Pre-Training Music Classification Models via Music Source Separation

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]Robotics Institute, Athena Research Center, Athens, Greece

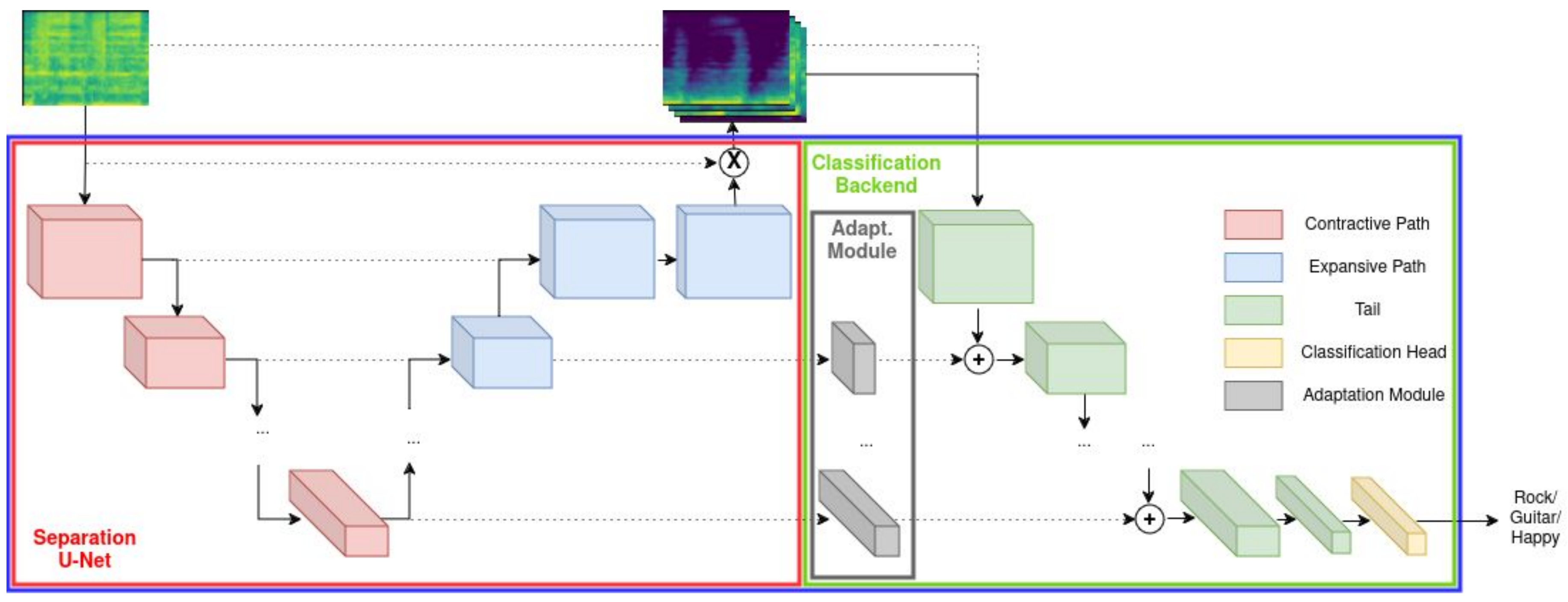
[3]School of ECE, National Technical University of Athens, Athens, Greece

EUSIPCO 2024

1. Introduction

- **Motivation:** The task of **music source separation**, i.e., decomposing a musical piece into its vocal or instrumental components, can aid in **music representation learning**:
  - Particular timbral or semantic attributes are mostly tied to different sources.
  - Features learned by separation networks might also contain high-level information.
- **Contribution:** A two-stage framework for music classification tasks:
  - **Pre-train** a **U-Net frontend** on music source separation.
  - Connect a **classification backend** to the U-Net, and **finetune them jointly** on the desired task.
- Experiments with two music classification datasets and two different backends:
  - Improved performance compared to 1) the bare backend, 2) the joint network without music source separation pre-training.
  - The increased performance can be traced to specific **source-related tags**.

