Codec-SUPERB: An In-Depth Analysis of Sound Codec Models

Haibin Wu[1] *, **Ho-Lam Chung**[1,3] *, **Yi-Cheng Lin**[1] †, **Yuan-Kuei Wu**[1] †, **Xuanjun Chen**[1] †, **Yu-Chi Pai**[1], **Hsiu-Hsuan Wang**[1], **Kai-Wei Chang**[1], **Alexander H. Liu**[2], **Hung-yi Lee**[1]

[1]National Taiwan University
[2]Massachusetts Institute of Technology
[3]ASUS Intelligent Cloud Services
hungyilee@ntu.edu.tw

Abstract

The sound codec's dual roles in minimizing data transmission latency and serving as tokenizers underscore its critical importance. Recent years have witnessed significant developments in codec models. The ideal sound codec should preserve content, paralinguistics, speakers, and audio information. However, the question of which codec achieves optimal sound information preservation remains unanswered, as in different papers, models are evaluated on their selected experimental settings. This study introduces Codec-SUPERB, an acronym for **Codec S**ound processing **U**niversal **PER**formance **B**enchmark. It is an ecosystem designed to assess codec models across representative sound applications and signal-level metrics rooted in sound domain knowledge. Codec-SUPERB simplifies result sharing through an online leaderboard, promoting collaboration within a community-driven benchmark database, thereby stimulating new development cycles for codecs. Furthermore, we undertake an in-depth analysis to offer insights into codec models from both application and signal perspectives, diverging from previous codec papers mainly concentrating on signal-level comparisons. Finally, we will release codes, the leaderboard, and data to accelerate progress within the community.

1 Introduction

Neural sound codec models were initially introduced to compress sound for efficient data transmission. The encoder of the codec model encodes the sound into codec codes, which are then transmitted. Subsequently, the codec decoder then resynthesizes the sound using the received codes.

Neural codec codes can be utilized as tokens in sound language modeling (LM). LM has proven highly successful in Natural Language Processing (NLP). Sound data contains semantic content and rich information about speaker, emotion, and general audio, offering deeper possibilities for language model applications. Researchers recently explored the potential of neural codecs (Défossez et al., 2022; Zeghidour et al., 2021; Borsos et al., 2023b; Wu et al., 2023; Yang et al., 2023a; Du et al., 2023; Zhang et al., 2023a; Kumar et al., 2023) as suitable tokenizers for converting continuous sound into discrete tokens, which can be employed in sound LM (Wu et al., 2024; Borsos et al., 2023a; Rubenstein et al., 2023; Agostinelli et al., 2023; Wang et al., 2023a; Zhang et al., 2023b; Wang et al., 2023c; Yang et al., 2023b; Chen et al., 2023; Wang et al., 2023d; Copet et al., 2023; Lan et al., 2023; Kreuk et al., 2022). Numerous high-performance neural codecs have been developed.

The dual roles of minimizing data transmission latency and serving as sound LM tokenizers require an ideal codec to preserve content, paralinguistic, speaker, and audio information under low bitrate measured by thousand bits per second (kbps). However, the question of which codec achieves the most optimal information preservation across various aspects remains unanswered, as codec models are evaluated with various experimental settings in different papers. Furthermore, prior codec papers mainly compare performance based on signal-level metrics, neglecting downstream application angles.

To address the aforementioned limitations, we introduce Codec-SUPERB, shorted for **Codec S**ound processing **U**niversal **PER**formance **B**enchmark, which firstly provides a holistic comparison of the current state-of-the-art codecs under the same, fair and comprehensive experimental setting. We highlight the following features of Codec-SUPERB:

1. **Diverse angles**: Codec-SUPERB conducts a comprehensive analysis to provide insights into codec models from both application and

*equal first contribution, †equal second contribution, order is sorted randomly.

Findings of the Association for Computational Linguistics ACL 2024, pages 10330–10348
August 11-16, 2024 ©2024 Association for Computational Linguistics

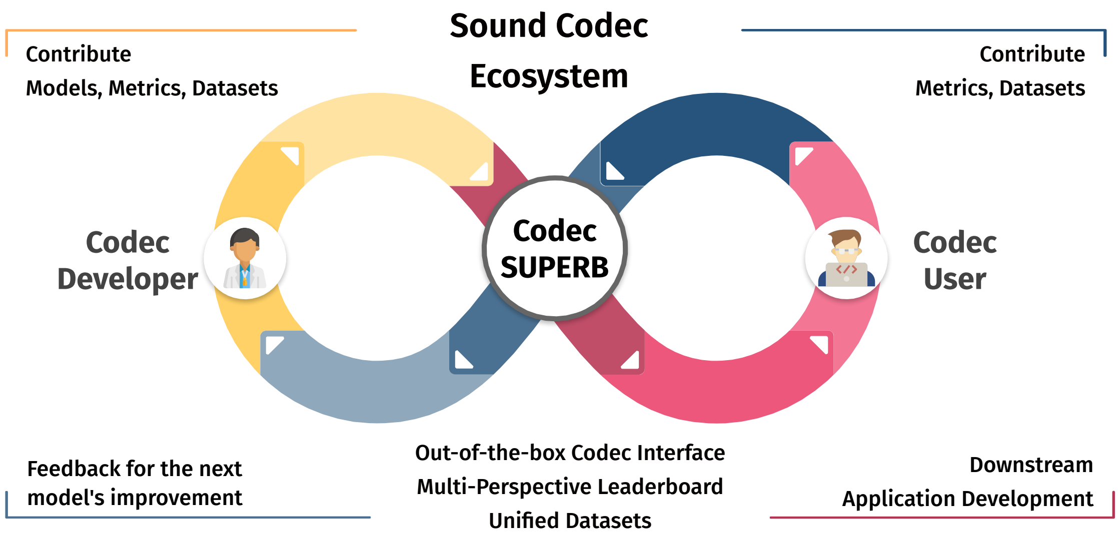


Figure 1: Illustration of the Codec-SUPERB platform from two angles: developers and users. From the perspective of developers, they develop and evaluate new codec models across a spectrum of sound applications and signal-level metrics defined in our codebase. Developers then submit their prediction files to the online leaderboard to expand the benchmark database and facilitate comparisons with other codec models. Ultimately, developers utilize the codebase's visualization and statistical tools to analyze performance discrepancies among Codec-SUPERB applications and metrics, thereby gaining invaluable insights for future improvement directions. From the users' perspective, they can contribute datasets and metrics and pick codec models for their downstream application usage.

signal perspectives, diverging from previous codec papers that predominantly focus on signal-level comparisons.

2. **Extensive coverage**: Codec-SUPERB exhaustively standardizes the comparison of codec models across six distinct codec models, each with its unique training settings, resulting in 19 distinct codec models. We evaluate 19 codec models across four applications to include comparison for content, speaker, paralinguistic, and audio information. Furthermore, we conduct signal-level comparisons across 20 datasets spanning speech, audio, and music data categories.

3. **Community collaboration**: We established an online leaderboard to showcase results, facilitating easy integration of future codec models for public submissions and supporting comparative analysis with statistical and visualization tools (Section 2). We make all resources in Section 2 open-source, welcoming researchers to contribute and promote advancements within the codec community.

2 Codec-SUPERB Platform design

As shown in Figure 1, Codec-SUPERB is designed to foster sound codec development by providing a platform for connecting codec developers and codec users. Codec-SUPERB is user-friendly for reproducing the model evaluation, assessing the custom codec models, contributing datasets and metrics, and conducting comparative analyses for model characteristics. This is facilitated by three core components: an easy-to-follow codebase, a community-driven leaderboard website, and well-selected datasets.

