All-In-One Audio Transformer

Sameer Khurana, Antoine Laurent*, Hugo Riguidel (phD)*, Salima Mdhaffar*,
Mickaël Rouvier*, Adel Mounen*, Lucas Maison (phD)*, Yannick Estève*,
Yuan Gong, Yuhang He
*_Esperanto_

November 2023

1 Introduction

Inspired by the remarkable capabilities of the human auditory system, this proposal outlines the development of an AI system for universal audio processing. Emulating aspects of human auditory processing, such as its efficiency in handling diverse auditory tasks, the proposed system leverages a modular transformer architecture, like a Mixture of Experts (MoE)[2, 1] and Self-Supervised Learning (SSL)[5], to create a versatile and efficient multi-task learning framework. We aim to build an **All-In-One audio (AIO) transformer** that can perform diverse sequence generation tasks such as Automatic Speech Recognition and Translation, Sound and Music Source Separation, Speech Separation and Enhancement.

Objectives.

- **Biologically Inspired Audio Processing.** Design an AI system that mirrors the human auditory system's ability to process and interpret a wide range of auditory stimuli (Speech, Music, General Sounds).
- **Development of a Univeral Front-End Audio Encoder.** Use self-supervised learning [5] to build an audio encoder that can encode any audio signal (Speech, Music, and Sounds) into a shared representation space. This encoder is used as a Front-End for the backend MoE multi-task transformer.
- **Efficient Multitask Learning with MoE.** Implement an MoE-based transformer model to learn diverse audio processing tasks efficiently. This work would introduce the first MoE audio transformer to the research community.
- **Unified Output Encoding.** Unify the target space for different audio processing tasks using audio codecs [3]. The model is required to generate different output signals for different tasks. We will unify the output space using audio codecs to convert all target audio signals into discrete code sequences using audio codecs.
- **Comprehensive Evaluation.** Evaluate AIO transformer on several audio tasks, such as Speech Recognition, Translation, Sound/Music Source Separation, Speech Separation, and Enhancement.

2 Methodology

Self-Supervised Pre-Training. Design self-supervised learning tasks to train an encoder to project different audio signals (Speech, Sound, Music) into a shared embedding space. The features discovered by this general audio encoder will feed into the MoE transformer. The pre-trained Descript Audio Codec (DAC) encoder [3] is a good candidate.

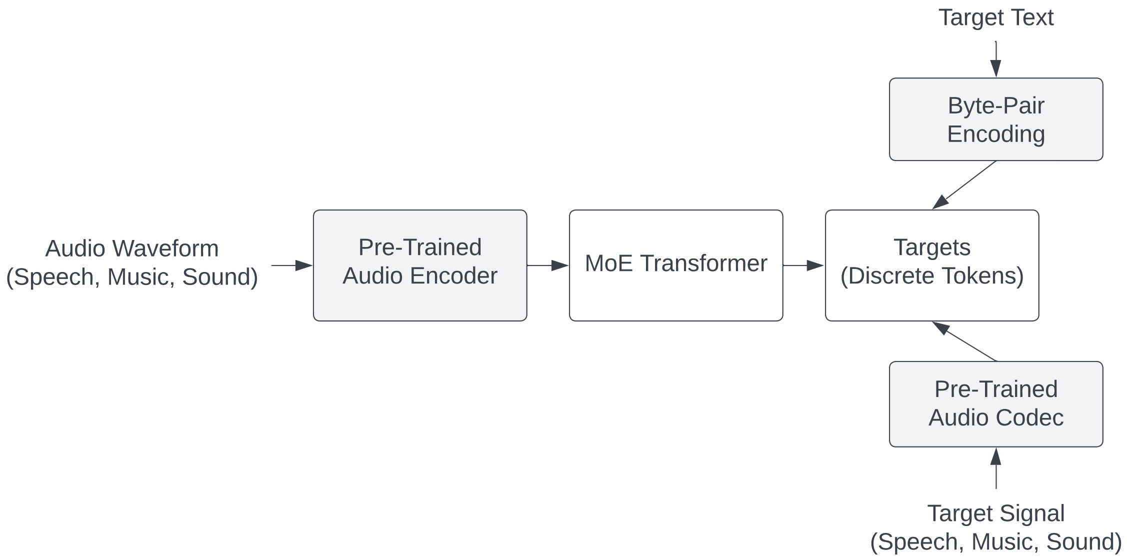
Target Text
Byte-Pair Encoding
Audio Waveform (Speech, Music, Sound)
Pre-Trained Audio Encoder
MoE Transformer
Targets (Discrete Tokens)
Pre-Trained Audio Codec
Target Signal (Speech, Music, Sound)

Figure 1: A Sketch of our proposed multitask learning framework. The gray blocks are frozen during multitask fine-tuning.

