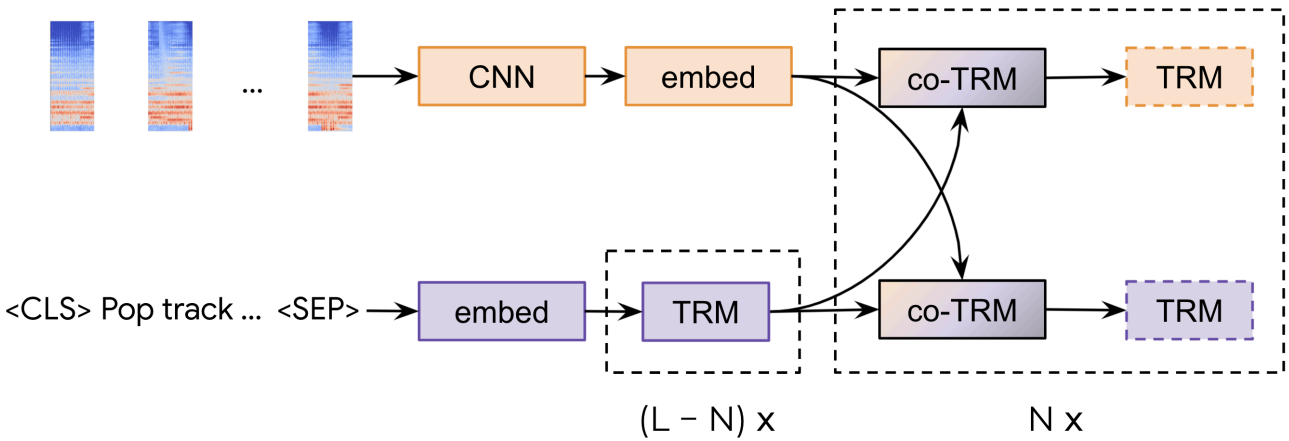
- MuLaP: leveraging weakly aligned natural language and audio to learn general-purpose music representations
- Can attain similar or better downstream performance when compared to traditional supervised techniques



I. Manco, E. Benetos, E. Quinton, G. Fazekas, “Learning music audio representations via weak language supervision”, in ICASSP, 2022.

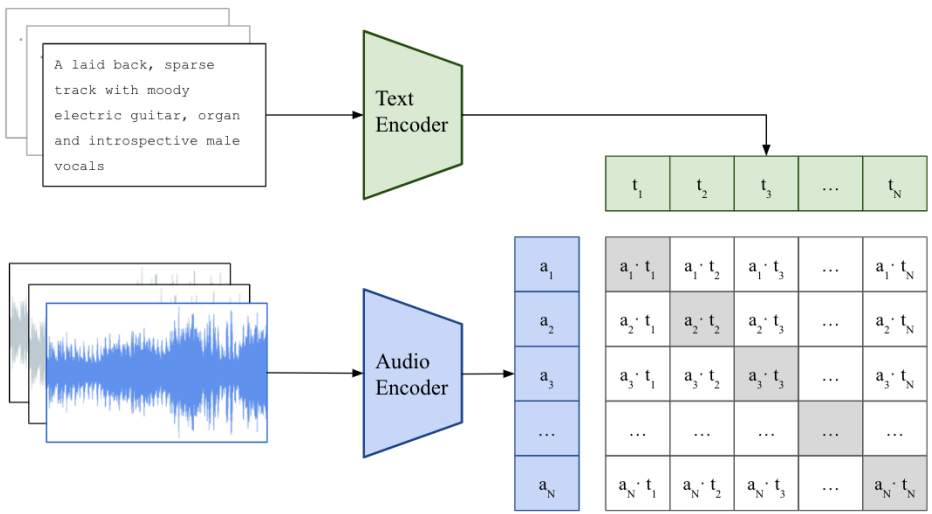
MuLaP: music and language pretraining

- We design a multimodal Transformer made of two modality-specific branches (audio and text) and a co-attentional module
- We pre-train using three learning objectives: masked language modelling, masked audio modelling, audio-text matching



