IEEE ICME 2025 Audio Encoder Capability Challenge

Junbo Zhang[1]*, Heinrich Dinkel[1], Qiong Song[2], Helen Wang[2], Yadong Niu[1], Si Cheng[1], Xiaofeng Xin[2], Ke Li[2], Wenwu Wang[3]†, Yujun Wang[1], Jian Luan[1]

[1]Xiaomi Corporation, China [2]Dataocean AI Inc., USA [3]University of Surrey, UK

Abstract

This challenge aims to evaluate 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. This challenge provides a platform for evaluating and advancing the state-of-the-art in audio encoder design.

1 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 [13, 14, 15] 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 provided[1,2].

The submitted models will be evaluated on diverse audio tasks, including human voice, environmental sounds, and music, using an open-source evaluation system[3]. Participants may test and optimize models independently, but the final rankings will be based on the evaluations by the organizers.

1.1 Tracks

The challenge consists of two tracks, each evaluating the pre-trained models in different ways.

Track A: Linear Fine-Tuning on Task-Specific Data. A linear layer will be trained using the provided user embeddings, optimized with predefined hyperparameters for each task. This approach assesses how effectively the fixed representations can be adapted to specific tasks by training an additional linear layer, using predefined hyperparameters tailored for each task. This task evaluates the adaptability and

*zhangjunbo1@xiaomi.com

†w.wang@surrey.ac.uk

[1] https://github.com/jimbozhang/xares-template/blob/main/examples/dasheng/dasheng_encoder.py

[2] https://github.com/jimbozhang/xares-template/blob/main/examples/wav2vec2/wav2vec2_encoder.py

[3] https://github.com/jimbozhang/xares

Table 1: Datasets for Fine-tuning and evaluation. The hidden datasets marked by †. All shown tasks are evaluated in Track A. For Track B (unparameterized evaluation, see Section 1.1), a subset of Track A's utterance-level classification tasks are selected. The preprocessed version of the datasets will be provided on Zenodo.

Domain	**Dataset**	**Task Type**	**Metric**	**#**	**Track B**
Speech	Speech Commands [16]	Keyword spotting	Acc	30	✓
	LibriCount [17]	Speaker counting	Acc	11	✓
	VoxLingua107 [18]	Language identification	Acc	33	✓
	VoxCeleb1 [19]	Speaker identification	Acc	1251	✓
	LibriSpeech [20]	Gender classification	Acc	2	✓
	Fluent Speech Commands [21]	Intent classification	Acc	248	✓
	VocalSound [22]	Non-speech sounds	Acc	6	✓
	CREMA-D [23]	Emotion recognition	Acc	5	✓
	speechocean762 [24]	Phoneme pronunciation	MSE	3	✗
	ASV2015 [25]	Spoofing detection	EER	2	✓
Sound	ESC-50 [26]	Environment classification	Acc	50	✓
	FSD50k [27]	Sound event detection	mAP	200	✗
	UrbanSound 8k [28]	Urban sound classification	Acc	10	✓
	DESED [29]	Sound event detection	Segment-F1	10	✓
	FSD18-Kaggle [30]	Sound event detection	mAP	41	✗
	Clotho [31]	Sound retrieval	Recall@1	-	✗
	Inside/outside car†	Sound event detection	Acc	2	✓
	Finger snap sound†	Sound event detection	Acc	2	✓
	Key scratching car†	Sound event detection	Acc	2	✓
	Subway broadcast†	Sound event detection	Acc	2	✓
	LiveEnv sounds†	Sound event detection	mAP	18	✗
Music	MAESTRO [32]	Note classification	Acc	88	✓
	GTZAN Genre [33]	Genre classification	Acc	10	✓
	NSynth-Instruments [34]	Instruments Classification	Acc	11	✓
	NSynth-Pitch [34]	Pitches Classification	Acc	128	✓
	Free Music Archive Small [35]	Music genre classification	Acc	8	✓

effectiveness of the pre-trained models when applied to new, task-specific contexts without altering the original model parameters.

Track B: Unparameterized Evaluation. Pre-trained model embeddings will be used directly for K-nearest neighbor (KNN) classification without training. This track aims to evaluate the inherent quality of the audio representations without any fine-tuning. While this approach may not always yield the highest performance in real-world applications, it serves as a rigorous test of the fundamental representational power of the embeddings. By avoiding parameterized layers, this track provides a clear view of how well the model captures essential features of the audio data.

1.2 Training Dataset

The challenge places a significant emphasis on data collection and utilization, which is a crucial component of the competition. The organizers do not prescribe a specific training dataset. Instead, participants are free to use any data for training, as long as it meets the following conditions:

- All training data must be publicly accessible.
- Data in Table.1 must be excluded from training.

