

Discrete vs. Continuous: A Comprehensive Study of Unified Audio Understanding in LALMs

Jing Peng^{1,*}, Zichao Nie^{1,*}, Zhisheng Zhang¹, Jingran Xie¹, Zhiyong Wu^{1,**}

¹ Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

zywu@se.cuhk.edu.hk

Abstract

Large Audio Language Models (LALMs) utilize either continuous features or discrete tokens, yet the optimal representation paradigm for general audio understanding remains debated. Existing benchmarks often focus on narrow domains or evaluate encoders outside LALM contexts. To address these gaps, we systematically evaluate continuous and discrete representations across speech, sound and music. Utilizing our UniARC framework with dual evaluation strategies across model scales from SmolLM2-135M to Llama-3-8B, we analyze the dynamic relationships of data volume, model capacity, and computational efficiency. Our results reveal the pivotal role of semantic constraints in tokenization for audio understanding and demonstrate that scaling backbones fail to compensate for information loss in audio representation, especially in data-limited tasks. These findings offer practical guidance for balancing semantic density, fidelity, and efficiency in future LALMs.

Index Terms: large audio language models, audio encoders, benchmarking, unified audio understanding

1. Introduction

In recent years, Large Audio Language Models (LALMs) have advanced significantly [1, 2]. The underlying representation schemes of these models have undergone several developments, evolving from continuous features to discrete tokens, and currently towards hybrid approaches where both paradigms coexist or are integrated [3, 4, 5]. Discrete representations, owing to their structural consistency with text tokens, have greatly simplified the process of cross-modal alignment [6, 7]. However, the quantization bottleneck inherent in discretization often leads to the loss of crucial acoustic details, prompting researchers to re-examine the value of continuous features.

While numerous studies have emerged in this field, they often exhibit specific limitations. Some works focus on comparing semantic discrete tokens with continuous features, neglecting acoustic-based discrete representations [3, 4]. Others conduct evaluations outside the context of large language models (LLMs) [8, 9], or concentrate on discrete tokens without a comprehensive benchmark against continuous features [10, 11]. Notably, most research focuses narrowly on speech, overlooking general audio understanding [3]. Furthermore, existing evaluations are often confined to constrained or “toy” settings, neglecting the complex interplay between representation paradigms, model capacity, and data scalability [4]. Therefore, a systematic comparison addressing both cross-domain capabilities and real-world scaling dynamics is urgently needed.

To bridge these gaps, we propose the **Unified Audio Representation Comparison (UniARC)** framework, specifically designed for audio understanding. In this study, we unify the terminology by referring to audio representation modules as encoders, encompassing continuous features from Self-Supervised Learning (SSL), and discrete tokens derived from either clustering-based or neural codec-based paradigms. To ensure the robustness of our findings and simulate real-world deployment constraints, UniARC evaluates encoders under two distinct strategies: (1) a parameter-efficient fine-tuning setup leveraging SmolLM2-135M/360M [12], and (2) a frozen-backbone probing pipeline using Llama-3-1B/8B [13]. Beyond a direct comparison of encoding paradigms, we leverage the UniARC to investigate the multi-dimensional interplay between data volume, model capacity, and computational efficiency.

Our analysis reveals that regardless of the representational form, the efficacy of audio encoders for understanding tasks is primarily determined by the richness of their semantic information. Specifically, we find that discrete tokens can surpass continuous features when they incorporate strong semantic constraints, whereas high-fidelity representations focused purely on signal reconstruction often fail in understanding tasks. These findings suggest that the semantic compatibility of audio representations with the language model’s reasoning space is the primary determinant of performance in LALMs. Furthermore, we demonstrate that for audio understanding, scaling the language backbone cannot effectively compensate for insufficient representational quality. Our findings reveal that the front-end features effectively set a performance ceiling that remains unchanged or even declines despite increases in model scale. The primary contributions of this work are as follows:

- We comprehensively evaluate across three diverse audio domains, i.e., speech, general sound, and music, to compare LLM-based continuous and discrete representations specifically for audio understanding tasks.
- We analyze how representation paradigms, model scales and data scalability interact, demonstrating that for audio understanding, aligning audio encoders with textual semantics is more decisive than scaling LLM backbones.
- We identify the task-dependent strengths of different paradigms within the understanding context, providing guidelines for balancing semantic density, acoustic fidelity, and training efficiency in future LALMs designs.