2.1 Codebase

The Codec-SUPERB evaluation processes are conducted through our GitHub repository [1]. Within this repository, codec models are referred to as *base_codec* models, intentionally designed to be disentangled from evaluations for downstream applications and signal-level metrics. This disentanglement enables users to seamlessly switch between various *base_codec* and evaluation combinations or add their own *base_codec* model for evalu-

[1] Codec-SUPERB codebase

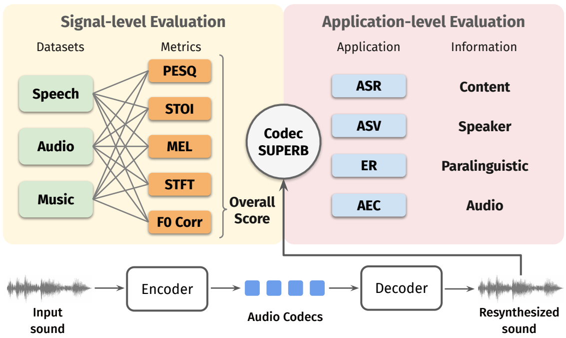


Figure 2: The input sound is compressed using the codec encoder and resynthesized using the codec decoder. Then the resynthesized sound is evaluated from signal-level and application-level angles. Three categories of dataset, speech, audio, and music, are evaluated using five signal-level metrics and one overall score. Also, 4 downstream applications are evaluated.

ation across all sound applications and signal-level metrics. The codebase is closely combined with the Codec-SUPERB official leaderboard website [2], enabling the automatic generation of corresponding submission files upon completion of each codec evaluation. Users can then upload these submission files, thus effortlessly contributing to expanding the benchmark database.

2.2 Website

Our leaderboard website plays a pivotal role in the Codec-SUPERB by continuously expanding the benchmark database, ensuring Codec-SUPERB remains more than just a static leaderboard to showcase our own evaluation results. Initially, we evaluate 19 codec models and submit them to our online leaderboard. To lower the participation threshold, the website also accepts submissions with partial results based on the developers' interests from specific angles when evaluating all results is not necessary and cost-prohibitive. Additionally, the website offers helpful visualization tools for comparing detailed characteristics of different models, as demonstrated in the experiment section.

2.3 Datasets

Neural codec models are challenging to compare, even when the source code is available, due to slight variations in evaluation dataset settings. This issue is obvious in the sound domain, where differences in sound sampling rates, data partition rules, and waveform preprocessing methods can yield significantly different results. To address this challenge and align with sound domain expertise, we curate a comprehensive dataset spanning 20 datasets, comprising a diverse array of speech, music, and audio data. These datasets are partitioned according to sound domain knowledge. This extensive dataset is publicly accessible and readily available through our leaderboard. We strive to cover as many different datasets and commonly used metrics as possible, presenting all results so users can select the metrics they wish to see and require according to their needs.

[2]Codec-SUPERB leaderboard

3 Holistic evaluation in Codec-SUPERB

Codec models are assessed across diverse experimental settings in various papers. Moreover, previous codec studies primarily focus on comparisons using signal-level metrics on their *selected* datasets, overlooking evaluations from downstream application angles. Thus, we make efforts to address the above limitations by including diverse datasets, comprehensive signal-level metrics, and mainstream sound applications. The detailed evaluation process is illustrated in Figure 2.

3.1 Signal-level evaluation

We utilize the codec models to resynthesize the datasets in Section 3.1.1. Additionally, we employ carefully selected objective metrics outlined in Section 3.1.2 and a well-designed overall score detailed in Section 3.1.3 to do signal-level comparisons for different codec models.

3.1.1 Datasets

Previous studies typically focus on non-comprehensive categories of data and often rely on a limited number of datasets for evaluation. We select representative sound datasets spanning three mainstream sound categories: speech, general audio, and music. This is because they offer comprehensive perspectives on sound. We incorporate all categories across a total of 20 datasets. To ensure fair comparisons, we standardize the dataset settings, including sampling rate and partition rules. All datasets and partition rules are released on our website. This diverse dataset comprehensively evaluates each codec's performance across various sound types.

Speech: Sound generated by human articulation. It is typically characterized by producing

Speech dataset	Features
Librispeech	diverse speaker, read audiobooks
VoxCeleb1	diverse speaker, celebrities on YouTube
Fluent Speech Commands	spoken keyword commands
QUESST	multi-lingual, low resource language
VoxLingua107 Top 10	multi-lingual, YouTube content
Audio SNIPS	spoken commands, crowdsourced
IEMOCAP	affective speech
CREMA-D	affective speech
Libri2Mix	multi-speaker scenarios
LibriCount	multi-speaker scenarios
Audio dataset	**Features**
ESC-50	diverse audio source
FSD-50K	diverse audio source
Gunshot Triangulation	diverse audio source
Vocal Imitations	human imitation of sound
Music dataset	**Features**
OpenSinger	singing voice, Chinese song
M4Singer	singing voice, Chinese song
VocalSet	singing skill
NSynth	instrument notes
GTZAN Genre	diverse music genre
GTZAN Music Speech	instrument note

Table 1: Dataset information.

specific sounds and patterns to convey meaningful messages. We select speech datasets based on two perspectives: enhancing the diversity of the dataset (i.e., speaker diversity, language variety, and duration) and expanding the range of information preserved in the speech (i.e., emotion and multi-speaker scenarios).

Music: Pattern of sounds created through pitch, tone, and timbre manipulation. We select music datasets to enhance the diversity of music categories, encompassing various singing voices of different levels of professionalism, music notes played by multiple instruments, and music spanning a variety of genres.

Audio: Any sound that humans can hear apart from speech and music. We chose audio datasets in order to increase the diversity of general audio categories and their applications.

We briefly describe the key information for the selected datasets in Table 1. Detailed descriptions, including the partition rules of datasets used in our evaluation, can be found in Appendix A.1. Also, Table 5 in Appendix A.1 summarizes the license for each dataset.

3.1.2 Signal-level metrics

We assess the quality of resynthesized sound using a comprehensive set of signal-level metrics grounded in sound domain expertise. These metrics include the Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), Short-Time Objective Intelligibility (STOI) (Taal et al., 2010), STFT distance (STFTDistance), Mel distance (MelDistance), and F0CORR (F0 Pearson Correlation Coefficient) (Jadoul et al., 2018). The features of the adopted signal-level metrics to assess sound quality are shown in Table 2. STFTDistance analyzes frequency content and temporal dynamics, while MelDistance focuses on spectral fidelity and timbral texture, reflecting the Mel scale's relevance to human hearing. PESQ provides a subjective quality score, capturing the perceptual quality of speech. STOI measures speech intelligibility in noise, essential for clear communication. F0CORR evaluates pitch accuracy, crucial for naturalness and expressiveness in sound. This diverse set of metrics enables us to conduct a thorough evaluation of sound quality across various dimensions, encompassing spectral fidelity, temporal dynamics, perceptual clarity, and intelligibility. Details for these metrics are shown in Appendix A.3

Metric	Functionality	Range
STFTDistance	Frequency content discrepancies.	$[0, \infty)$
MelDistance	Gauges the fidelity of spectral features.	$[0, \infty)$
PESQ	Rates the perceptual quality of speech.	$[-0.5, 4.5]$
STOI	Evaluates speech intelligibility.	$[0, 1]$
F0CORR	Measures pitch accuracy.	$[0, 1]$

Table 2: Summary of signal-level metrics

3.1.3 Overall score for Signal-level metrics

Currently, no single overall score exists to evaluate signal-level metrics of resynthesized sound produced by codec models. What's particularly innovative is our introduction of a unified overall score, which integrates all signal-level metrics in Section 3.1.2 for improved visualization. Notably, the overall score demonstrates strong correlations with each individual metric as shown in Section 4.2.