Table 2: The hidden datasets provided by the Challenge organizers. These datasets are concealed from participants.

Dataset	**Size**	**Description**
Inside/outside car	15k samples	Inside car or outside car, for security threat prevention
Finger snap sound	15k samples	Wake-up word alternative for smart speakers
Key scratching car	5k samples	Car vandalism through key scratching
Subway broadcast	125 hours	Subway announcements broadcasting
LiveEnv sound	25k samples	Environmental sounds recorded from real scenarios

1.3 Datasets for Fine-tuning and Evaluation

The datasets outlined in Table 1 encompass a diverse range of audio data across multiple domains, including human voice, environmental sounds, and music. We utilize each dataset's native train-test split to fine-tune and evaluate the participant-submitted models. All datasets are open-source, except for six hidden datasets, detailed in Table 2, which focus on real-world industrial scenarios provided by the challenge organizers and are marked with † in Table 1.

2 Registration

To participate, registration is required. Complete the registration form, accessible at [4], by the registration deadline of **April 1, 2025**. Note that this does not means the challenge starts on April 1. The challenge begins on February 7, 2025.

3 Submission Guide

Participants are required to submit a pre-trained model encapsulated within the specified API. The model should accept a single-channel audio signal, represented as a PyTorch tensor with shape $[B, T]$, where B denotes the batch size and T represents the number of samples in the time domain. The model should output a frame-level prediction of shape $[B, T', D]$, where T' can be different from the input T and D is the embedding dimension defined by the participant.

While there are no strict limitations on model size, submitted models must be able to be run successfully in a Google Colab T4 environment, where the runtime is equipped with a 16 GB NVIDIA Tesla T4 GPU, 12GB RAM.

Participants are also required to submit a technical report along with their submission.

The submission steps are as follows:

1. Clone the audio encoder template from the GitHub repository[5].
2. Implement your own audio encoder following the instructions in `README.md` within the cloned repository. The implementation must pass all checks in `audio_encoder_checker.py` provided in the repository.
3. Before the submission deadline, email the organizers [6] a ZIP file containing the complete repository. Additionally, please attach a technical report paper (PDF format) not exceeding 6 pages describing your implementation. Pre-trained model weights can either be included in the ZIP file or downloaded automatically from external sources (e.g., Hugging Face) during runtime. If choosing the latter approach, please implement the automatic downloading mechanism in your encoder implementation.

[4] https://forms.gle/VGgRQdPLs9f72UM8A

[5] https://github.com/jimbozhang/xares-template.git

[6] 2025icme-aecc@dataoceanai.com

4 Evaluation and Ranking

The performance metrics for each task are normalized to a 0-1 scale, and the final score is computed based on these normalized metrics.

4.1 Normalization of Metrics

Each task in Table 1, i.e. T_i, has an associated metric M_i (e.g., accuracy, EER, mAP, F1). To normalize these metrics, we use the following formula:

$$\hat{M}_i = \frac{M_i - M_i^{\min}}{M_i^{\max} - M_i^{\min}} \tag{1}$$

where $\hat{M}_i$ is the normalized metric for task T_i, and M_i is the raw metric value for task T_i. $M_i^{\min}$ and $M_i^{\max}$ are the worst and best possible values of the metric M_i, respectively.

For instance, the accuracy, EER, and F1 scores range from 0 to 1, so their $M_i^{\min}$ and $M_i^{\max}$ are 0 and 1, respectively; mAP ranges from 0 to 100, so for mAP tasks, $M_i^{\min} = 0$ and $M_i^{\max} = 100$.

4.2 Final Score and Ranking

The final score for each participant for Track A and Track B is calculated as the weighted average of the normalized metrics across all tasks applicable to the respective task, where the weight is determined by the size of the test set for each task. This approach ensures that tasks with larger test sets have a greater impact on the final score, reflecting their significance in evaluating the model's performance. The final scores S_A and S_B for Track A and Track B are given by:

$$S_{\text{track}} = \frac{\sum_{i=1}^{N_{\text{task}}} n_i \hat{M}_i}{\sum_{i=1}^{N_{\text{task}}} n_i} \tag{2}$$

where N_{task} is the total number of tasks applicable to the respective task, n_i is the size of the test set for task T_i, and $\hat{M}_i$ is the normalized metric for task T_i.

Participants are ranked within each track based on their final scores, S_A and S_B, respectively. The overall performance of the participants will be showcased in two separate leaderboards, one for Track A and one for Track B, to accurately reflect competencies in both parameterized and unparameterized evaluation methodologies.