2. Evaluation Methodology

2.1. Evaluation Framework

Building upon the open-source XARES-LLM project [9], we propose **UniARC**, a modular framework that systematically

*These authors contributed equally.

**indicates the corresponding author.

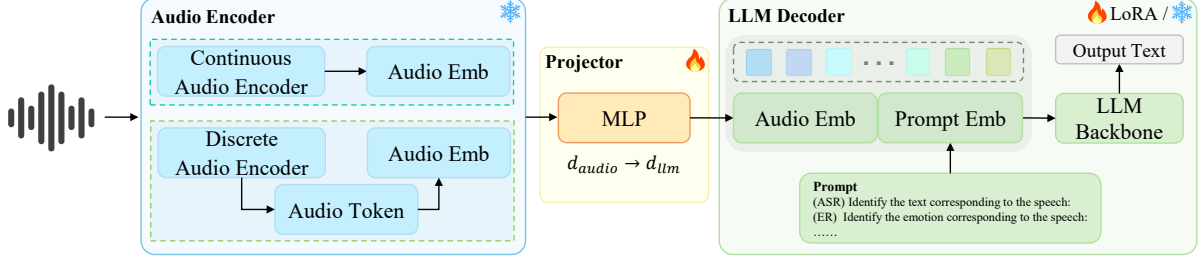


Figure 1: *The UniARC evaluation framework. The modular design supports flexible configurations of encoders, projectors, and LLM backbones. Fire and snowflake icons represent the training and frozen status of components, respectively.*

evaluates audio encoders via an instruction-tuning paradigm, as shown in Figure 1. Audio embeddings are appended to task-specific prompts, guiding the LLM within a unified sequence-to-sequence generation framework. To simulate diverse real-world deployment scenarios, we implement two strategies.

The first scenario targets domain-specific experts via parameter-efficient fine-tuning. For lightweight applications where backbone updates are computationally feasible, we employ SmolLM2 (135M/360M) with a LoRA-based instruction-tuning strategy [14], which updates only the adapters while mapping audio embeddings via a projector, maximizing performance on specific tasks within limited parameter budgets.

The second scenario evaluates general-purpose LLMs via frozen-backbone probing. To preserve inherent reasoning capabilities and minimize confounding variables from model adaptation, we utilize Llama-3 (1B/8B) with frozen parameters. To bridge the modality gap without updating the LLM, we employ a robust two-layer MLP projector and a 10-frame temporal concatenation to handle long acoustic sequences efficiently. This design isolates the adaptive capacity of LLM, directly exposing the intrinsic semantic quality of frontend encoders.

2.2. Processing of Audio representations

To ensure fair comparison, UniARC standardizes the conversion of diverse audio inputs into a unified format compatible with the LLM projector.

For **continuous representations**, raw audio is processed by encoders into high-dimensional dense embeddings, preserving representational density within a continuous latent space. For **discrete representations** derived from semantic clustering (K-means) or neural codecs, inputs are first converted to token indices and then mapped to codebook embeddings, effectively transforming discrete units into a continuous vector space.

In both paradigms, the resulting embeddings are processed by a projector to align dimensions with the LLM’s text space, enabling end-to-end differentiable training. Finally, task-specific prompts guide the model across all downstream tasks.

3. Experiments

3.1. Encoder Selection

We include a diverse set of encoders to cover different representation paradigms. For continuous features, we select HuBERT [15], WavLM [16], Wav2Vec 2.0 [17], and Whisper [18] to represent SSL and weakly supervised approaches.

