Codec-SUPERB @ SLT 2024: Codec Speech processing Universal PERformance Benchmark

Codec-SUPERB team

Abstract

We present the Codec-SUPERB challenge at SLT 2024, which aims to enable a fair comparison of all current existing codec models and stimulate the development of more advanced codecs. In recent years, significant developments in codec models have been witnessed and numerous high-performance neural audio codecs have been developed. The ideal neural audio codec models should preserve content, paralinguistics, speakers, and audio information under low *bitrate* measured by thousand bits per second (*kbps*). However, the question of which codec achieves optimal audio information preservation remains unanswered, as in different papers, models are evaluated on their selected experimental settings. The challenge, built upon the Codec-SUPERB benchmark [1], gathers representative speech applications and objective metrics to comprehensively measure the capacity of neural audio codec models to preserve content, paralinguistics, speakers, and audio information under different bitrates.

Index Terms: Neural audio codec, Codec-SUPERB

1. Introduction

Neural audio codecs are initially introduced to compress audio data into compact codes to reduce transmission latency. Researchers recently discovered the potential of codecs as suitable tokenizers for converting continuous audio into discrete codes, which can be employed to develop audio language models (LMs) [1]. The neural audio codec's dual roles in minimizing data transmission latency and serving as tokenizers underscore its critical importance. In recent years, significant developments in codec models have been witnessed [2–17]. Numerous high-performance neural audio codecs have been developed within the current three years. The ideal neural audio codec models should preserve content, paralinguistics, speakers, and audio information. However, the question of which codec achieves optimal audio information preservation remains unanswered, as in different papers, models are evaluated on their selected experimental settings. There's a lack of a challenge to enable a fair comparison of all current existing codec models and stimulate the development of more advanced codecs. To fill this blank, we propose the Codec-SUPERB challenge.

2. Challenge overview

The goal of this challenge is to encourage innovative methods and a comprehensive understanding of the capability of codec models. This challenge will conduct a comprehensive analysis to provide insights into codec models from both application and signal perspectives [2]. We prepare an easy-to-follow script [2] to participants, which includes open dataset download, environment installment, and evaluation.

2.1. Dataset

To facilitate the development of codec techniques and fair comparison over challenge submissions, we plan to have two datasets for each task: open set and hidden set. The hidden set will always be hidden for participants. The open set serves as the development set. Participants can use the open set to evaluate and develop their models. The final results are evaluated based on the hidden set.

2.1.1. Open set

We list the datasets we used in this challenge below. To resolve the license issue, we replace and remove some of the datasets in the original paper [2]. We also only do sub-sampling to make the evaluation faster.

QUESST 2014 dataset [18] contains 23 hours of spoken documents in six low-resource languages, encoded at 8 KHz and 16-bit resolution, sourced from various speech types and acoustic environments.

Fluent Speech Commands [19] comprises 30,043 spoken utterances from 97 individuals, recorded as single-channel .wav files at a 16 kHz sampling rate. Each file captures a distinct utterance intended for the operation of smart-home devices or a virtual assistant. For example, an utterance might be "turn on the light in the bedroom." We use the test set for codec evaluation.

LibriSpeech [20] is a highly utilized corpus of English speech data, comprising roughly 1000 hours of audio recordings. These recordings are characterized by a reading style, as they consist of utterances read from audiobooks. We use test-clean and test-other sets for codec evaluation.

Audio SNIPS [21] utilizes a text-to-speech (TTS) system to synthesize the SNIPS dataset into utterances with different speakers and accents. The dataset is designed for speech recognition and natural language understanding simultaneously. We use test and valid splits for codec evaluation.

VoxCeleb1 [22] is an audio-visual dataset featuring short segments of human speech sourced from interview videos on YouTube. It includes over a million real-world utterances from more than 6000 speakers. We use the test set for evaluation.

Libri2Mix [23] is a synthesized corpus featuring mixtures of two speakers' speech intertwined with background noise. The speech segments are sourced from LibriSpeech, while the ambient noise is taken from the WHAM! dataset. The corpus is

[1] https://codecsuperb.github.io/

[2] Codec-SUPERB script

Speech dataset	Features	app	obj
Librispeech	diverse speaker, read audiobooks	✓	✓
VoxCeleb1	diverse speaker, celebrities on YouTube	✓	✓
QUESST	multi-lingual, low resource language		✓
VoxLingua107 Top 10	multi-lingual, YouTube content		✓
Fluent Speech Commands	spoken keyword commands		✓
Audio SNIPS	spoken commands, crowdsourced		✓
CREMA-D	affective speech		✓
RAVDESS	affective speech	✓	
Libri2Mix	multi-speaker scenarios		✓
Audio dataset	**Features**		
ESC-50	diverse audio source	✓	✓
FSD-50K	diverse audio source		✓
Gunshot Triangulation	diverse audio source		✓

Table 1: *Dataset information.* ***app*** *implies the dataset is used in application-level evaluation.* ***obj*** *implies the dataset is used in objective metrics evaluation.*

organized into four subsets: train-360, train-100, dev, and test, cumulatively encompassing 300 hours of speech. We use the test set for codec evaluation.