5 Challenge Organizers

This Challenge is organized by teams from three institutions: Xiaomi Corporation, the University of Surrey and Dataocean AI Inc.

Xiaomi Corporation is a renowned technology company established in 2010. It is widely known for its diverse product range including smartphones, cars, tablets, laptops, wearables, and smart home devices, to form a platform of more than 800 million active devices. The company emphasizes innovation and user experience, is dedicated to fundamental technologies, blends into open-source. AI has been fully integrated into to reinforce Xiaomi's machie intelligence and service efficiency, ranging from user interaction, imaging, auto pilot, to internet sales, delivery, and service. Among them, the acoustic and speech team of the AI lab is committed to us large audio and speech models to boost the research and development in speech recognition, speech synthesis, microphone array based noise reduction, voice trigger, extraction and understanding of rich language, and acoustic measurement.

Dr. Junbo Zhang is an AI Research Scientist at Xiaomi Corporation. He earned his Ph.D. from the Institute of Acoustics at the Chinese Academy of Sciences. With years of experience in developing acoustic and speech algorithms, Dr. Zhang has made significant contributions to various fields, including speech recognition, pronunciation evaluation, speech synthesis, audio tagging, sound separation, and noise reduction. He has authored over 30 papers in prestigious journals and top-tier conferences. As a code contributor to the open-source project Kaldi, he also wrote the book "Kaldi Speech Recognition Practice",which has sold more than ten thousand copies. At Xiaomi, he was instrumental in developing and launching the company's initial speech recognition system, the wake word detection for "Xiao Ai" (Xiaomi's AI assistant), and the voiceprint recognition system. Currently, he leads several pioneering projects in the large model technology domain, pushing the boundaries of what is possible in consumer electronics.

University of Surrey The Machine Audition Lab within the Centre for Vision Speech and Signal Processing at the University of Surrey, led by Prof Wenwu Wang, is a leading research lab in audio signal processing and machine learning, consisting more than 30 researchers. They have developed several widely used audio representation models such as PANNs, AudioLDM, AudioLDM 2, AudioSep, etc. They have been contributing to the activities in Detection and Classification of Acoustic Scenes and Events (DCASE) challenges and workshops since 2013, including the organisation of two tasks of the DCASE 2024 Challenges, i.e. Task 6 - Automated Audio Captioning and Task 9 - Language-Queried Audio Source Separation.

Dr. Wenwu Wang is a Professor in Signal Processing and Machine Learning, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 300 papers in these areas. He has been recognized as a (co-)author or (co)-recipient of more than 15 accolades, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing, and an Associate Editor (2024-2026) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He is the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, a Board Member (2023-2024) of IEEE SPS Technical Directions Board, the elected Chair (2025-2027) and Vice Chair (2022-2024) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He has been on the organising committee of INTERSPEECH 2022, IEEE ICASSP 2019 & 2024, IEEE MLSP 2013 & 2024, and SSP 2009. He is Technical Program Co-Chair of IEEE MLSP 2025. He has been an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops.

Dataocean AI Inc. is a global data collection and labeling services provider that combines technology with a diverse network of millions data contributors, scientists, and engineers. The company delivers cutting-edge data solutions across multiple domains, including text, audio, image, and multimodal for foundation models or GenAI applications. With over 1,600 off-the-shelf datasets and a proven track record of delivering thousands of customized data projects, Dataocean AI have been trusted by of over 1,000 global AI leading enterprises and institutions. The company cover more than 200 languages around the world. Its self-developed data platform ensures precision and efficiency in tasks such as collection, cleansing, labeling and evaluation. With nearly two decades of experience, Dataocean AI has established itself as a trusted partner in the AI ecosystem, consistently delivering excellence and earning global recognition.

Helen Wang is marketing head of Dataocean AI, mainly responsible for open data community and challenge operation.

6 Challenge Schedule

The Challenge will follow this schedule:

- February 7, 2025: Challenge announcement.
- April 30, 2025: Submission deadline.
- May 27, 2025: Results announcement.

References

[1] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[2] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, “Finite scalar quantization: Vq-vae made simple,” *arXiv preprint arXiv:2309.15505*, 2023.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved rvqgan,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, “Snac: Multi-scale neural audio codec,” *arXiv preprint arXiv:2410.14411*, 2024.

[5] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, “Moshi: a speech-text foundation model for real-time dialogue,” Tech. Rep., 2024. [Online]. Available: https://arxiv.org/abs/2410.00037

[6] D. Wang, M. Cui, D. Yang, X. Chen, and H. Meng, “A comparative study of discrete speech tokens for semantic-related tasks with large language models,” *arXiv preprint arXiv:2411.08742*, 2024.

