The ICME 2025 Audio Encoder Capability Challenge

Abstract

This challenge aims to establish a benchmark for evaluating the capabilities of audio encoders, especially in the context of multi-task learning and real-world applications. Participants are invited to submit pre-trained audio encoders that map raw waveforms to continuous embeddings. These encoders will be assessed across a diverse range of tasks spanning voice, environmental sounds, and music, with a focus on their effectiveness as components in larger systems. The challenge features two tracks: Track A for parameterized evaluation, and Track B for parameter-free evaluation. To emphasize practical relevance, the organizers are releasing several novel, open-source datasets featuring diverse audio recordings from real-world scenarios and users. This challenge provides a critical platform for benchmarking and advancing the state-of-the-art in audio encoder design.

Challenge Description

The field of audio representation learning has advanced significantly in recent years, enabling models to extract meaningful features from audio data effectively. While much of the focus has been on discrete representations and tokenization [1, 2, 3, 4, 5] recently, continuous representations remain crucial for many tasks. Unlike discrete tokens,

continuous embeddings retain the nuanced information within audio signals, leading to better performance [6] in downstream tasks like fine-grained classification, regression, and time-series analysis. Moreover, continuous audio encoders play a key role in multimodal large language models, facilitating the integration of audio with other modalities [7, 8, 9]. Models such as wav2vec2 [10], Data2vec2 [11], and Dasheng [12] have demonstrated strong performance across various audio tasks. While there are some existing benchmarks for evaluating these models, they leave room for further refinement, and a comparison with similar benchmarks can be found in Appendix A. This challenge provides a platform for participants to showcase innovative approaches in model design and data utilization, pushing the boundaries of audio representation learning. Participants are required to submit a single pre-trained encoder that processes audio waveforms and generates two outputs: a sequence of continuous embedding vectors for frame-level tasks and a fixed-dimension embedding vector for utterance-level tasks. The model should comply with an API specified by the organizers, with examples provided1,2 . The submitted models will be evaluated on diverse audio tasks, including human voice, environmental sounds, and music, using an open-source evaluation system3 . Participants can test and optimize models independently, but final rankings will be based on evaluations conducted on the organizers' servers. This challenge

aims to advance the state-of-the-art in continuous audio representation learning.

Track A

Linear Fine-Tuning on Task-Specific Data

Track B

Unparameterized Evaluation

Challenge Website: Coming Soon...

Organizers

- Xiaomi Corporation
- Dataocean AI Inc
- University of Surrey - w.wang@surrey.ac.uk
- Dr. Junbo Zhang
- Dr. Wenwu Wang - zhangjunbo1@xiaomi.com