RAVDESS [24] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) stands as a renowned emotional dataset, licensed under CC BY-NC-SA 4.0. It is acted by 24 professional actors (12 female, 12 male)in North American accents. Within the dataset, speech encompasses expressions of calm, happiness, sadness, anger, fear, surprise, and disgust.

CREMA-D [25] has 7,442 clips from 91 actors (48 male and 43 female). Each clip is annotated with six distinct emotions. The professional actors, guided by experienced theatre directors, skillfully express a designated emotion while delivering specific sentences.

VoxLingua107 Top 10 [26] comprises audio segments for spoken language identification, encompassing 107 distinct languages. The audio clips in this dataset are automatically extracted from YouTube videos. We use the audio clips from the top 10 most frequent languages.

ESC-50 [27] encompasses 2000 environmental sounds categorized into 50 classes. The clips within this dataset are manually selected from public field recordings compiled by the Freesound.org project.

FSD50K [28] is an open collection of human-labeled sound events. It comprises 51,197 Freesound clips distributed across 200 classes, selected from the AudioSet Ontology. We use a test and valid set for codec evaluation.

Gunshot Triangulation [29] collect the audio of seven distinct firearms—comprising four pistols and three rifles—each fired a minimum of three times. The shots were directed sequentially toward a target positioned 45 meters away from the shooter in an open field. The sound associated with these firings was captured using four separate iPod Touch devices.

2.1.2. Hidden set

The other dataset is newly created by us and maintained as a hidden set. The hidden set will include counterparts for all kinds of datasets in the open set. To construct these hidden datasets, we collaborate with LxT[3], to engage 60 human speakers, ensuring gender balance, to recite sentences and record the audio.

[3]https://www.lxt.ai/

2.2. Objective metrics

The diverse set of signal-level metrics, including Perceptual Evaluation of Speech Quality (PESQ) [30][4], Short-Time Objective Intelligibility (STOI) [32][5], Signal-to-distortion ratio (SDR), Mel Spectrogram Loss (MelLoss) [10] [6], enable us to conduct a thorough evaluation of audio quality across various dimensions, encompassing spectral fidelity, temporal dynamics, perceptual clarity, and intelligibility.

2.3. Application

The application angle evaluation will comprehensively analyze each codec's ability to preserve crucial audio information, encompassing content (word error rate (WER) for automatic speech recognition (ASR)), speaker timbre (equal error rate (EER) for automatic speaker verification (ASV)), emotion (accuracy for speech emotion recognition), and general audio characteristics (mean average precision (mAP) for audio event classification).

2.3.1. Automatic speech recognition (ASR)

For the ASR evaluation, we use the Whisper model [33] to assess how well various codecs preserve context information within speech. We use the word error rate (WER) and edit distance as primary metrics. This evaluation is conducted on the LibriSpeech dataset [20], specifically focusing on the test-clean and test-other subsets. These metrics help determine the effectiveness of codecs in maintaining the clarity and accuracy of spoken content during resynthesis.

2.3.2. Automatic speaker verification (ASV)

Speaker information represents a distinct and unique aspect of speech. We employ ASV to assess the degree of speaker information loss in the resynthesized speech generated by neural codecs. We utilize the cutting-edge speaker verification model, ECAPA-TDNN [34] [7], for the pre-trained ASV model. We adopt equal error rate (EER) as the evaluation metric to evaluate the performance of ASV on Voxceleb test-O set [22]. EER provides a balance between false acceptances and rejections.

2.3.3. Emotion recognition

In addition to speaker information, speech conveys affective information, including emotions. We employ ER to quantify the degree of paralinguistic information loss due to speech resynthesis by codec models. We utilize the emotion2vec [35] [8] to evaluate the one famous emotion dataset, RAVDESS [24].

2.3.4. Audio event classification

The purpose of the AEC task is to evaluate how well different codecs maintain audio event information. This is done by utilizing a pre-trained AEC model to classify audio events of re-synthesized audio. We use the pre-trained Contrastive Language-Audio Pretraining (CLAP) model [36,37][9] test on the ESC-50 dataset [27].