MusCALL: contrastive audio-language learning for music

- Exploring multimodal contrastive learning for music audio language
- Applied to cross-modal retrieval for music, transferred to music classification tasks in a zero-shot setting



I. Manco, E. Benetos, G. Fazekas, and E. Quinton, “Contrastive audio-language learning for music”, in ISMIR, 2022.

MusCALL: failure cases

Query Text	Text of the Top-1 Audio
An atmospheric and introspective orchestral track featuring strings, piano, and synth.	*An inspirational and moody orchestral track featuring strings and choir.*
Deep chilled out space jazz with crisp beats and lush electronics.	*Jaunty swing featuring trumpet.*
Up tempo, pumping dance pop with female vocals.	*Quirky, fun, positive disco party music.*

In some “failure” cases, MusCALL still retrieves items that are semantically related to the query

Song Describer Dataset

- **Song Describer dataset**: a new crowdsourced corpus of high-quality audio-caption pairs
- 1.1k human-written natural language descriptions of 706 music recordings, all publicly accessible and released under CC licenses

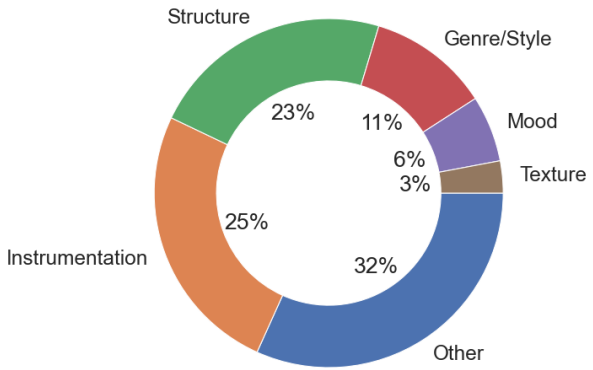
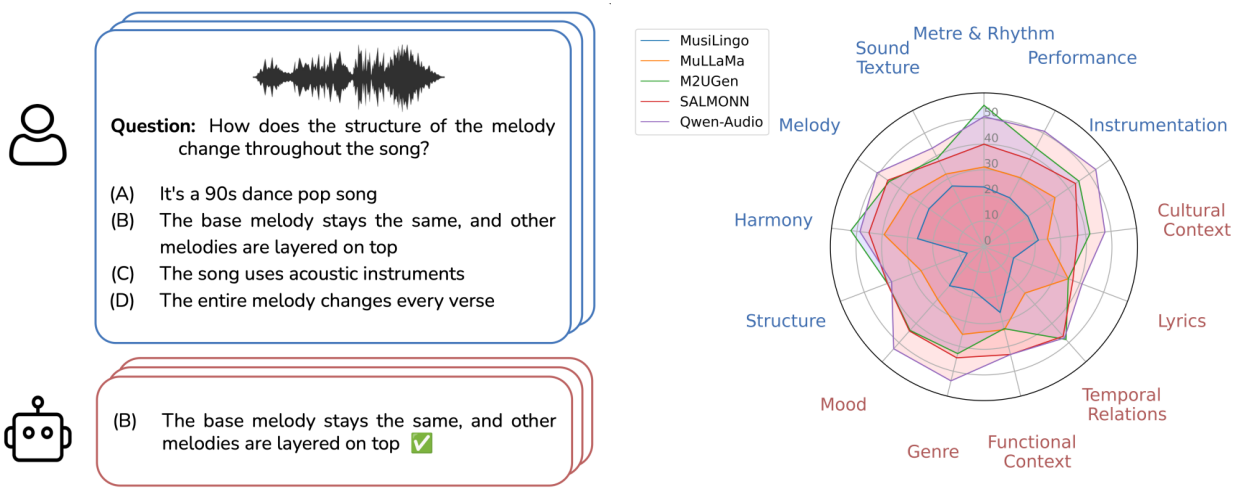


Figure: Distribution of music aspects by the most frequent word stems in the collected captions.

