Multi-Source Contrastive Learning for Musical Audio

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

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

[1]Institute of Language and Speech Proc., Athena Research Center, Athens, Greece

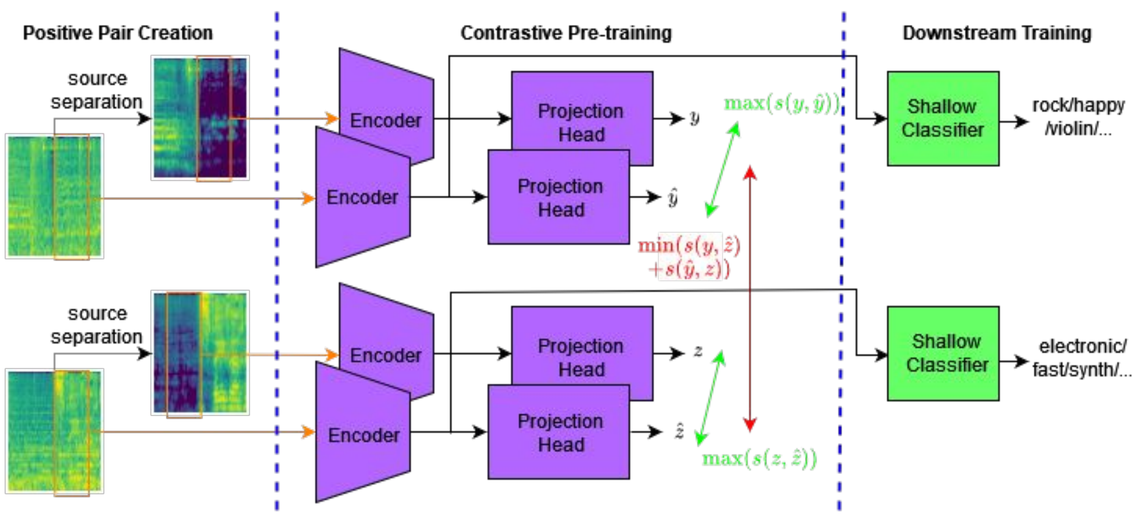
[2]Institute of Robotics, Athena Research Center, Athens, Greece

[3]School of ECE, National Technical University of Athens, Athens, Greece

1. Introduction

- **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
  - Pretraining: Associate a music excerpt with a randomly selected extracted source.
  - Semantic information incorporated by modeling the existence or absence of sources in the musical piece.
  - **Competitive results** to self-supervised alternatives in three downstream tasks.
  - Insights about the **correlation** between particular tags and musical sources.

2. Methodology



- **Positive Pair Creation:** Match each musical piece excerpt with a **time-shifted source excerpt** from the same piece.
  - Each batch contains instances from the **same source**.
  - Some batches also include **silent** instances of the target source.
- **Backbone Encoder:** Based on **EfficientNet-B0** (similar to COLA [1])
  - Series of inverted residual depth-wise convolutional blocks, reduce input resolution.
  - Max-pooling and linear layer for flattening the representation → **512-D embedding**.
- **Contrastive Loss Objective:** Variant of the **NT-XEnt Loss**: $\mathcal{L} = -\sum_{y \in S} \log \frac{\exp(s(y,\hat{y})/T)}{\sum_{z \in S^{*-}} (\exp(s(y,z)/T))}$, due to the false positives from silent source excerpts.
  - S contains the **anchor embeddings** per batch.
  - S^{*-} contains the **non-silent** positive embeddings and the **centroid** of the silent ones.
  - Applied upon linear projections of embedding pairs in a batch-wise fashion.
  - **Bilinear** similarity used as the similarity function.

3. 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 songs, 30 sec each, 188 tags
  - The top-50 most frequent are used as a popular benchmark
  - We further filter the dataset, using only songs with at least one top-50 tag.
- MTAT does not include source tracks → acquisition of source tracks (bass, drums, vocals, accompaniment) via an **automatic source separation system** (open-unmix).
- **Data preprocessing**: Mel-spectrogram computation (4-sec length, 25ms window, 10ms overlap, 64 mel bands), 1-sec segments cropped during training
  - Source segments with low mean energy are replaced with **silence**.
- **Training**: 10000 training steps, batch size 128, Adam (lr = 0.001, halved at 5000 steps)

Shallow classifier training:

- **Downstream tasks:** Music auto-tagging (MTAT), instrument classification (NSynth), music genre recognition (FMA) – commonly used train/validation/test splits.
  - Reporting results for both filtered (MTAT) and unfiltered (MTAT*) dataset versions.
- **Classifiers:** Linear (MTAT and FMA), 1-layer MLP (NSynth)
- Training/inference in spectrogram-level, aggregation per musical piece via averaging
- Results reported are the average of 5 runs for each model.

Baselines (trained with the same pre-training dataset and protocol)

- The unmodified **COLA** [1] framework.
- The COLA framework [1], along with the data-driven methodology for pair creation devised in Zhao et al. (**masking, warping and shifting** - MWS) [2].
- The COLA framework, but with sources extracted by **random soft masking**.
- **CLMR** [3] (results mined from [2] and [3]).

