Multi-Source Contrastive Learning for Musical Audio

Christos Garoufis[1,2,3], Athanasia Zlatintsi[1,2,3], Petros Maragos[2,3]

christos.garoufis@athenarc.gr, athanasia.zlatintsi@athenarc.gr, maragos@cs.ntua.gr

[1]Institute of Language and Speech Proc., Athena Research Center, Athens, Greece
[2]Institute of Robotics, Athena Research Center, Athens, Greece
[3]School of ECE, National Technical University of Athens, Athens, Greece

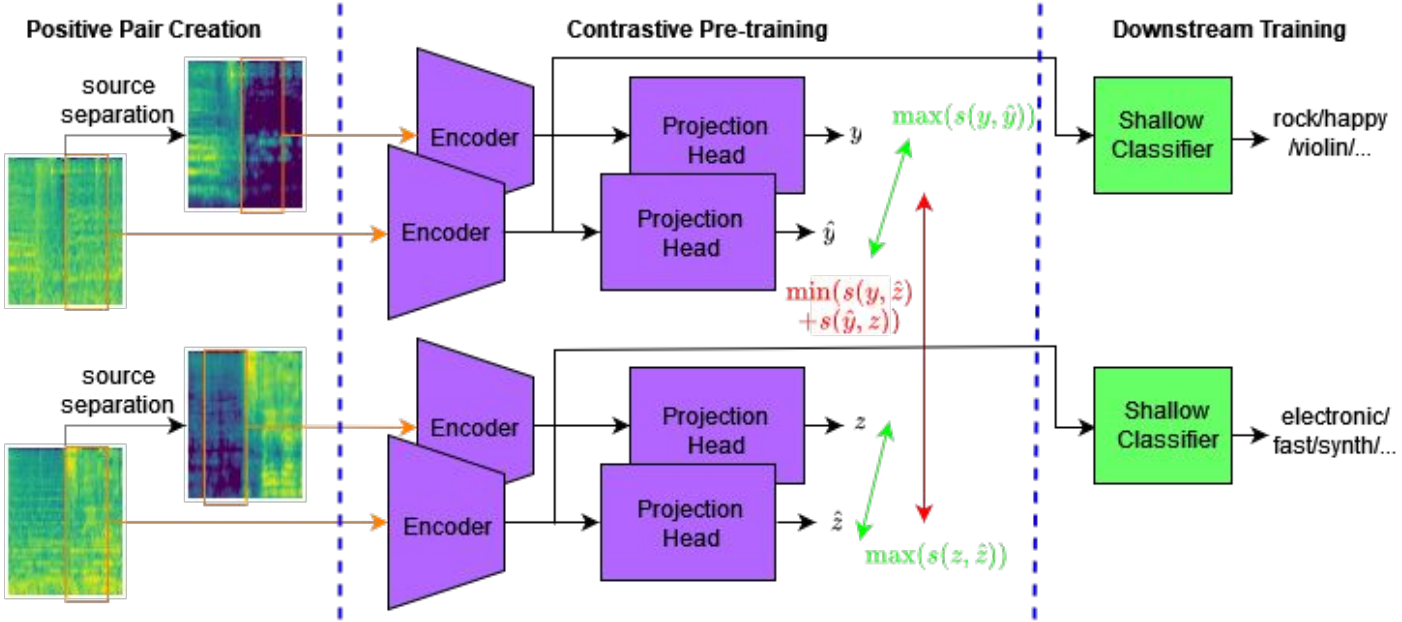
Overview

- **Contrastive self-supervised learning**: Representation learning from augmented **anchor-positive pairs** of large, unlabelled data collections
  - Representations of each pair are enforced to be **as close as possible**.
  - Representations of different pairs are enforced to **deviate from each other**.

- **Motivation:** The various co-playing sources in musical pieces are **harmonically** and **rhythmically** coordinated, and their existence/absence carries **semantic information**.