For discrete representations, we investigate two primary categories of encoders based on discretization mechanisms: 1) Clustering-based tokens, derived via K-means (1,000 centroids)

on SSL features (HuBERT, WavLM) to prioritize high-level semantic abstraction [3]; and 2) Neural Codec tokens, learned through end-to-end reconstruction. For the latter, we evaluate DAC [19] and WavTokenizer [20], alongside SpeechTokenizer [21], a specialized codec that integrates semantic distillation to decouple semantic content from acoustic details.

To ensure consistent feature quality, we operate all encoders at 16 kHz, with the exception of WavTokenizer, which utilizes its native 24 kHz configuration. For continuous encoders (HuBERT, WavLM), we consistently extract the last hidden state. For clustering-based discrete tokens, we set K-means clusters to 1000 for SmolLM2-135M/360M lightweight decoders with LoRA fine-tuning, balancing speech unit complexity, training stability, inference efficiency and computational cost [3].

3.2. Tasks and Datasets

To align with the two evaluation strategies defined in UniARC, we adopt a stratified task selection protocol covering Speech, Sound, and Music. A comprehensive overview of all evaluated datasets, task types, and metrics is summarized in Table 1.

Table 1: *Overview of datasets, tasks, and metrics utilized in the UniARC framework. **Mode** denotes the evaluation setup: Tuned (T) with SmolLM2, Frozen (F) with Llama-3, or Both.*

Dataset	Task Type (Abbr.)	Metric	Mode
<i>Speech</i>			
LibriSpeech (LS) [22]	Speech Recognition (ASR)	WER/iWER	Both
AISHELL-1 [23]	Speech Recognition (ASR)	iCER	T
Fluent Speech Commands [24]	Intent Classification (IC)	Acc	T
SLURP [25]	Intent Classification (IC)	Acc	F
Speech Commands V1 [26]	Keyword Spotting (KS)	Acc	T
CREMA-D [27]	Emotion Recognition (ER)	Acc	Both
IEMOCAP [28]	Emotion Recognition (ER)	Acc	F
VoxCeleb1 [29]	Speaker Identification (SI)	Acc	T
VoxLingua33 [30]	Language Identification (LID)	Acc	T
LibriCount [31]	Speaker Counting (SC)	Acc	T
VocalSound [32]	Human Sound Classification (HSC)	Acc	T
ASVspoof2015 [33]	Spoofing Detection (SD)	Acc	T
<i>Sound</i>			
ESC-50 [34]	Environment Classification (EnvC)	Acc	T
UrbanSound8K [35]	Urban Sound Classification (USC)	Acc	Both
FSD50K [36]	Sound Event Detection (SED)	mAP	T
FSDKaggle2018 [37]	Sound Event Detection (SED)	mAP	T
Clotho [38]	Sound Captioning (SCap)	FENSE	Both
MECAT [39]	General Caption (GCap)	DATE	T
<i>Music</i>			
GTZAN [40]	Genre Classification (GC)	Acc	Both
Free Music Archive (FMA) [41]	Genre Classification (GC)	Acc	T
NSynth [42]	Instrument Classification (InstC)	Acc	T
Song Describer (SDD) [43]	Music Captioning (MCap)	FENSE	Both

For parameter-efficient fine-tuning evaluations, we utilize a broad suite of instruction-tuning tasks adhering to XARES-LLM to establish a fundamental baseline. Conversely, for