The overall score is calculated through normalization and harmonic mean combining all metrics. Normalization ensures metrics are comparable and less affected by outliers. For bounded metrics, PESQ, STOI, and F0CORR, we normalize them by subtracting the min and dividing by the range. For unbounded metrics, STFTDistance and MelDistance, we normalize them by the Sigmoid function. Inspired by F1 score (Chicco and Jurman, 2020), the harmonic mean is used to aggregate the normalized scores, prioritizing balanced performance across metrics. Similar to the F1 score (Chicco and Jurman, 2020), which harmonizes precision and recall, the harmonic mean in our context ensures a balanced evaluation, preventing any single signal-level metric from disproportionately influencing the overall score.

3.2 Application-level evaluation

Beyond previous works mainly focusing on signal-level comparison, we expand our evaluation to include application-level metrics. This step is essential for comprehensively understanding each codec's ability to preserve crucial sound information, encompassing content, speaker timbre, emotion, and general audio characteristics. For downstream application evaluation, we utilize pre-trained models to analyze the quality of resynthesized sound. Details are shown below.

3.2.1 Automatic speech recognition (ASR)

We use ASR to evaluate the content information loss of the codec resynthesis process. Our study evaluates the "whisper-large" variant of the Whisper ASR model (Radford et al., 2023), renowned for its robust performance across multiple languages and tasks, utilizing an encoder-decoder Transformer architecture. We use the most common metric, Word Error Rate (WER), and the most common dataset, LibriSpeech. This evaluation aims to showcase Whisper's proficiency in handling diverse speech qualities and accents, underscoring its potential in real-world speech recognition applications. More details can be found in Appendix A.2.1.

3.2.2 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. As the pre-trained ASV model, we utilize the cutting-edge speaker verification model, ECAPA-TDNN (Desplanques et al., 2020). We adopt equal error rate (EER) and minimum decision cost function (minDCF) as two evaluation metrics to evaluate the performance of ASV. EER provides a balance between false acceptances and rejections, and minDCF allows for a more nuanced assessment of system performance by considering the costs associated with different types of errors (false acceptances and rejections). More details can be found in Appendix A.2.2

3.2.3 Emotion recognition (ER)

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 WavLM-Large (Chen et al., 2022) self-supervised model for feature extraction and train an emotion classification model on the most famous emotion dataset, IEMOCAP. This setting achieves robust and nearly SOTA results. More details on ER downstream task setting can be found in Appendix A.2.3

3.2.4 Audio event classification (AEC)

The goal of adopting AEC is to assess the fidelity of various codecs in preserving audio event information by leveraging a pre-trained AEC model to classify sound events for audio re-synthesized by these codecs. We leverage the pre-trained Audio Spectrogram Transformer (AST) (Gong et al., 2021) model and test on the original AudioSet (Gemmeke et al., 2017) evaluation set as the baseline. More AEC downstream task setting details can be found in Appendix A.2.4.

4 Experiments

4.1 Experimental setup

We adopt six open-source codec models, SpeechTokenizer (Zhang et al., 2023a), AudioDec (Wu et al., 2023), AcademiCodec (Yang et al., 2023a), Descript-audio-codec (DAC) (Kumar et al., 2023) Encodec (Défossez et al., 2022), and FunCodec (Du et al., 2023), each with its own distinct training specifications, yielding a total of 19 unique codec models for comparison. The column (a) of Table 3 provides brief information regarding these models. Detailed information is in Appendix A.4

We select different objective metrics based on the nature of different types of sound to evaluate different categories of sound. Speech data adopts STFTDistance, MelDistance, PESQ, and STOI. Audio data adopts STFTDistance and MelDistance. Music data includes all metrics, particularly F0CORR, for fidelity and expressiveness.

4.2 Signal-level evaluation

To conduct signal-level evaluation, we employ the "Overall score" as the principal metric. To affirm the overall score as a reliable measure of codec performance to consider diverse signal-level metrics, we conduct a correlation analysis, summarized in Table 4. This analysis aims to find the correlation scores between the overall score rankings and those from individual signal-level metric scores.

(a) Codec Information			(b) Signal-level Evaluation			(c) Application-level Evaluation				
	kbps	**Other Configuration**	**Speech↑**	**Audio↑**	**Music↑**	**WER↓ (ASR)**	**EER↓ (ASV)**	**minDCF↓ (ASV)**	**ACC ↑ (ER)**	**mAP ↑ (AEC)**
None	-	-	-	-	-	2.96	0.86	0.07	69.84	45.68
A	4	16k	0.644	0.581	0.585	4.02	3.31	0.24	65.49	15.11
B1	2	16k_320d	0.610	0.574	0.601	4.94	4.43	0.29	65.96	16.19
B2	2	16k_320d_large_uni	0.617	0.574	0.630	6.26	5.22	0.38	64.63	28.65
B3	3	24k_320d	0.611	0.592	0.604	4.49	6.16	0.36	65.95	14.01
C	6.4	24k_320d	0.596	0.602	0.572	3.94	5.22	0.30	65.70	17.41
D1	6	16k	0.798	0.591	0.749	3.26	1.59	0.12	68.81	41.08
D2	24	24k	**0.864**	0.636	**0.815**	**2.96**	2.24	0.14	**69.56**	**41.37**
D3	8	44k	0.802	**0.702**	0.770	3.18	3.59	0.26	69.18	32.04
E1	1.5	24k	0.579	0.594	0.568	9.21	13.88	0.68	58.84	18.84
E2	3	24k	0.636	0.599	0.621	4.34	6.85	0.39	63.54	26.63
E3	6	24k	0.697	0.602	0.669	3.49	4.28	0.27	66.18	32.43
E4	12	24k	0.748	0.606	0.710	3.22	3.44	0.21	67.63	35.84
E5	24	24k	0.775	0.609	0.732	3.17	3.15	0.19	68.26	36.64
F1	16	en_libritts_16k_gr1nq32ds320	0.724	0.582	0.667	3.21	**1.50**	**0.10**	63.54	37.31
F2	16	en_libritts_16k_gr8nq32ds320	0.704	0.583	0.668	3.16	1.81	**0.10**	66.18	37.77
F3	16	en_libritts_16k_nq32ds320	0.705	0.581	0.649	3.28	1.76	0.12	67.63	25.52
F4	8	en_libritts_16k_nq32ds640	0.678	0.578	0.632	3.43	2.04	0.13	68.26	21.43
F5	16	zh_en_16k_nq32ds320	0.726	0.583	0.665	3.21	1.52	0.11	69.25	26.42
F6	8	zh_en_16k_nq32ds640	0.718	0.583	0.667	3.27	1.60	0.11	69.55	33.59

Table 3: Comparison between codec models. (a) Codec information. "A" denotes the Speech Tokenizer, "B∼" signifies the AcademiCodec, "C" is associated with AudioDec, "D∼" represents the DAC, "E∼" refers to the EnCodec, and "F∼" indicates the FunCodec. (b) Signal-level evaluation. (c) Application-level evaluation. "None" means that no codec has been applied.

Metric	**Dataset**	**Mean Correlation**	**Mean p-value**
Mel	Audio	-0.91	2.7×10^{-7}
STFT	Audio	-0.94	1.7×10^{-9}
Mel	Music	-0.65	2.4×10^{-2}
STFT	Music	-0.58	4.6×10^{-2}
PESQ	Music	0.95	3.1×10^{-9}
STOI	Music	0.83	2.8×10^{-3}
F0CORR	Music	0.74	2.7×10^{-2}
Mel	Speech	-0.77	6.1×10^{-4}
STFT	Speech	-0.71	1.0×10^{-2}
PESQ	Speech	0.97	1.8×10^{-11}
STOI	Speech	0.84	2.9×10^{-5}

Table 4: Consolidated average correlation coefficients and p-values across three kinds of datasets. A correlation value above 0.7 (below -0.7) indicates a strong positive (negative) correlation. A p-value less than 0.05 denotes significance (cor, 2015).