- **Contribution:** A framework for music representation learning, using **music source association** (MSA) as a pretext task in a contrastive learning setup.
  - Competitive performance to self-supervised baselines in three downstream tasks.

Methodology



Experimental Setup

Two-stage training:

- **Pre-training** the encoder, with a contrastive loss objective, to associate music pieces with source excerpts from a pre-training dataset.
- Training **shallow classifiers** on top of the **frozen encoder** in downstream tasks.

Pre-Training:

- **Dataset:** Magna-Tag-A-Tune (MTAT): 25863 song pieces, 30 sec each, 188 tags
- MTAT does not include source tracks → acquisition of source tracks (bass, drums, vocals, accompaniment) via an **automatic source separation system** (open-unmix).

Shallow classifier training:

- **Downstream tasks:** Music auto-tagging (MTAT), instrument classification (NSynth), music genre recognition (FMA)

Results: Multi-Source Models

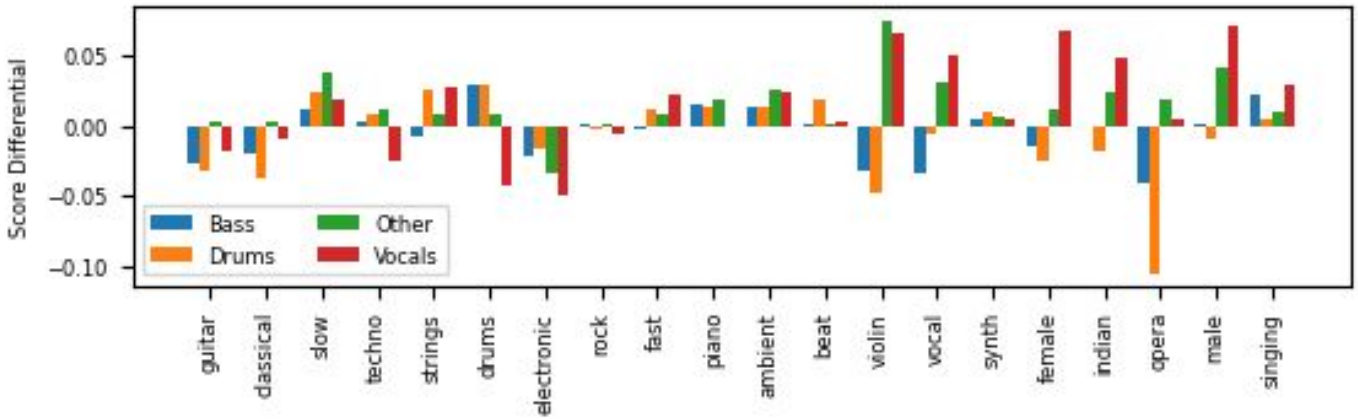
SSL Framework	MTAT		MTAT*		NSynth	FMA
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	WA (%)	WA (%)
CLMR [9]	-	-	0.887	0.356	-	0.484
COLA [7]	0.886	0.396	0.880	0.334	0.593	0.460
COLA [7] + MWS [10]	0.898	0.425	0.892	0.358	**0.645**	0.493
COLA [7] + Random Mask	0.883	0.390	0.880	0.337	0.632	0.476
COLA [7] + MSA (ours)	**0.900**	**0.429**	**0.895**	**0.361**	0.627	**0.510**

- MSA **outperforms** the COLA baseline in all three downstream tasks.
- **Comparable** performance to the data-driven MWS method, as well as CLMR.

- **Faster convergence** than MWS during the early stages of pre-training, but performance balances out as pre-training progresses.
- The **quality of the separated sources** impacts the downstream performance.

Results: Source-Targeted Models

- On average, the **vocal** and **accompaniment**-based models perform the best.
- In general, the **multi-source** model performs better than all targeted ones.
- Models display a **specialization**, according to the target pre-training source.



Thank you for your attention!

For the source code and pre-trained models, visit our github page! https://github.com/cgaroufis/MSCOL_SMC23

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers” (Project Number: 7773). For more information: https://i-mreplay.athenarc.gr/