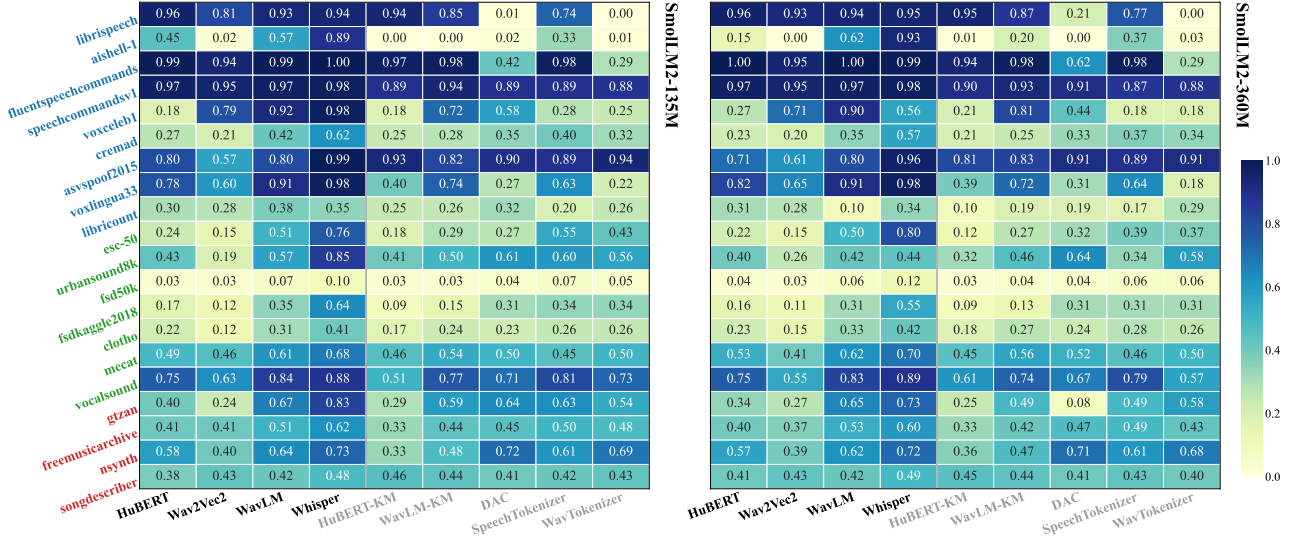


Figure 2: Performance of representation paradigms within the UniARC fine-tuning pipeline (KM stands for K-means).

frozen-backbone probing evaluations, we strategically curate a representative subset of core tasks. Notably, SLURP and IEMOCAP are introduced here to rigorously probe encoders’ intrinsic semantic reasoning.

We employ metrics tailored to each task type: Accuracy (Acc) for single-label classification, mAP for multi-label event detection, FENSE [44] for audio/music captioning to evaluate semantic fluency, and DATE [39] for general captioning. For speech recognition, standard Word/Character Error Rates (WER/CER) are calculated. To facilitate intuitive visualization of fine-tuning evaluations, WER/CER are inverted into Inverse Word/Character Error Rate (iWER/iCER), ensuring that higher scores consistently indicate better performance.

3.3. Experimental Setup

Our experiments are conducted on NVIDIA A100 GPUs using PyTorch. For Fine-Tuning Strategy, we apply LoRA ($r = 8, \alpha = 32$) to LLM backbone on single GPU, maintaining consistent default hyperparameters on a single GPU. For Frozen-Probing Strategy, we freeze LLM and optimize only projector using AdamW optimizer [45] with a learning rate of 1×10^{-5} across four GPUs. To ensure fair comparison of representation quality, all task-specific configurations within each strategy are kept strictly identical across different audio encoders.

4. Results and Analysis

We comprehensively evaluate continuous features and discrete tokens across diverse tasks under UniARC framework. Results from the parameter-efficient fine-tuning strategy are visualized in Figure 2, where datasets are color-coded by domain (speech in blue, sound in green, music in red) on y-axis, and encoders are color-coded by representation type (continuous in black, discrete in gray). These results are numerically summarized in Table 2 by averaging the scores within each domain. Conversely, detailed findings from the strict frozen-backbone probing strategy are presented in Table 3. The following subsections systematically analyze these results regarding representation paradigms, scaling effects, and training efficiency.

Table 2: Encoder Performance on Different Backbones. Higher scores indicate better performance ($\uparrow$).