[4]We use the implementation from [31]
[5]https://github.com/mpariente/pystoi
[6]https://github.com/descriptinc/descript-audio-codec/tree/main
[7]https://github.com/TaoRuijie/ECAPA-TDNN
[8]https://github.com/ddlBoJack/emotion2vec
[9]https://github.com/microsoft/CLAP

3. Registration process

Please use the following Google Form to register: `https://forms.gle/sBRB4VsoDKkNYQQ98`

4. Submission of results

Our main focus is sharing observations and insights with the community, rather than just ranking.

4.1. Open set

Participants should submit the evaluation results by creating a GitHub issue `https://github.com/voidful/Codec-SUPERB/tree/SLT_Challenge` for all objective metrics and applications, as well as the adopted bitrate.

4.2. Hidden set

Participants have two choices:

- If the model checkpoint can be released, the participants can submit a script that indicates the available bitrate choices, takes the waveform path as an input argument, and re-synthesizes the waveform.
- If the model checkpoint can not be released, the participants can provide an API. Organizers will use the API to select available bitrates supported by the submitted codec model, input the waveform, and re-synthesize the waveform.

5. Paper submission

A special session dedicated to the Codec-SUPERB challenge will be featured at SLT 2024. Participants in the Codec-SUPERb challenges may choose to submit papers via the regular submission system, which will go through SLT peer review process. Additionally, challenge participants have the option to submit a paper describing their systems to distinct Challenge Proceedings. The challenge organizers will review these submissions. While accepted system description papers will not be indexed by IEEE, authors will be given the opportunity to showcase their work during a specific session at the workshop.

6. References

[1] H. Wu, X. Chen, Y.-C. Lin, K.-w. Chang, H.-L. Chung, A. H. Liu, and H.-y. Lee, "Towards audio language modeling-an overview," *arXiv preprint arXiv:2402.13236*, 2024.

[2] H. Wu, H.-L. Chung, Y.-C. Lin, Y.-K. Wu, X. Chen, Y.-C. Pai, H.-H. Wang, K.-W. Chang, A. H. Liu, and H.-y. Lee, "Codec-superb: An in-depth analysis of sound codec models," *arXiv preprint arXiv:2402.13071*, 2024.

[3] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[4] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[5] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[6] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[8] Z. Du, S. Zhang, K. Hu, and S. Zheng, "Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," *arXiv preprint arXiv:2309.07405*, 2023.

[9] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

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

[11] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.

[12] S. Ji, M. Fang, Z. Jiang, R. Huang, J. Zuo, S. Wang, and Z. Zhao, "Language-codec: Reducing the gaps between discrete codec representation and speech language models," *arXiv preprint arXiv:2402.12208*, 2024.

[13] Y.-C. Wu, D. Marković, S. Krenn, I. D. Gebru, and A. Richard, "Scoredec: A phase-preserving high-fidelity audio codec with a generalized score-based diffusion post-filter," *arXiv preprint arXiv:2401.12160*, 2024.

[14] Y. Zheng, W. Tu, L. Xiao, and X. Xu, "Srcodec: Split-residual vector quantization for neural speech codec," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 451–455.

[15] ——, "Supercodec: A neural speech codec with selective back-projection network," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 566–570.

[16] L. Xu, J. Wang, J. Zhang, and X. Xie, "Lightcodec: A high fidelity neural audio codec with low computation complexity," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 586–590.

[17] H. Yang, I. Jang, and M. Kim, "Generative de-quantization for neural speech codec via latent diffusion," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1251–1255.

[18] X. Anguera, L.-J. Rodriguez-Fuentes, A. Buzo, F. Metze, I. Szöke, and M. Penagarikano, "Quesst2014: Evaluating query-by-example speech search in a zero-resource setting with real-life queries," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5833–5837.

[19] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. of Interspeech*, G. Kubin and Z. Kacic, Eds., 2019.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[21] C. Lai, Y. Chuang, H. Lee, S. Li, and J. R. Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *ICASSP*. IEEE, 2021, pp. 7468–7472.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*. ISCA, 2017, pp. 2616–2620.

[23] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[24] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[25] 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.

[26] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[27] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[28] 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, 2022.

[29] S. Cooper and S. Shaw, "Gunshots recorded in an open field using ipod touch devices," *Dryad, Dataset*, 2020.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] M. Wang, C. Boeddeker, R. G. Dantas, and ananda seelan, "ludlows/python-pesq: supporting for multiprocessing features," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6549559

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *PREPRINT*, 2022.

[34] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[35] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[36] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," 2023. [Online]. Available: https://arxiv.org/abs/2309.05767

[37] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.