I. Manco et al, "The Song Describer Dataset: a corpus of audio captions for music-and-language evaluation", in NeurIPS ML4Audio Workshop, 2023.
https://github.com/mulab-mir/song-describer-dataset

MuChoMusic Benchmark

- Benchmark for music understanding in audio-language models
- 1,187 multiple-choice questions about 644 music tracks
- Challenging and robust to unimodal shortcuts, exposes both audio and language hallucinations



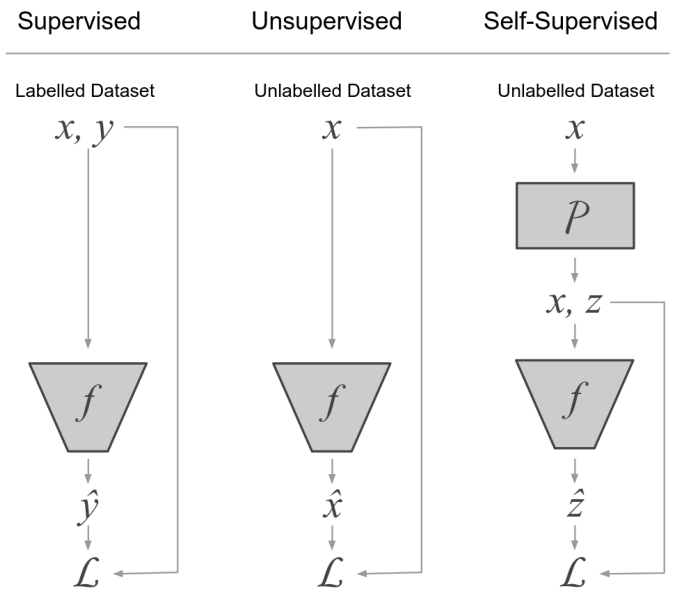
B. Weck et al, "MuChoMusic: Evaluating music understanding in multimodal audio-language models", in ISMIR 2024. **Best paper award**
https://github.com/mulab-mir/muchomusic

Self-supervised learning for machine listening

Self-supervised learning

Self-supervised learning

Special case of unsupervised learning, relying on pretext tasks that exploit knowledge about the data modality used for training.

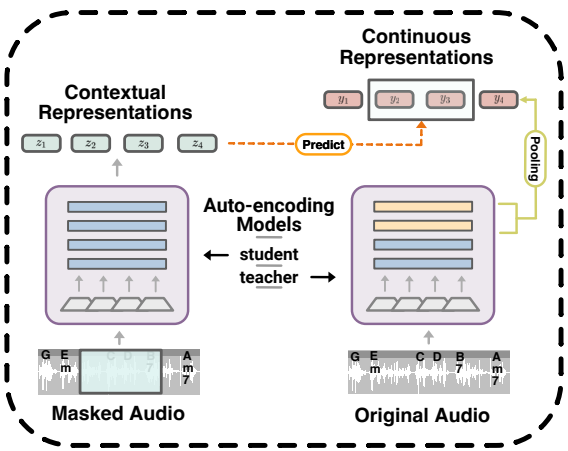


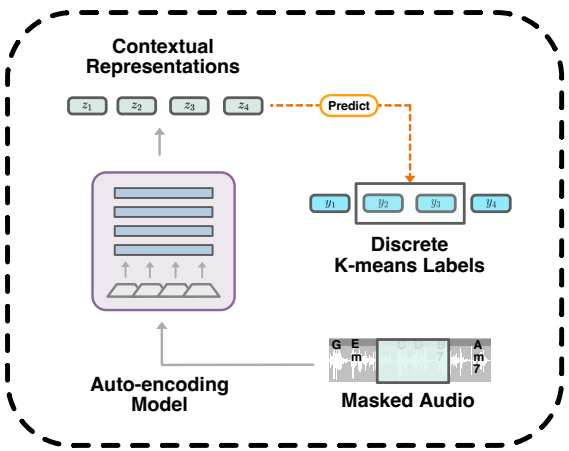
[Source: Ericsson et al, IEEE SPL 2022]

Common SSL approaches: transformation prediction, masked prediction, instance discrimination, and clustering