Encoder	Speech	Sound	Music	Overall
<i>SmolLM2-135M</i>				
HuBERT	0.634	0.333	0.445	0.491
Wav2Vec2	0.575	0.243	0.370	0.418
WavLM	0.766	0.466	0.561	0.620
Whisper	0.858	0.618	0.666	0.735
HuBERT+Kmeans	0.534	0.263	0.356	0.403
WavLM+Kmeans	0.620	0.361	0.488	0.503
DAC	0.419	0.381	0.554	0.432
SpeechTokenizer	0.593	0.440	0.541	0.529
WavTokenizer	0.353	0.408	0.535	0.409
<i>SmolLM2-360M</i>				
HuBERT	0.603	0.331	0.432	0.474
Wav2Vec2	0.589	0.238	0.368	0.422
WavLM	0.732	0.438	0.555	0.594
Whisper	0.807	0.561	0.635	0.687
HuBERT+Kmeans	0.502	0.304	0.366	0.407
WavLM+Kmeans	0.665	0.407	0.457	0.533
DAC	0.435	0.391	0.417	0.416
SpeechTokenizer	0.583	0.375	0.505	0.494
WavTokenizer	0.347	0.378	0.523	0.393

4.1. Analysis of Representation Paradigms

Based on the experimental results in Figure 2 and Table 3, we find that whether an encoder is discrete or continuous is not the primary factor governing performance; rather, semantic compatibility fundamentally dictates its efficacy for LLM-based understanding, far outweighing raw signal fidelity.

This principle is evident among continuous encoders. Weakly supervised Whisper aligns features with textual semantics, leads across most benchmarks. Conversely, Wav2Vec2’s self-supervised focus on low-level acoustics limits its efficacy in non-speech semantic tasks. This confirms that while SSL models excel at signal modeling, Whisper’s text-aligned features are more effective for high-level LLM reasoning.

A similar trend governs discrete representations, explaining

Table 3: Experimental results across Speech, Sound, and Music domains. Best results within each model scale (1B/8B) are **bolded**.

Model	Speech					Sound		Music	
	ASR (WER % ↓)		ER (ACC % ↑)		IC (ACC % ↑)	USC (ACC % ↑)	SCap (FENSE ↑)	GC (ACC % ↑)	MCap (FENSE ↑)
	LS-960 (clean/other)	LS-100 (clean)	IEMOCAP	CREMA-D	SLURP	UrbanSound8K	Clotho	GTZAN	SDD
<i>1B Models</i>									
HuBERT	2.31/4.62	3.45	52.30	50.12	71.23	55.56	0.196	30.69	0.453
SpeechTokenizer	40.62/70.08	80.82	57.70	51.97	41.82	60.33	0.143	51.03	0.476
WavLM	4.02/6.55	6.00	64.06	64.62	68.67	69.89	0.164	61.72	0.545
DAC	171.02/177.48	208.46	47.46	36.24	7.90	51.14	0.211	35.17	0.478
<i>8B Models</i>									
HuBERT	2.24/4.38	3.13	47.62	52.21	70.13	48.03	0.129	45.19	0.507
SpeechTokenizer	18.33/38.75	83.40	56.49	54.30	41.47	71.92	0.083	51.72	0.392
WavLM	2.83/5.57	5.49	62.29	67.81	65.39	71.45	0.177	59.31	0.523
DAC	147.35/148.41	166.57	46.01	29.85	7.42	48.86	0.101	36.55	0.501

the divergence among tokenization paradigms. While K-means quantization typically lags behind continuous features due to information loss, encoders that incorporate explicit semantic constraints defy this trend. Notably, SpeechTokenizer consistently surpasses its continuous teacher, HuBERT, in specific tasks such as ER. This demonstrates that semantically-aware discretization can act as a regularizer, distilling high-density semantic features optimally tailored for LLM comprehension.

Conversely, reconstruction-oriented codecs (e.g., DAC and WavTokenizer) struggle with semantic understanding tasks despite their superior fidelity. The preservation of low-level acoustic details, such as phase and microstructural information, may introduce irrelevant noise that complicates the mapping of audio features into the LLM’s semantic space. However, this high fidelity remains valuable for timbre-sensitive tasks like music classification (NSynth), where fine-grained frequency details are more critical than high-level semantic abstractions.