We show the mean correlation values for each kind of dataset in Table 4. The results of all metrics for all datasets are presented in Figure 5 - Figure 7 in Appendix B. Key findings include:

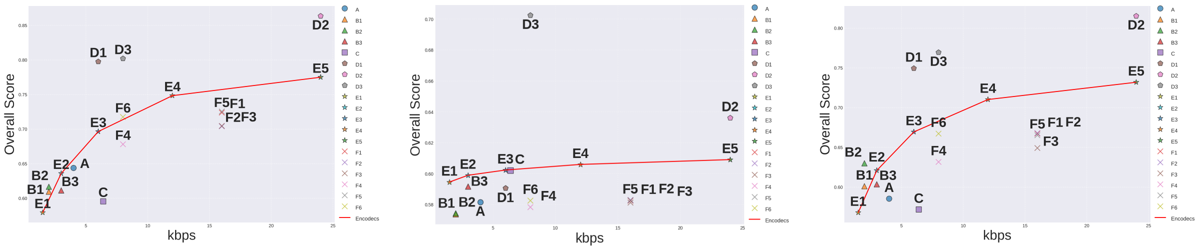
- MelDistance and STFTDistance have strong negative relations with the overall score.
- PESQ, STOI, and F0CORR have strong positive relations with the overall score.
- Mean p-values confirm the above correlations are significant.

This correlation analysis, detailed in Table 4, establishes the overall score as a comprehensive indicator of codec quality, effectively encompassing various signal-level metrics. We can discern which codec achieves superior performance at a given bitrate by comparing the “Overall Score” against the bitrate (kbps).

Table 6 to Table 8 in Appendix B show the results for each dataset. We only show the average performance below due to space limitations. The performance trends are similar

4.2.1 Speech Dataset

As shown in Figure 3a, for the Speech dataset, it’s clear that the Encodec (E1-E5) sets a strong baseline, with only the DAC codec (D1-D3) notably surpassing it at a similar bitrate. Other codecs don’t show a significant advantage. In addition, at very low bitrates, the Academicodec (B1-B3) achieves improved performance.



(a) Speech Overall Score vs bitrate. (b) Audio Overall Score vs bitrate. (c) Music Overall Score vs bitrate.

Figure 3: Points in the upper left corner represent a better trade-off between performance and bitrate.

4.2.2 Audio Dataset

As shown in Figure 3b, for the Audio dataset, Encodec again proves to be a strong baseline, with DAC being the only codec to surpass its performance significantly. Other models do not markedly exceed the performance of Encodec.

4.2.3 Music Dataset

As shown in Figure 3c, the results within the music dataset seem to consolidate the findings from the previous two datasets. Academicodec outperforms Encodec at low bitrates and can even surpass Encodec models when the bitrate is doubled. DAC also maintains a leading position.

4.2.4 Takeaways

The observations across the three datasets indicate that DAC achieves a well-balanced trade-off between performance and bitrate. In contrast, Academicodec demonstrates the capability to maintain superior performance even at a significantly lower bitrate. The early-stage model, Encodec, remains a solid baseline.

4.3 Application-level evaluation

4.3.1 Automatic speech recognition

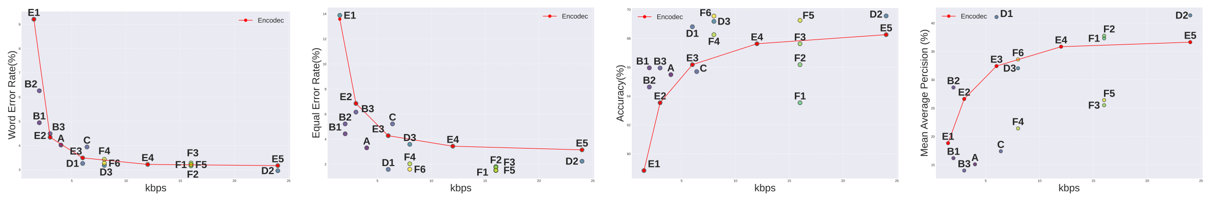
The column (c) of Table 3 (Table 3-c) indicates that the process of codec resynthesis process typically leads to a loss of contextual information in speech, adversely affecting the Word Error Rate (WER). However, an intriguing exception to this pattern is observed with the D2 codec (dac_24k), which maintains ASR performance comparable to the original, unprocessed speech. This suggests that the D2 codec's resynthesis process uniquely preserves the integrity of the speech content. As depicted in Figure 4a, we observe that: (1) WER consistently decreases as bitrate increases, signifying that a higher bitrate contributes to preserving content information within the codec; (3) the DAC codecs(D1-D3) exhibit the lowest WER across the board, indicating their effectiveness in maintaining content information during resynthesis; (4) in contrast, with a bitrate of around 6kbps, the AudioDec codec (C) obtained the highest WER, which indicates a significant loss of content information.

4.3.2 Automatic speaker verification

As presented in Table 3-c, adopting codecs A to F6 to the original audio leads to some loss of speaker information. Based on Figure 4b and Table 3-c, we can observe that: (1) when comparing Encodec E1 to E5, we observe an increase in bitrate and a decrease in EER, indicating that a higher bitrate can better preserve speaker information; (2) Funcodec F2 has the lowest EER and minDCF, resulting in the least degradation of speaker information. Funcodec F2 also attains the highest bitrate, which results in preserving more information than other codec models; (3) B1, A, and D1 attained the optimal Pareto balance, effectively striking a fine trade-off between EER and bitrate; (4) DAC D1 attained an impressively low EER while maintaining a reasonable bitrate.

4.3.3 Emotion recognition

Our observations are drawn from Figure 4c and Table 3-c: (1) when comparing Encodec from E1 to E5, we observe an increase in bitrate and Accuracy (%), indicating that a higher bitrate can preserve more information; (2) DAC model D2 has the highest ACC 69.56%. D2 only drops in accuracy by 0.38% compared to the original audio; (3) Encodec E1 has the worst accuracy. When comparing E1 with B1 and B2, both having similar bitrate, a significant decrease in accuracy is observed; (4) under the same bitrate, DAC outperforms all the other codecs; F2 outperforms F1, F3, and F5; AcademiCodec B1 outperforms B2 even though they have the same architecture while B2 is trained on a



(a) ASR WER (%) vs bitrate. (b) ASV EER (%) vs bitrate. (c) ER ACC (%) vs bitrate. (d) AEC mAP (%) vs bitrate.

Figure 4: The trade-off between bitrate and performance of different applications. Models located in the lower/upper left corner of (a)(b)/(c)(d) indicate a more favorable trade-off between performance and bitrates.

larger set of data; AcademiCodec B3 outperforms Encodec E3; (5) in general, Encodec models and AudioDec have the lowest accuracy under a similar bitrate, suggesting more loss of emotional information; (6) even for the least performing model, E1, with a bitrate of only 1.5kbps, it can still maintain the accuracy drop from 69.84% to 58.84%. This implies that emotional information can be retained even with a very low bitrate.