[7] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.

[8] Z. Xie and C. Wu, “Mini-omni: Language models can hear, talk while thinking in streaming,” *arXiv preprint arXiv:2408.16725*, 2024.

[9] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, “Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation,” *arXiv preprint arXiv:2411.18138*, 2024.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” in *International Conference on Machine Learning.* PMLR, 2022, pp. 1298–1312.

[12] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, “Scaling up masked audio encoder learning for general audio classification,” in *Interspeech 2024*, 2024.

[13] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, “HEAR: Holistic evaluation of audio representations,” in *NeurIPS 2021 Competitions and Demonstrations Track.* PMLR, 2022, pp. 125–145.

[14] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, “SUPERB: Speech processing universal performance benchmark,” *Interspeech 2021*, 2021.

[15] P. Mousavi, L. Della Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, “DASB–discrete audio and speech benchmark,” *arXiv preprint arXiv:2406.14294*, 2024.

[16] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” *arXiv preprint arXiv:1804.03209*, 2018.

[17] F.-R. Stöter, S. Chakrabarty, E. Habets, and B. Edler, “Libricount, a dataset for speaker count estimation,” 2018.

[18] J. Valk and T. Alumäe, “Voxlingua107: a dataset for spoken language recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2021, pp. 652–658.

[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2015, pp. 5206–5210.

[21] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[22] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[24] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Proc. Interspeech 2021*, 2021.

[25] T. Kinnunen, Z. Wu, E. Nicholas Evans, and J. Yamagishi, "Automatic speaker verification spoofing and countermeasures challenge (asvspoof 2015) database," 2018.

[26] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[27] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[28] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[29] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[30] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[31] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 736–740.

[32] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019.

[33] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[34] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.

[35] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

Appendices

A Related Work

Our challenge broadens the scope by including non-speech related tasks, enabling a more comprehensive evaluation of audio encoders, which are pivotal in both continuous and discrete audio processing contexts. Here, we discuss three of the existing benchmarks, highlighting the unique contributions and improvements of our proposed challenge.

A.1 HEAR: Holistic Evaluation of Audio Representations

Our proposed challenge is strongly inspired by the HEAR benchmark [13], which assesses audio representations across environmental sound and music tasks. While HEAR provides an excellent foundation, our challenge introduces several enhancements:

Diverse task set: HEAR comprises 19 tasks in total, 17 of which are unique, while two tasks differ only in their available training data. While the tasks in HEAR encompass various application scenarios for sound event detection and music processing, they lack variety in human voice processing. Our challenge offers a more comprehensive and balanced distribution of tasks across human voice, music, and environmental sound domains, leveraging a suite of open-source datasets that reflect real-world scenarios and user experiences, including unique datasets such as car scratching, inside/outside car environments.

Focus on real-world applications: Some tasks in HEAR, although interesting, may have limited applications and high variance during testing (e.g., Gunshot Triangulation and Beehive) due to the factors such as small sample sizes, which have led to many follow-up works discarding those tasks. We seek to balance task variety, real-world impact, and robust performance estimation, ensuring the evaluated representations are relevant to industrial use, providing reliable performance metrics.

Evaluation methods: In addition to linear projection, we utilize unparameterized methods for classification. This evaluation aims at investigating the use of features for cases such as unsupervised clustering.

Efficient system: We propose an open-sourced, efficient evaluation system, incorperating a simple pipline that can be run without any prequisites.

A.2 SUPERB: Speech processing Universal PERformance Benchmark

SUPERB [14] and its derivatives primarily focus on speech processing tasks using self-supervised learning (SSL) representations. In recent years, SUPERB also included additional tasks such as emotion recognition and sound codecs, but notably, it does not include environmental audio or music related tasks.

Our challenge broadens this scope with the inclusion of non-speech related tasks (environmental audio, music), enabling a more comprehensive evaluation of audio representations.

A.3 DASB: Discrete Audio and Speech Benchmark

DASB [15] benchmarks discrete audio tokens across various tasks, mainly focuses on the speech domain. While discretization is an important research field, continuous representations offer complementary advantages. Continuous representations directly addresses the need for robust audio encoders in multimodal applications, where continuous embeddings are often preferred for seamless integration and efficient processing [6, 9]. The output of this challenge can be used to complement discrete representation research by, for example, injecting general semantic information into codecs [5], or evaluating the loss of information during the discretisation process. Our challenge prioritizes established methods for continuous representations.