4.2. Scaling, Data Diversity, and Efficiency

Backbone Scaling and the Inverse Scaling Phenomenon. We analyze scaling behavior separately within each model family: SmoLLM2 (135M to 360M) and Llama-3 (1B to 8B). As shown in Tables 2 and 3, increasing capacity does not yield universal improvements. For SmoLLM2, scaling from 135M to 360M provides marginal gains or fluctuations, suggesting capacity saturation for the evaluated encoders. For Llama-3, while scaling laws [46] hold for the data-abundant ASR task (LS-960), performance regression occurs in many data-limited scenarios. Overall, across both frameworks, performance decreased in approximately 53% of the evaluated encoder-task configurations. These results indicate that for audio understanding, scaling backbones cannot compensate for low feature quality or limited training data. While we focus on representations rather than projector design, our simple MLP might act as a bottleneck limiting these scaling benefits, which future context-aware architectures could mitigate by preventing surjective mappings.

Impact of Pre-training Data Diversity. Encoder understanding performance depends on the breadth of pre-training. While semantic compatibility governs overall efficacy, the task-level breakdowns in Tables 2 and 3 further elucidate the critical impact of pre-training data diversity. As observed in Table 2, encoders pre-trained on diverse, multi-domain data (Whisper and WavLM) show a clear advantage over those trained exclusively on speech (HuBERT and SpeechTokenizer). Furthermore, results in Table 3 provide a more nuanced perspective: although HuBERT excels in tasks heavily reliant on speech-

specific semantics, such as ASR and IC—sometimes even surpassing WavLM—the latter maintains its dominance across most of other general-audio categories. This gap underscores that multi-domain pre-training is essential for universal audio understanding, since instruction-tuning alone cannot bridge representation gaps inherited from narrow pre-training data.

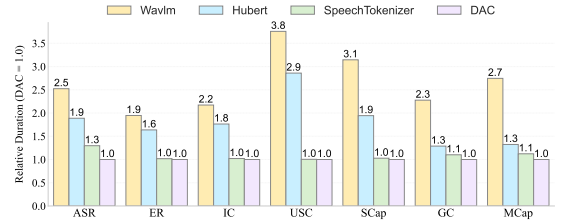


Figure 3: Total training time until convergence for encoders.

Computational Efficiency and Training Convergence. As illustrated in Figure 3, the computational cost of continuous features is a practical challenge. The sequence processing latency for continuous encoders (HuBERT and WavLM) significantly exceeds that of discrete tokenizers (SpeechTokenizer and DAC) across all benchmarks. Notably, discrete representations converge much faster during modality alignment, with validation performance stabilizing in fewer training steps. Despite the performance edge of continuous features in certain domains, the rapid convergence and lower overhead of discrete tokens make them an attractive paradigm for resource-constrained research and rapid system deployment.

5. Conclusion

This paper systematically evaluates continuous and discrete representations for audio understanding across speech, sound, and music. Results reveal that discrete and continuous encoders exhibit distinct and complementary strengths across task requirements. We highlight the necessity of incorporating semantic constraints into discrete representations. Furthermore, our study underscores multi-domain pre-training data in achieving universal audio intelligence. Crucially, for audio understanding, backbone scaling fails to consistently compensate for an encoder’s low semantic density, often leading to performance regression in data-limited tasks. These insights guide future LALM development, suggesting a focus on balancing semantics and fidelity through multi-domain pre-training.

6. Acknowledgments

This work is supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (JCYJ20220818101014030).

7. Generative AI Use Disclosure

During the preparation of this manuscript, the authors used generative AI tools exclusively for the purpose of language editing and manuscript polishing to improve readability. These tools were not used to generate any core scientific ideas, experimental data, or technical contributions. All authors have thoroughly reviewed and approved the final version of the manuscript, and assume full responsibility for the integrity and entirety of its content.