4.3.4 Audio event classification

Based on Figure 4d and Table 3-c, we can observe that: (1) when comparing Encodec from E1 to E5, we observe an increase in bitrate and mAP (%), indicating that a higher bitrate can better preserve audio event information; (2) DAC model D2 has the highest mAP 41.37%, resulting in the least degradation of audio event information. (3) EnCodec and DAC models are trained using AudioSet training set, so when comparing mAP (%) for AudioSet testing set under different bitrates, EnCodec often exceeds SpeechTokenizer (around 4kbps), AcademiCodec (around 3kbps), and AudioDec (around 6kbps); (4) AcademiCodec B2 significantly surpassed EnCodec at a similar rate as it is trained on a diverse speech dataset. (5) FunCodec was not trained using the AudioSet training set. When comparing model F6 with F4, it is evident that F6's performance closely approaches that of EnCodec (which is trained using Audioset). The primary distinction between them is that F6 utilizes a multi-domain speech dataset, whereas F4 relies solely on LibriTTS (Zen et al., 2019).

4.3.5 Takeaways

Here we summarize our findings with lessons:

- Emotion information can be conserved even at a remarkably low bitrate of 1.5kbps.
- Despite not being trained on Audioset, some models (F6) trained on diverse enough speech data can generalize and effectively maintain audio information. This suggests that future efforts to develop a universal codec model may not necessarily require audio data, but rather diverse speech data.
- There exists a clear trade-off between bitrate and the quality of codec resynthesis in terms of all downstream tasks we covered.
- Among the higher bitrate (6kbps~24kbps) models, DAC outperforms other codecs in retaining content, emotion, speaker, and audio information under similar bitrate.
- The best low-bitrate model is Academicodec, which performs excellently in preserving content, emotion, speaker, and audio information from 2 - 3kpbs.

4.4 Discussions

Another way to evaluate the codec models for different applications is to extract the codec codes and train application models upon the extracted codes. Training application models using codec codes requires significant computational resources. We aim to create a lightweight benchmark for users to facilitate easier evaluation of metrics for those proposed codecs, allowing them to obtain preliminary results for later development reference. The users can quickly get insights based on the fast evaluation pipeline.

5 Limitations

Our evaluation data spans 20 datasets, each containing numerous testing samples, resulting in a substantial amount of data that requires significant computation time and resources. As a solution, we aim to devise criteria for identifying representative data points within each dataset to expedite the evaluation process.

6 Conclusion

This study presents Codec-SUPERB, a public framework tailored to fairly and comprehensively evaluate codec models. Comprised of a user-friendly codebase, a collaborative benchmark driven by the community, and meticulously crafted metrics paired with curated datasets, Codec-SUPERB streamlines result comparisons through an interactive online leaderboard. Additionally, our comprehensive analysis provides valuable insights into codec models, examining both application and signal perspectives. This departure from prior codec research, primarily focusing on signal-level comparisons, allows for a richer understanding of codec performance. Our innovative introduction of a unified overall score sets us apart from previous works, seamlessly integrating all signal-level metrics to enhance visualization. Remarkably, this overall score exhibits robust correlations with each individual metric, underscoring its reliability. Lastly, we commit to releasing codes, the leaderboard, and data resources to expedite progress and foster growth within the community.

7 Acknowledgment

We thank the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

2015. Learn about pearson's correlation coefficient in spss with data from the global health observatory data (2012).

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325.

Xavier Anguera, Luis-J Rodriguez-Fuentes, Andi Buzo, Florian Metze, Igor Szöke, and Mikel Penagarikano. 2015. 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), pages 5833–5837. IEEE.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023a. Audiolm: a language modeling approach to audio generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. arXiv preprint arXiv:2305.09636.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42:335–359.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4):377–390.

Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. arXiv preprint arXiv:2310.04673.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics, 21(1):1–13.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622.

Seth Cooper and Steven Shaw. 2020. Gunshots recorded in an open field using ipod touch devices. Dryad, Dataset.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. arXiv preprint arXiv:2306.05284.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. arXiv preprint arXiv:2005.11262.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2023. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. arXiv preprint arXiv:2309.07405.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with WaveNet autoencoders. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1068–1077. PMLR.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:829–852.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In Proc. IEEE ICASSP 2017, New Orleans, LA.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In Proc. Interspeech 2021, pages 571–575.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In Proceedings of the 29th ACM International Conference on Multimedia, pages 3945–3954.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71:1–15.

Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. 2018. Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology. In DCASE, pages 148–152.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33:17022–17033.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. arXiv preprint arXiv:2209.15352.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems, 32.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. arXiv preprint arXiv:2306.06546.

Cheng-I Lai, Yung-Sung Chuang, Hung-Yi Lee, Shang-Wen Li, and James R. Glass. 2021. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. In ICASSP, pages 7468–7472. IEEE.

Gael Le Lan, Varun Nagaraja, Ernie Chang, David Kant, Zhaoheng Ni, Yangyang Shi, Forrest Iandola, and Vikas Chandra. 2023. Stack-and-delay: a new codebook pattern for music generation. arXiv preprint arXiv:2309.08804.

Margaret Lech, Melissa Stolar, Christopher Best, and Robert Bolia. 2020. Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. Frontiers in Computer Science, 2:14.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In Proc. of Interspeech.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In INTERSPEECH, pages 2616–2620. ISCA.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In ICASSP, pages 5206–5210. IEEE.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd Annual ACM Conference on Multimedia, pages 1015–1018. ACM Press.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pages 28492–28518. PMLR.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. 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), volume 2, pages 749–752. IEEE.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5329–5333. IEEE.

Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Edler, and Emanuël AP Habets. 2018. Classification vs. regression in supervised learning for single channel speaker count estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 436–440. IEEE.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing, pages 4214–4217. IEEE.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. Seanet: A multimodal speech enhancement network. arXiv preprint arXiv:2009.02095.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5):293–302.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In Proc. IEEE SLT Workshop.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111.

Jun-You Wang, Hung-Yi Lee, Jyh-Shing Roger Jang, and Li Su. 2023b. Zero-shot singing voice synthesis from musical score. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 1–8. IEEE.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023c. Viola: Unified codec language models for speech recognition, synthesis, and translation. arXiv preprint arXiv:2305.16107.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. 2023d. Speechx: Neural codec language model as a versatile speech transformer. arXiv preprint arXiv:2308.06873.

Pete Warden. 2017. Speech commands: A public dataset for single-word speech recognition. Dataset available online.

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. 2018. Vocalset: A singing voice dataset. In International Society for Music Information Retrieval Conference.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024. Towards audio language modeling-an overview. arXiv preprint arXiv:2402.13236.

Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An open-source streaming high-fidelity neural audio codec. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023a. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. arXiv preprint arXiv:2305.02765.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023b. Uniaudio: An audio foundation model toward universal audio generation. arXiv preprint arXiv:2310.00704.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:495–507.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. arXiv preprint arXiv:1904.02882.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, and Zhou Zhao. 2022. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechtokenizer: Unified speech tokenizer for speech large language models. arXiv preprint arXiv:2308.16692.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023b. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. arXiv preprint arXiv:2303.03926.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. 2020. Neural networks fail to learn periodic functions and how to fix it. Advances in Neural Information Processing Systems, 33:1583–1594.

A Technical Appendix

A.1 Dataset description

21 public datasets are adopted in this work for zero-shot evaluation, including 9 datasets for speech, 4 datasets for audio, and 8 datasets for music. If not specified, we use the whole dataset for evaluation. All the dataset licenses are shown in Table 5.

A.1.1 Speech

Speech Commands v1 Google Speech Command v1 (Warden, 2017) is a dataset designed for recognizing spoken commands, consisting of 64,727 utterances from 1,881 speakers, with each utterance normalized as a 1-second waveform.

QUESST The QUESST 2014 dataset (Anguera et al., 2015) 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 The Fluent Speech Commands dataset (Lugosch et al., 2019) 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 LibriSpeech (Panayotov et al., 2015) 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 The Audio SNIPS dataset (Lai et al., 2021) 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 VoxCeleb (Nagrani et al., 2017) 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 codec evaluation.