2. Methodology



The proposed framework employs a **composite architecture**, entailing a **separation network**, a **classification backend** and a **feature adaptation module**, inspired by [1].

- **Separation Network:** Convolutional U-Net network, based on [2]:
  - STFT magnitude as network input.
  - **Contractive Path (Encoder)**: 6 blocks, with 2 convolutional layers and max pooling.
  - **Expansive Path (Decoder)**: Symmetric to the encoder, contains transposed convolutions and upsampling layers. Connected to the encoder via skip connections.
- **Classification Backend:** Experiments with two different architectures:
  - **Short-Chunk CNN** [3]: 7 convolutional blocks with 2 layers in each block, followed by a 2-layer MLP for classification.
  - **Audio Spectrogram Transformer** (AST) [4]: Input patchfied into 16x16 patches, using a 48x8 grid. 12 Transformer blocks, linear layer for classification
    - **ImageNet weights** used for Transformer initialization.
- Input internally transformed into the mel scale (128 bands).
- **Adaptation Module:** Convolutional layers that change the dimensionality of the expansive features, **propagating** them to the classification backend.
  - **CNN**: Convolutions with **1x1 kernels** and max pooling across the frequency dimension before feature summation, before each convolutional block.
  - **AST**: Intermediate representations are **patchified**, with appropriate kernels to align them with the **internal AST resolution** (48x8 grid), and inserted into the AST every two Transformer blocks.
- **Training Scheme:** We devise a three-stage training protocol:
  - **Pre-train the U-Net** with a music source separation objective.
  - (Optionally) **pre-train the classification backend** in the target task until convergence.
  - **Jointly finetune** the complete network at the desired task.

3. Experimental Setup

U-Net Pre-Training:
- Dataset: musdb18
  - 150 songs, sampled at 44.1 kHz, total duration of approx. 10 hours.
  - Apart from the full tracks, contains vocal, drum, bass and melodic accompaniment stems.
- U-Nets pre-trained for all **uni-source cases**, as well as **multi-source separation**.

Downstream Training:
- Music auto-tagging: Magna-Tag-A-Tune (MTAT)
  - 25,863 29-sec song excerpts, sampled at 16 kHz (duration: 210 hours)
  - **50 most frequent tags** commonly used as a classification benchmark.
  - **Multi-class classification** problem → BCE loss, ROC-AUC/PR-AUC as metrics.
- Genre classification: Free Music Archive (FMA, *medium* subset):
  - 25,000 30-sec song excerpts, sampled at 44.1 kHz (duration: 208 hours)
  - Each excerpt is annotated with 1 out of **16 root genres.**
  - **Binary classification** problem → CCE loss, weighted accuracy (%) as the metric.

Data Preprocessing:
- All audio were resampled to 16 kHz for compatibility purposes.
- STFT computation parameters: 512-sample window length, 160-sample hop size.

4. Results & Discussion

- We compare our framework, over various source pre-training configurations, to the following baselines (standard training-validation-testing splits):
  - **Bare classification backends** (without the U-Net) – top row.
  - The proposed joint architecture, **without a pre-training scheme** – second row.
- Segment-level training and inference; per-excerpt results obtained via averaging.