Self-supervised learning for music

- Lack of music domain knowledge, lack of long-term sequence modelling
- Retraining speech SSL models (data2vec and HuBERT) on music data
- Pretraining on speech is helpful; pretraining on music is better; but models suffer on modelling polyphony and harmony

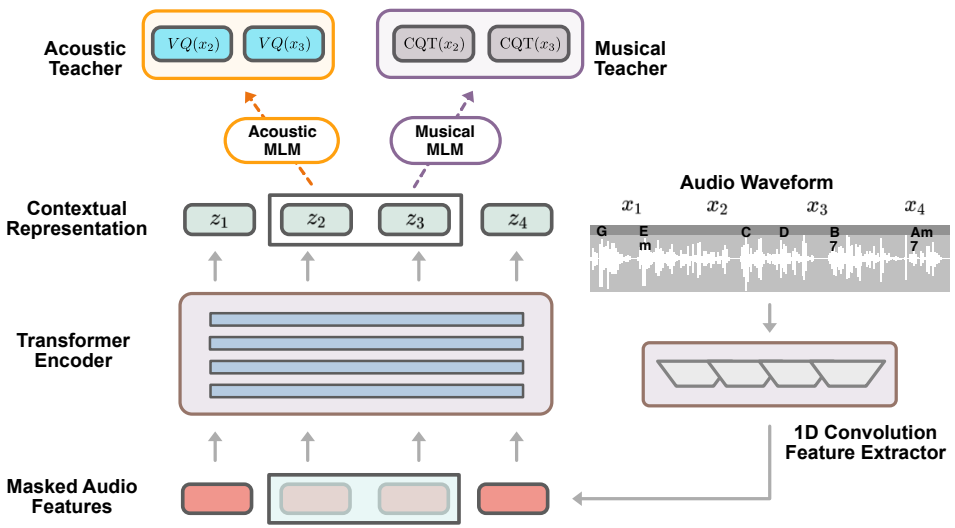




Y. Ma et al, “On the effectiveness of speech self-supervised learning for music”, in ISMIR, 2023.

Self-supervised learning for music

- **MERT**: Music undERstanding model with large-scale self-supervised Training
- Multi-task paradigm to balance the acoustic and musical representation learning
- Overall architecture is similar to HuBERT, adapted to music: CQT reconstruction loss plus EnCodec (Défossez et al, 2022)



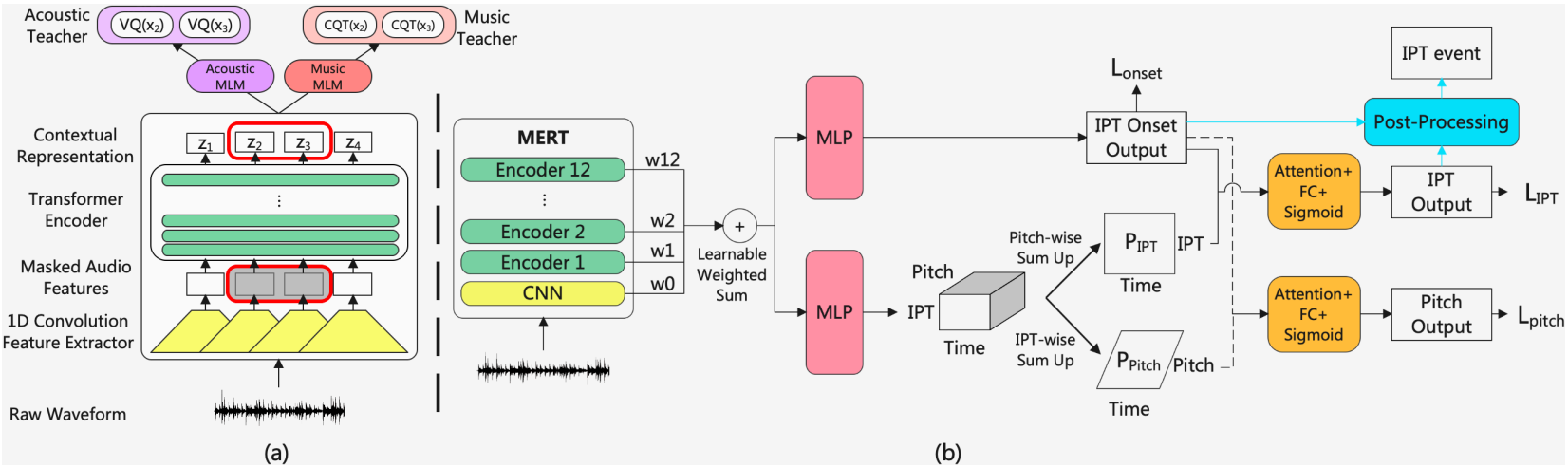
Y. Li et al, "MERT: acoustic music understanding model with large-scale self-supervised training", in ICLR, 2024.