IEMOCAP The IEMOCAP dataset (Busso et al., 2008), aimed at Multimodal Emotion Recognition, comprises 151 dialogue recordings, amounting to 302 videos due to the presence of two speakers in each session. It features annotations for 9 distinct emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral) and valence, arousal, and dominance.

Libri2Mix Libri2Mix (Cosentino et al., 2020) 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 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.

CREMA-D (Cao et al., 2014) is a 7,442 original 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.

LibriCount (Stöter et al., 2018) is a generated dataset where each audio clip simulates a cocktail party scenario, incorporating 0 to 10 speech segments from LibriSpeech Test-Clean mixed with a signal-to-noise ratio (SNR) of 0dB.

VoxLingua107 Top 10 (Valk and Alumäe, 2021) 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.

A.1.2 Audio

ESC-50 (Piczak, 2015) 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 (Fonseca et al., 2022) 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 (Cooper and Shaw, 2020) 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.

Vocal Imitations (Kim et al., 2018) comprises 11,242 crowd-sourced vocal imitations covering 302 sound event classes. The original sound recordings for these classes were sourced from Freesound, while their corresponding imitations were gathered through crowd-sourcing. We use the human imitation samples in the "included" split.

A.1.3 Music

OpenSinger (Huang et al., 2021), a Chinese multi-singer vocal dataset, features high-fidelity recordings by professional singers, free of noise and background interference. OpenSinger does not have a standard way of splitting datasets. We use the songs with male prefixes from 25 to 27 and female with prefixes from 45 to 47 in the dataset as our test set, which is based on the split method of a recent paper on zero-short singing voice synthesis (Wang et al., 2023b). We use the test set for codec evaluation.

M4Singer (Zhang et al., 2022) offers a rich collection of about 700 Chinese pop songs recorded by 20 professional vocalists, encompassing all four SATB voice types: soprano, alto, tenor, and bass. Following UniAudio (Yang et al., 2023b), we conduct experiments on the M4Singer test set. We use the test set for codec evaluation.

VocalSet (Wilkins et al., 2018) comprises 10.1 hours of recordings from 20 professional singers (11 male, 9 female), executing 17 distinct vocal techniques, which aids in the development of advanced machine learning models for applications like singer identification, vocal technique detection, and singing synthesis. We only use the test set of VocalSet for experiments.

NSynth (Engel et al., 2017) stands out as a large-scale, high-quality collection of musical notes, significantly surpassing similar public datasets in size. We only use the test set of the NSynth dataset for experiments.

GTZAN Genre (Tzanetakis and Cook, 2002) includes music samples categorized into 10 genres, each containing 100 audio files. All audio files within the dataset have a standardized length of 30 seconds.

GTZAN Music Speech (Tzanetakis and Cook, 2002) consists of both music and speech segments, with each category containing 60 samples having a standardized length of 30 seconds.

A.2 Downstream task description

A.2.1 Automatic Speech Recognition (ASR)

ASR is an essential component in speech processing, aiming to convert speech into text. ASR is dedicated to extracting and interpreting the content information embedded in speech. This involves understanding various linguistic elements within the spoken language, such as phonetics, syntax, and semantics. ASR systems are instrumental in numerous applications, including voice-activated assistants, transcription services, and interactive voice response systems, where accurate content interpretation is crucial.

Our study employs the Whisper model (Radford et al., 2023), specifically the “whisper-large" variant, the current state-of-the-art ASR system, for evaluation. Whisper stands out due to its robust and versatile architecture, which is capable of handling a broad range of speech recognition tasks across multiple languages. The model's foundation is an encoder-decoder Transformer, adept at learning from large, weakly supervised datasets. This enables Whisper to perform effectively in diverse scenarios without requiring specific dataset fine-tuning. The model's comprehensive training approach, focusing on generalization and robustness, positions it as a powerful tool for speech content interpretation.

To evaluate the performance of the Whisper model in ASR, we utilize the metric Word Error Rate (WER). WER measures the percentage of errors at the word level, offering insights into the model's accuracy in transcribing speech to text. The metric is critical in assessing the effectiveness of ASR systems, allowing for a detailed understanding of their capabilities in accurately capturing and converting spoken language into written form. In this evaluation, we analyze the ASR performance using subsets of the LibriSpeech dataset, precisely the test-clean and test-other subsets, to

Speech Dataset	License
LibriSpeech	CC BY 4.0
VoxCeleb1	CC BY 4.0
Fluent Speech Commands	CC BY-NC-ND 4.0
QUESST	Free For Research Purposes
VoxLingua107 Top 10	CC BY 4.0
Audio SNIPS	CC0 v1.0
IEMOCAP	SAIL Agreement
CREMA-D	Open Database License
Libri2Mix	MIT License
LibriCount	CC BY 4.0
Audio Dataset	**License**
ESC-50	CC BY-NC 3.0
FSD-50K	CC0, CC BY 4.0, CC BY-NC 4.0, CC Sampling+ 1.0
Gunshot Triangulation	CC0
Vocal Imitations	CC BY 4.0
Music Dataset	**License**
OpenSinger	CC BY-NC-SA 2.0
M4Singer	CC BY-NC-SA 4.0
VocalSet	CC BY 4.0
NSynth	CC BY 4.0
GTZAN Genre	CC BY 4.0, Apache License v.2.0
GTZAN Music Speech	CC BY 4.0, Apache License v.2.0

Table 5: The dataset license statistics.

ensure a comprehensive assessment of the model's transcription accuracy.

A.2.2 Automatic Speaker Verification (ASV)

In contrast to text, which primarily conveys content information, 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. ASV is a cutting-edge technology that plays a pivotal role in voice authentication and security systems. ASV is designed to verify a person's claimed identity by analyzing their unique vocal characteristics, such as pitch, tone, and speech patterns. It offers a seamless and secure method of confirming whether an individual is who they claim to be, making it a valuable tool in applications ranging from access control and secure transactions to law enforcement and customer service.

We utilize the cutting-edge speaker verification model, ECAPA-TDNN (Desplanques et al., 2020), which is pre-trained on the VoxCeleb2 dataset (Chung et al., 2018), as the pre-trained ASV model. Building upon the well-established x-vector architecture (Snyder et al., 2018), ECAPA-TDNN introduces several novel enhancements inspired by recent trends in face verification.

We adopt equal error rate (EER) as two evaluation metrics to evaluate the performance of ASV. EER provides a balance between false acceptances and rejections, and minDCF allows for a more nuanced assessment of system performance by considering the costs associated with different types of errors (false acceptances and rejections).

A.2.3 Emotion recognition (ER)

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. ER is an essential component in human-computer interaction, such as smart entertainment, healthcare, or e-learning. ER specifically identifies the emotional components of speech that are unrelated to semantic information (Lech et al., 2020).

We adopt the state-of-the-art self-supervised model WavLM-Large (Chen et al., 2022) as feature extractor, and train the downstream emotion classification model using the weighted sum of hidden states as the representation. Following the SUPERB benchmark, we employ mean-pooling followed by a linear layer for modeling purposes and utilize cross-entropy as the training loss function. As for the dataset, we select a subset of the IEMOCAP dataset (Busso et al., 2008), where we have excluded the unbalanced emotion classes, resulting in four remaining classes (neutral, happy, sad, angry). We further divided this subset into five folds for cross-validation purposes. We report the average classification accuracy across the five folds.