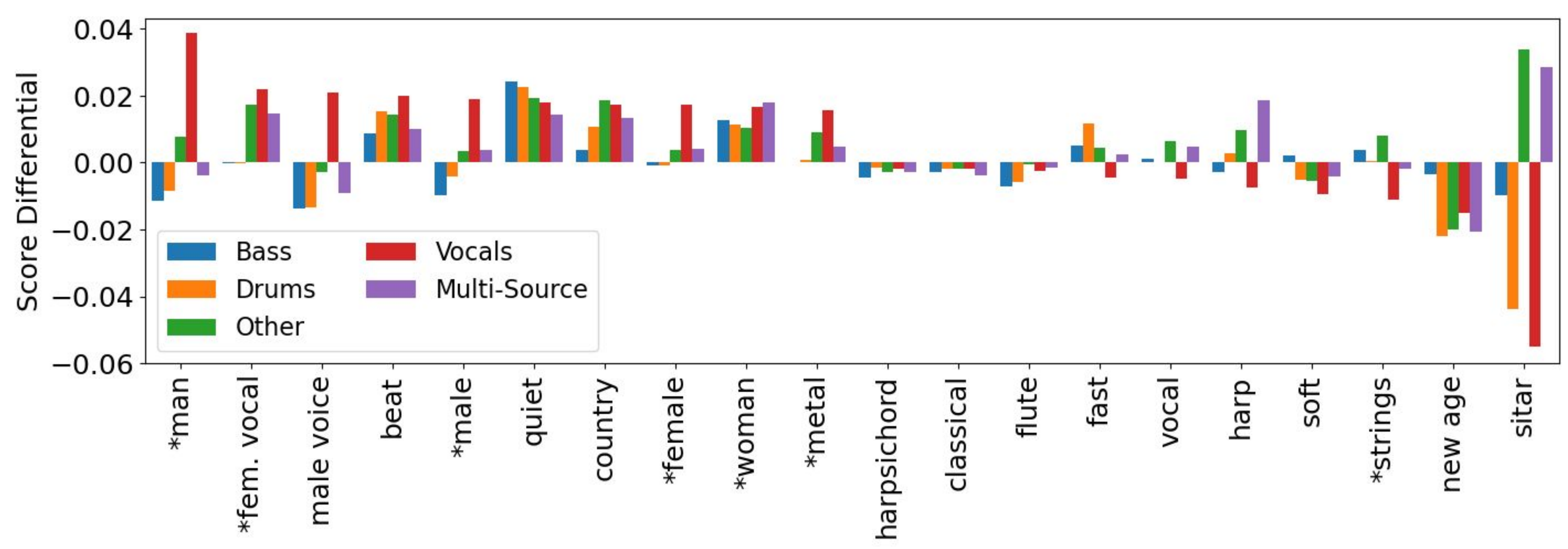
U-Net	Pre-Training	Source(s)	CNN Backend: MTAT ROC-AUC	CNN Backend: MTAT PR-AUC	CNN Backend: FMA WA (%)	AST Backend: MTAT ROC-AUC	AST Backend: MTAT PR-AUC	AST Backend: FMA WA (%)
✗	-	-	91.47 ± 0.03	46.50 ± 0.11	66.30 ± 0.31	91.65 ± 0.05	46.82 ± 0.20	67.22 ± 0.25
✓	✗	-	91.38 ± 0.08	46.32 ± 0.35	66.53 ± 0.13	91.58 ± 0.04	46.77 ± 0.12	67.10 ± 0.61
✓	✓	Bass	91.45 ± 0.08	46.48 ± 0.18	66.58 ± 0.24	91.69 ± 0.08	47.00 ± 0.24	67.89 ± 0.31
✓	✓	Drums	91.47 ± 0.07	46.68 ± 0.20	66.80 ± 0.19	91.57 ± 0.03	46.57 ± 0.20	66.11 ± 0.46
✓	✓	Other	91.59 ± 0.07	46.98 ± 0.13*	67.14 ± 0.12*	91.78 ± 0.12	47.49 ± 0.23	67.09 ± 0.33
✓	✓	Vocals	91.85 ± 0.03*	47.16 ± 0.17*	66.10 ± 0.41	91.89 ± 0.08*	47.21 ± 0.34	66.77 ± 0.14
✓	✓	Multiple	91.50 ± 0.02	46.65 ± 0.12	66.52 ± 0.21	91.87 ± 0.08*	47.31 ± 0.13*	67.40 ± 0.93

Main Takeaways:
- **Random initialization** of the U-Net **does not consistently improve** performance.
- On the other hand, using a **music source separation objective** to pre-train the U-Net can lead to **better results**, according to the pre-training source:
  - The most **consistent improvement** is achieved for **accompaniment** separation.
  - Significant boost in **auto-tagging** for **vocal separation** → vocal-specific tags?
  - Middling results for the multi-source case, contrary to self-supervised pre-training [5].
- **AST** appears to have a **higher ceiling** as a backend, in both datasets.