Self-supervised learning for music

- Model variant trained on publicly available data only (MERT-95M-public)
- Preliminary versions used k-means clustering with audio features: scaling issues
- Some capability in handling longer sequences, but still limited by the short 5-second training context
- Community impact: >750k downloads of MERT in Hugging Face (https://huggingface.co/collections/m-a-p)

Self-supervised learning for music

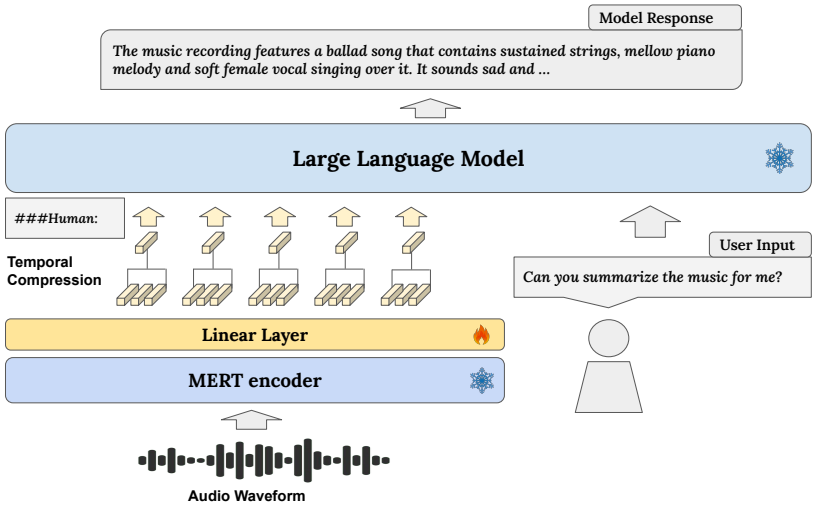
- Adapting MERT to a low-resource task: instrument playing technique (IPT) detection
- Pretraining on MERT, finetuning for IPT detection, pitch detection, and IPT onset detection



D. Li et al, "MERTech: instrument playing technique detection using self-supervised pretrained model with multi-task finetuning", in ICASSP, 2024.

Self-supervised + multimodal learning for music

- **MusiLingo**: a system for music caption generation and music-related query responses.
- Aligning audio representations from MERT with the frozen Vicuna-7B language model
- New dataset of 60k music Q&A pairs



Z. Deng et al, “MusiLingo: bridging music and text with pre-trained language models for music captioning and query response”, in NAACL, 2024.

Future perspectives

Future perspectives

- Significant performance gaps between audio LLMs and supervised MIR models
- Continual learning for audio and music
- Multimodal AI for acoustic and music understanding: beyond audio and text?
- Acoustic diversity & understanding low-resource corpora
- Resource-efficiency and sustainability

Many thanks to

- Helen Bear
- Simon Dixon
- George Fazekas
- Jinhua Liang
- Carlos Lordelo
- Yinghao Ma
- Ilaria Manco
- Charis Papaioannou
- Huy Phan
- Elio Quinton
- Shubhr Singh
- Benno Weck
- Wei Wei

SUPPORT:

UKRI Engineering and Physical Sciences Research Council

Royal Academy of Engineering