A.2.4 Audio event classification

ASE aims to automatically identify and categorize specific sound events or occurrences within an audio recording. These sound events can be various sounds, such as footsteps, car horns, dog barks, music genres, or any other acoustic events. We use AudioSet (Gemmeke et al., 2017) as the evaluation set. AudioSet offers a comprehensive library of sound events, categorized in a hierarchical structure that spans a broad spectrum of sounds, from human and animal noises to natural and environmental sounds and musical and miscellaneous audio events. The Audio Spectrogram Transformer (AST), proposed by (Gong et al., 2021), is a high-performance open-source AEC model. The model takes spectrograms as input features, divides them into patch embeddings, and adds a learnable position embedding for each patch. An extra classification token is added to the input sequence at the beginning. Subsequently, the feature sequence is fed into a Transformer Encoder to make predictions.

We employ the pre-trained AST model[3] for our AEC downstream task evaluation. We adopt mean average precision (mAP)[4] to evaluate the AEC performance. The AST model had undergone pre-training on the AudioSet training dataset and had been subjected to post-processing using a weight averaging strategy to obtain 45.9 mAP(%) at its evaluation set.

A.3 Signal-level metrics

Aligned with expertise in the speech domain, we assess codec models using a comprehensive set of Signal-level metrics, encompassing various aspects of audio quality. These include:

- **STFTDistance**: Evaluates frequency content by calculating the L1-loss over multi-scale STFT (Short-Time Fourier Transform) representations, measuring frequency discrepancies across multiple resolutions. This method provides a detailed assessment of frequency content and temporal dynamics.

- **MelDistance**: Employs the L1-loss between log Mel spectrogram representations to gauge the fidelity of spectral features, reflecting spectral fidelity and timbral texture in the audio.

- **PESQ**: Perceptual Evaluation of Speech Quality, Rates the perceptual quality of speech, mimicking human auditory perception to provide a subjective quality score.

- **STOI**: Short-Time Objective Intelligibility, Evaluates speech intelligibility, especially in noisy conditions, ensuring clarity and comprehensibility of the generated speech

- **F0CORR (F0 Pearson Correlation Coefficient)**: Evaluates the pitch accuracy between original and synthesized audio by aligning their fundamental frequency (F0) contours using dynamic time warping (DTW) and then computing the Pearson correlation. This metric highlights the codec's ability to maintain pitch, which is essential for audio naturalness and expressiveness.

These metrics provide a comprehensive evaluation of codec models, focusing on both quantitative accuracy and perceptual quality of audio.

A.4 Codec models

SoundStream: SoundStream (Zeghidour et al., 2021) stands as one of the pioneering implementations of neural codec models, embodying a classic neural codec architecture comprising encoder, quantizer, and decoder modules. It utilizes the streaming SEANets (Tagliasacchi et al., 2020) as its encoder and decoder. The quantizer incorporates a speech enhancement system with a Residual Vector Quantization (RVQ) (Kumar et al., 2019; Zeghidour et al., 2021) bottleneck to obtain parallel token streams. During training, the model parameters are optimized using a combination of reconstruction and adversarial loss. SoundStorm

[3]https://github.com/YuanGongND/ast
[4]https://scikit-learn.org/

(Borsos et al., 2023b) is an improved version of SoundStream to achieve both efficiency and high-quality audio generation. It accomplishes this by employing an architecture specifically tailored to the hierarchical structure of audio tokens. Moreover, it pioneers a parallel, non-autoregressive decoding scheme, which relies on confidence-based strategies for residual vector-quantized token sequences.

Encodec: Encodec (Défossez et al., 2022) builds upon a framework similar to SoundStream. Nonetheless, it further augments its capabilities by integrating supplementary LSTM (Hochreiter and Schmidhuber, 1997) layers and harnessing a Transformer-based language model (Vaswani et al., 2017) to model the RVQ codes, thereby amplifying its sequence modeling performance. Then, there is a stream of work aimed at making codec models more general and powerful. AudioDec (Wu et al., 2023) represents an enhanced version of Encodec, implementing a group convolution mechanism to facilitate real-time operation of the streamable network, while also harnessing the capabilities of HiFi-GAN (Kong et al., 2020) to effectively generate high-fidelity audio at a sampling rate of 48 kHz.

AcademiCodec: In the AcademiCodec model introduced by (Yang et al., 2023a), a novel technique known as group-residual vector quantization is presented. This technique is tailored explicitly for generation tasks. It aims to enhance the reconstruction performance using a limited number of codebooks, consequently achieving an impressively low bit rate per second (BPS). This low BPS is of utmost significance as it effectively addresses the challenge of lengthy speech tokens in speech-language modeling, resulting in reduced sequence lengths.

SpeechTokenizer: SpeechTokenizer (Zhang et al., 2023a) is a unified speech tokenizer designed for speech-language models. It implements an Encoder-Decoder architecture enhanced with RVQ. By integrating semantic and acoustic tokens, SpeechTokenizer hierarchically separates various facets of speech information across different RVQ layers. Specifically, SpeechTokenizer is designed to regularize the first RVQ layer to learn the Hubert tokens (Hsu et al., 2021). The authors claim that employing such techniques can lead to improved disentanglement of information across various RVQ layers.

Descript-Audio-Codec: Descript-audio-codec (DAC) (Kumar et al., 2023), another instance of a universal neural codec model, distinguishes itself through its exceptional ability to maintain high-fidelity audio quality across a broad spectrum of data types, encompassing audio, music, and speech. It accomplishes this feat by employing a multitude of training techniques, such as periodic activation functions (Ziyin et al., 2020), enhanced residual vector quantization using factorized and L2-normalized codes, random quantizer dropout to preserve audio reconstruction quality, as well as refining adversarial and reconstruction loss during the training process. Out of all the techniques employed, they emphasize the pivotal role played by the periodic activation function.

FunCodec: Unlike most models focusing on the time domain, FunCodec (Du et al., 2023) proposes a frequency-domain codec. The authors claim they can achieve comparable performance with fewer parameters and lower computation complexity. Meanwhile, it also finds that incorporating semantic information in the codec tokens improves speech quality at low bit rates.

B Additional experiment results

(Due to the space limitation, please refer to the next page.)

Codec	Librispeech	Fluent Speech Commands	QUESST	VoxCeleb1	SNIPS	IEMOCAP	Libri2Mix	CREMA D	LibriCount	VoxLingua107Top10	Overall
A	0.713	0.677	0.637	0.676	0.697	0.629	0.637	0.586	0.566	0.651	0.644
B1	0.679	0.633	0.615	0.632	0.666	0.579	0.589	0.550	0.544	0.641	0.610
B2	0.678	0.638	0.625	0.637	0.655	0.587	0.601	0.568	0.555	0.647	0.617
B3	0.673	0.633	0.618	0.631	0.662	0.584	0.592	0.555	0.549	0.643	0.611
C	0.638	0.634	0.599	0.612	0.639	0.589	0.582	0.544	0.529	0.610	0.596
D1	0.824	0.822	0.801	0.822	0.808	0.820	0.814	0.764	0.748	0.764	0.798
D2	0.873	0.876	0.907	0.874	0.868	0.898	0.876	0.834	0.822	0.818	0.864
D3	0.805	0.810	0.807	0.808	0.794	0.815	0.796	0.807	0.778	0.802	0.802
E1	0.592	0.599	0.590	0.589	0.596	0.565	0.570	0.561	0.542	0.595	0.579
E2	0.652	0.648	0.654	0.652	0.656	0.621	0.629	0.611	0.594	0.650	0.636
E3	0.708	0.697	0.716	0.718	0.716	0.678	0.700	0.672	0.661	0.705	0.697
E4	0.751	0.738	0.770	0.776	0.764	0.734	0.762	0.726	0.721	0.747	0.748
E5	0.773	0.761	0.798	0.802	0.787	0.767	0.793	0.755	0.750	0.768	0.775
F1	0.764	0.732	0.730	0.748	0.764	0.749	0.740	0.672	0.656	0.702	0.724
F2	0.759	0.726	0.706	0.729	0.771	0.733	0.707	0.641	0.608	0.694	0.704
F3	0.750	0.732	0.708	0.747	0.775	0.731	0.677	0.648	0.609	0.704	0.705
F4	0.736	0.700	0.690	0.723	0.753	0.684	0.656	0.607	0.587	0.686	0.678
F5	0.781	0.747	0.738	0.776	0.793	0.750	0.698	0.671	0.618	0.723	0.726
F6	0.773	0.736	0.735	0.769	0.769	0.732	0.696	0.661	0.617	0.723	0.718