Multi-Task Fine-Tuning with MoE. Implement fine-tuning of the MoE transformer using multi-task learning, simulating the brain's ability to process various specialized audio tasks efficiently.

Justification for using MoE.

- **Task Specialization:** Each 'expert' in the MoE model can be trained to specialize in different aspects of audio processing, similar to how different areas of the auditory cortex specialize in processing speech, music, or spatial cues.
- **Scalability and Efficiency:** MoE allow the model to scale efficiently, handling many tasks without a proportional increase in computational complexity. This reflects the human brain's ability to efficiently process complex auditory information without overburdening cognitive resources.
- **Dynamic Routing:** The dynamic routing mechanism in MoE, which directs input to the most relevant experts, mimics the brain's selective attention mechanism in auditory processing.
- **Adaptability and Learning Efficiency:** The modular nature of MoE facilitates learning efficiency and adaptability, allowing the model to adapt to new tasks or changes in data distribution, much like the human auditory system adapts to new sounds or environments.
- **Cross-Task Knowledge Transfer:** The shared underlying structure in the MoE model allows for cross-task knowledge transfer, enhancing overall learning and performance, akin to how learning in one auditory domain can benefit perception in another in the human brain.

3 Evaluation.

- **Task-Specific Performance:** Assess the model's performance on individual tasks using task-specific metrics such as Word Error Rate for ASR, BLEU for Translation, and SI-SDR [4] for Enhancement and Source Separation.

- **Cross-Task Transfer Analysis:** Analyze how well the pre-trained front-end audio encoder and MoE architecture facilitate transfer across different audio tasks.

4 Expected Outcomes.

- A sophisticated AI model that efficiently processes a wide range of audio processing tasks using a modular transformer approach.
- Insight into the effectiveness of combining self-supervised pre-training with an MoE architecture in a multi-task learning environment.
- Benchmarks comparing this integrated approach against traditional and non-modular multi-task learning frameworks.

5 Potential Challenges.

- Optimizing the MoE model to ensure the right balance between shared and task-specific learning without overcomplicating the architecture.
- Ensuring the discrete code representation outputted by audio codecs retains sufficient detail and specificity for diverse tasks.
- Developing an effective mechanism for routing tasks to the appropriate experts within the MoE model.

6 Timelines and Milestones

Pre-Workshop Phase. Establish a robust baseline for the Mixture of Experts (MoE) model across a suite of audio processing tasks.

- **T-6 to T-3 Months**: 1) Assemble and preprocess a comprehensive dataset for the targeted audio tasks. 2) Finalize the MoE architecture, ensuring it's equipped to handle the diversity of audio inputs and tasks.
- **T-3 to T-1 Months:** 1) Initiate self-supervised pre-training of the audio encoder to develop a rich, shared audio embedding space. 2) Train the baseline MoE model on multiple audio processing tasks, including speech recognition, translation, sound source separation, and music source separation.
- **T-1 Month to Workshop:** 1) Conduct a preliminary evaluation of the MoE model to establish baseline performance metrics. 2) Identify and implement refinements in the model based on initial performance feedback.

Workshop Phase. Enhance the MoE model's performance and conduct an in-depth analysis of the learned representations.

- **Workshop Kickoff:** Present the baseline MoE model, outline its current capabilities and set expectations for workshop activities.
- **During Workshop:** 1) Engage in focused sprints with workshop participants to refine the model's performance. 2) Conduct collaborative sessions to dissect and understand the learned representations in the shared audio embedding space.

References

[1] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. G. Learned-Miller, and C. Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023.

[2] Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, Z. Wang, et al. M^{3}vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar. High-fidelity audio compression with improved rvqgan. *arXiv preprint arXiv:2306.06546*, 2023.

[4] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[5] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.