4. Multi-Source Models: Results

Comparison to baselines:

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**

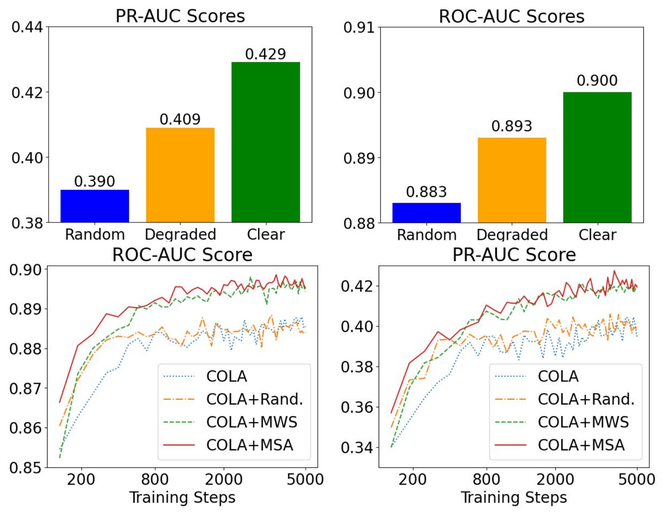
- MSA **outperforms** the COLA [1] baseline in all three downstream tasks.
- **Comparable** performance to the data-driven MWS method [2], as well as CLMR [3].
- Improved results compared to random masking → **musically meaningful soft masking** is necessary to learn useful representations.

Ablation study:

- Inclusion of **non-silent** source segments during pre-training improves performance in auto-tagging and genre classification, not in instrument classification.
- Similar effect observed regarding intra-batch **source homogeneity** and the proposed **loss function**.
- **Time shifting** of the source excerpts is critical to the performance.

Qualitative Analysis:

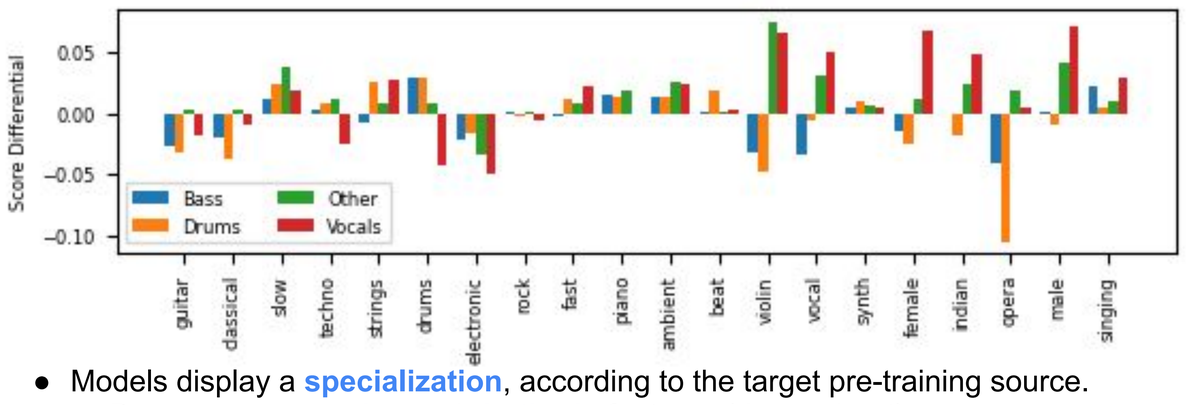
- **Faster convergence** than MWS during the early stages of pre-training, but
- Performance balances out as pre-training progresses due to the **multitude** of augmented examples MWS generates.
- In contrast to environmental sounds [4], the **quality of the separated sources** impacts the performance.



5. Source-Targeted Models: Results

Source	MTAT		MTAT*		NSynth	FMA
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	WA (%)	WA (%)
Bass	0.885	0.392	0.880	0.332	0.624	0.487
Drums	0.881	0.385	0.876	0.328	0.611	**0.503**
Accomp.	0.892	0.410	0.888	**0.348**	**0.636**	0.480
Vocals	**0.896**	**0.412**	**0.891**	**0.348**	0.632	0.481
None (COLA)	0.886	0.396	0.880	0.334	0.593	0.460
Multi-Source	0.900	0.429	0.895	0.361	0.627	0.510

- On average, the **vocal** and **accompaniment**-based models perform the best.
- **Drum** and **bass**-based models do not beat COLA (exception: genre recognition)
- In general, the **multi-source** model performs better than all targeted ones.
- Tag-wise average PR-AUC scores for source-targeted models, relative to COLA:



- Models display a **specialization**, according to the target pre-training source.
  - Drums → **percussive** and **rhythmic** features (drums, beat).
  - Accompaniment → **accompanying instruments** (violin, piano, guitar).

6.Conclusions

- Proposed a **contrastive learning framework** for learning representations for musical audio, using **musical source association** as a pretext task.
  - **Competitive performance** to other methods in a number of downstream tasks.
  - Can be **steered** towards specific features, based on the selected musical source.
- **Future work:**
  - Explore the **scalability** of the framework in larger datasets or with different encoders
  - Examine the feasibility of the learned embedding subspace for source-wise **music recommendation** and **similarity**.

References

[1] A. Saeed et al., "Contrastive Learning of General-Purpose Audio Representations," in Proc. ICASSP 2021

[2] H. Zhao et al., "S3T: Self-Supervised Pre-training with Swin Transformer for Music Classification," in Proc. ICASSP 2022

[3] J. Spijkervet et al., "Contrastive Learning of Musical Representations," in Proc. ISMIR 2021

[4] E. Fonseca, et al., "Self-Supervised Learning from Automatically Separated Sound Scenes," in Proc. WASPAA 2021

Acknowledgements

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/