Table 6: Signal level overall scores across speech datasets

Codec	ESC-50	FSD50K	Gunshot Triangulation	Vocal Imitations	Overall
A	0.575	0.577	0.594	0.580	0.581
B1	0.566	0.562	0.594	0.573	0.574
B2	0.567	0.563	0.595	0.573	0.574
B3	0.585	0.576	0.620	0.588	0.592
C	0.599	0.597	0.614	0.598	0.602
D1	0.583	0.585	0.606	0.589	0.591
D2	0.634	0.628	0.652	0.630	0.636
D3	0.699	0.705	0.707	0.699	0.702
E1	0.596	0.592	0.605	0.585	0.594
E2	0.600	0.595	0.611	0.589	0.599
E3	0.604	0.599	0.614	0.593	0.602
E4	0.609	0.602	0.618	0.595	0.606
E5	0.612	0.604	0.623	0.597	0.609
F1	0.574	0.579	0.595	0.582	0.582
F2	0.576	0.580	0.593	0.583	0.583
F3	0.576	0.578	0.594	0.578	0.581
F4	0.575	0.577	0.585	0.577	0.578
F5	0.576	0.578	0.596	0.581	0.583
F6	0.577	0.579	0.594	0.580	0.583

Table 7: Signal level overall scores across audio datasets

Codec	OpenSinger	m4singer	VocalSet	GTZAN Genre	GTZAN Music Speech	Overall
A	0.717	0.711	0.565	0.481	0.531	0.585
B1	0.714	0.730	0.653	0.492	0.527	0.601
B2	0.709	0.719	0.650	0.564	0.576	0.630
B3	0.704	0.710	0.649	0.489	0.548	0.604
C	0.680	0.679	0.579	0.473	0.514	0.572
D1	0.840	0.840	0.731	0.726	0.685	0.749
D2	0.878	0.866	0.783	0.793	0.785	0.815
D3	0.806	0.798	0.789	0.706	0.752	0.770
E1	0.628	0.616	0.519	0.499	0.580	0.568
E2	0.680	0.663	0.581	0.548	0.632	0.621
E3	0.724	0.699	0.625	0.589	0.697	0.669
E4	0.762	0.730	0.658	0.623	0.756	0.710
E5	0.784	0.749	0.675	0.642	0.784	0.732
F1	0.773	0.765	0.672	0.557	0.627	0.667
F2	0.796	0.796	0.705	0.560	0.585	0.668
F3	0.786	0.772	0.672	0.533	0.581	0.649
F4	0.775	0.766	0.658	0.512	0.555	0.632
F5	0.820	0.806	0.703	0.537	0.582	0.665
F6	0.789	0.797	0.686	0.549	0.609	0.667

Table 8: Signal level overall scores across music datasets

Figure 5: Speech overall score correlation

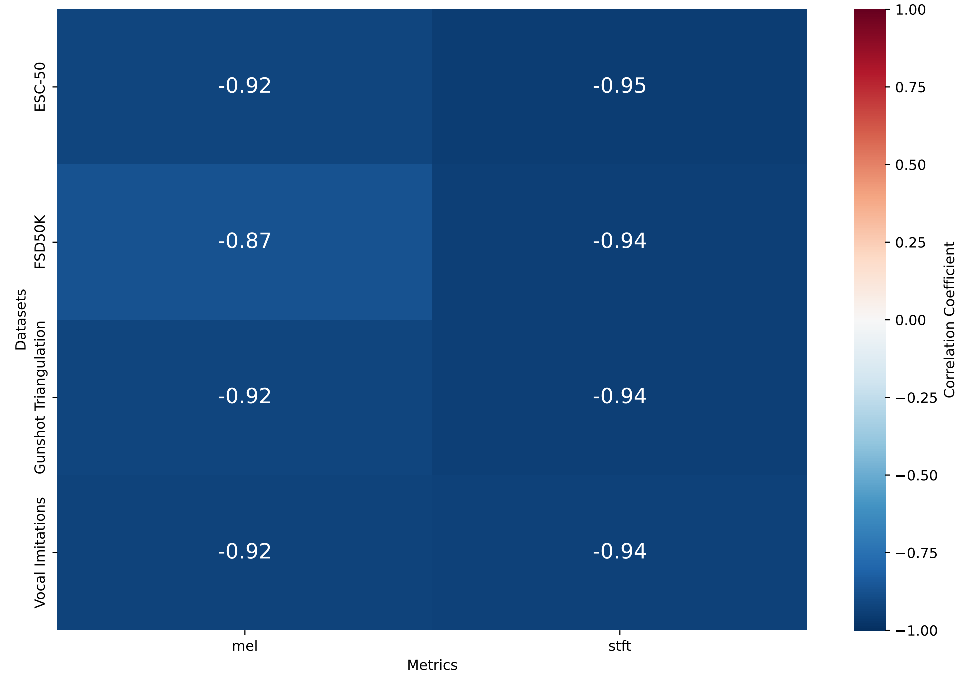


Figure 6: Audio overall score correlation

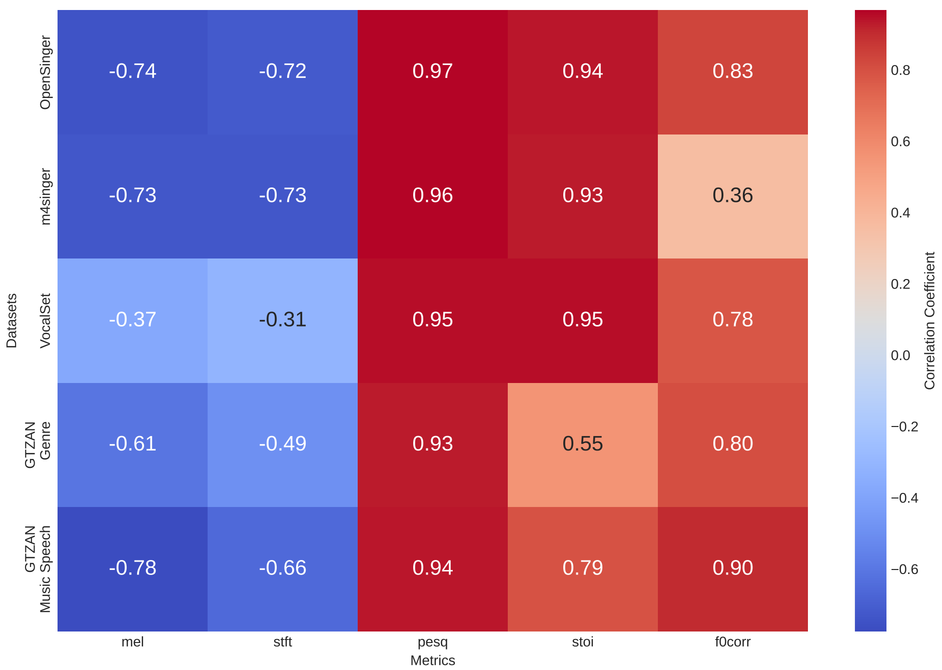


Figure 7: Music overall score correlation