8. References

- [1] 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.
- [2] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. MA, and C. Zhang, “Salmonn: Towards generic hearing abilities for large language models,” in *International Conference on Learning Representations (ICLR)*.
- [3] D. Wang, J. Li, M. Cui, D. Yang, X. Chen, and H. Meng, “Speech discrete tokens or continuous features? a comparative analysis for spoken language understanding in speechllms,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2025, pp. 24924–24935.
- [4] Y. Xu, S.-X. Zhang, J. Yu, Z. Wu, and D. Yu, “Comparing discrete and continuous space llms for speech recognition,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2024, pp. 2509–2513.
- [5] Y. Zhou, G. Zeng, X. Liu, X. Li, R. Yu, Z. Wang, R. Ye, W. Sun, J. Gui, K. Li *et al.*, “Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning,” *arXiv preprint arXiv:2509.24650*, 2025.
- [6] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, “Audiolm: A language modeling approach to audio generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 2523–2533, 2023.
- [7] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, “Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,” in *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP Findings)*, 2023, pp. 15 757–15 773.
- [8] 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.
- [9] J. Zhang, H. Dinkel, Y. Niu, C. Liu, S. Cheng, A. Zhao, and J. Luan, “X-ares: A comprehensive framework for assessing audio encoder performance,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2025, pp. 4868–4872.
- [10] P. Mousavi, G. Maimon, A. Moumen, D. Petermann, J. Shi, H. Wu, H. Yang, A. Kuznetsova, A. Ploujnikov, R. Marxer *et al.*, “Discrete audio tokens: More than a survey!” *Transactions on Machine Learning Research (TMLR)*, 2025.
- [11] Y. Guo, Z. Li, H. Wang, B. Li, C. Shao, H. Zhang, C. Du, X. Chen, S. Liu, and K. Yu, “Recent advances in discrete speech tokens: A review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- [12] L. B. allal, A. Lozhkov, E. Bakouch *et al.*, “SmolLM2: When smol goes big — data-centric training of a fully open small language model,” in *Second Conference on Language Modeling (COLM)*, 2025.
- [13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” *International Conference on Learning Representations (ICLR)*, p. 3, 2022.
- [15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 3451–3460, 2021.
- [16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, p. 1505, 2022.
- [17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 12 449–12 460.
- [18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 28 492–28 518.
- [19] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved rvqgan,” 2023, pp. 27 980–27 993.
- [20] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li *et al.*, “Wavtokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling,” in *International Conference on Learning Representations (ICLR)*.
- [21] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Speechtokenizer: Unified speech tokenizer for speech language models,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [24] M. Ravanelli, V. Tomar, P. Ignoto, Y. Rachmad, and L. Lugosch, “Speech model pre-training for end-to-end spoken language understanding,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2019, pp. 814–818.
- [25] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, “Slurp: A spoken language understanding resource package,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.
- [26] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” *arXiv preprint arXiv:1804.03209*, 2018.
- [27] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced emotional multi-modal actors dataset,” *IEEE Transactions on Affective Computing (TAC)*, pp. 377–390, 2014.
- [28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation (LRE)*, pp. 335–359, 2008.

- [29] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language (CSL)*, vol. 60, p. 101027, 2019.
- [30] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658, 2020.
- [31] F.-R. Stöter, S. Chakrabarty, E. Habets, and B. Edler, "Libricount, a dataset for speaker count estimation," 2018.
- [32] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.
- [33] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 11, no. 4, pp. 588–604, 2017.
- [34] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *ACM international conference on Multimedia (MM)*, 2015, pp. 1015–1018.
- [35] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM international conference on Multimedia (MM)*, 2014, pp. 1041–1044.
- [36] 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 (TASLP)*, vol. 30, pp. 829–852, 2021.
- [37] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset and baseline," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2018.
- [38] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.
- [39] Y. Niu, T. Wang, H. Dinkel, X. Sun, J. Zhou, G. Li, J. Liu, X. Liu, J. Zhang, and J. Luan, "Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks," *arXiv preprint arXiv:2507.23511*, 2025.
- [40] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in *International Symposium on Music Information Retrieval (ISMIR)*, 2001.
- [41] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2016.
- [42] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning (ICML)*. JMLR. org, 2017, pp. 1068–1077.
- [43] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos *et al.*, "The song describer dataset: A corpus of audio captions for music-and-language evaluation," in *Workshop on Machine Learning for Audio, Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems, 2023.
- [44] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 981–985.
- [45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.
- [46] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.