Impact of the training scheme (accompaniment separation, MTAT):
- **CNN: Three-stage scheme** → better results.
- **AST:** Better results via a **two-stage** scheme!
  - Pre-trained ImageNet weights provide a **robust starting point** for training.
  - **No inductive biases** → overfitting?

Backend	Pre-Training	ROC-AUC	PR-AUC
CNN	✗	91.35 ± 0.15	46.15 ± 0.44
CNN	✓	91.59 ± 0.07	46.98 ± 0.13
AST	✗	91.78 ± 0.12	47.49 ± 0.23
AST	✓	91.47 ± 0.08	46.58 ± 0.16

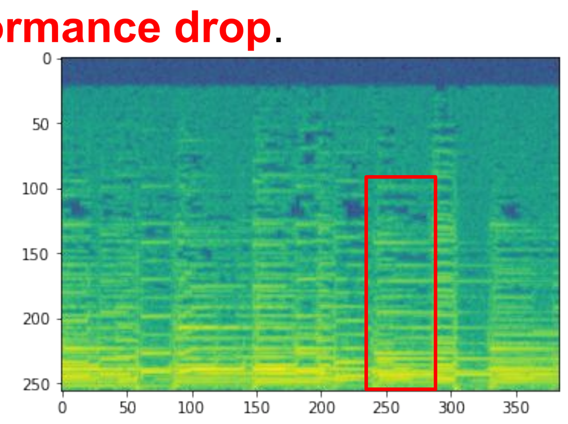
- Per-tag differential for selected MTAT tags, for various separation pre-training objectives, between the proposed framework and the bare backend:



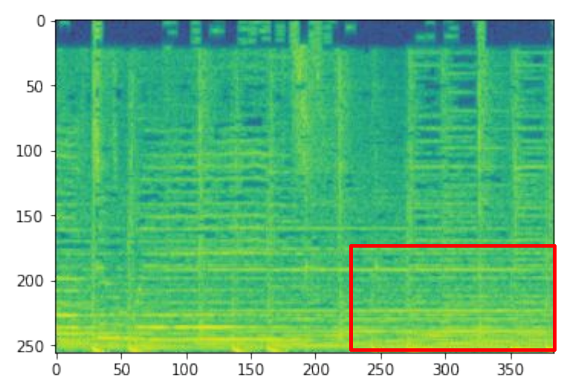
- Pre-training in particular sources can lead in **increased tag-wise performance**:
  - Tags related to **attributes** or the **appearance** of vocals mostly **benefit** from vocal pre-training.
  - Conversely, **instrument tags** report a **slight performance drop**.

Qualitative Analysis (MTAT):

- The baseline backend may mistake **high-pitched instrumental parts** for **vocals**, while the vocally pre-trained composite network is **more confident** about the absence of vocals.



- Conversely, some instances of **sustained female notes** will get detected by the vocally pre-trained network but **discarded** by the baseline backend.



5. Conclusions

- Investigated the suitability of **music source separation** as a **partial pre-training strategy** for music classification models, in two different downstream tasks.
  - Improved performance over 1) bare classification backends and 2) non pre-trained composite architectures, for **accompaniment** and **vocal** (auto-tagging) separation.
  - **Steerable** to particular attributes, according to the pre-training sources.
- **Future work:**
  - **Increase the pre-training scale** using automatically estimated source excerpts.
  - Explore **in-domain pre-trained weights** for the backend network.

References

[1] M. Velez-Vasquez et al., "Tailed U-Net: Multi-Scale Music Representation Learning", Proc. ISMIR 2022

[2] Q. Kong et al., "Decoupling Magnitude and Phase Estimation with Deep Res-U-Net for Music Source Separation", Proc. ISMIR 2021

[3] M. Won et al., "Evaluation of CNN-Based Automatic Music Tagging Models", Proc. SMC 2020

[4] Y. Gong et al., "AST: Audio Spectrogram Transformer", Proc. Interspeech 2021

[5] C. Garoufis et al., "Multi-Source Contrastive Learning from Musical Audio", Proc. SMC 2023

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/

Scan me for the source code and pre-trained models!