



Analysis of Semantic and Acoustic Token Variability Across Speech, Music, and Audio Domains

Takanori Ashihara, Marc Delcroix, Tsubasa Ochiai, Kohei Matsuura, Shota Horiguchi

NTT, Inc., Japan

takanori.ashihara@ntt.com

Abstract

Techniques for discrete audio representation, which convert an audio signal into a sequence of audio tokens using neural audio codecs or self-supervised speech models, have gained attention for offering the possibility of modeling audio with large language models (LM) efficiently. While these audio tokens have been studied in various domains (e.g., speech, music, and general sound), their encoding properties across domains remain unclear. This paper examines several audio token types to analyze cross-domain variations. Our major findings include that audio tokens exhibit consistent statistical structures and probabilistic predictability deduced from rank-frequency distribution and perplexity, regardless of the domain. However, the token usage pattern is somewhat domain-dependent. This result underpins the steady success of the versatile audio LM, while also suggesting that domain-aware LM could further optimize performance by better capturing domain-specific token usage distributions.

Index Terms: speech, music, audio, neural audio codec, self-supervised learning, acoustic token, semantic token

1. Introduction

Recent audio-processing models have increasingly shifted towards using learned features as inputs, rather than relying on hand-crafted features (e.g., log mel-filterbank outputs) [1, 2]. In particular, the discretized representation of audio waveforms (hereinafter, audio tokens) has garnered significant interest recently [2–8]. Indeed, audio tokens can serve as an efficient gateway to language models (LMs), bridging the audio-text modality gap as well as improving efficiency in transmission, data storage, and training cost.

Audio tokens can be broadly categorized into two groups: acoustic tokens and semantic tokens [9]. *Acoustic tokens* refer to general-purpose codes produced by neural audio codecs (NACs) [2, 10–12], which are designed to compress audio waveforms across various domains, including *speech*, *music*, and general sound event (hereafter referred simply to as *sound*). These NAC models typically use an encoder-decoder architecture with residual vector quantization (RVQ), which consists of multi-level codebooks, trained with a reconstruction loss. On the other hand, *semantic tokens* are derived from discretized representation obtained by clustering hidden vectors from self-supervised learning (SSL) speech models [1, 13, 14], which typically consist of multi-layer Transformer blocks trained with a masked prediction task. Although speech SSL models were originally developed for speech-processing tasks, several studies have explored their applicability in other domains [7, 15].

Despite rapid progress in audio tokens, a fundamental understanding of their properties and domain-specific statistical and structural variations remains limited. Since audio tokens

function as a universal representation across different domains (speech [16, 17], music [18, 19], and sound [20, 21]), their widespread applicability raises fundamental questions: *how do their statistical and structural properties vary across domains?* and *why can music and sound be treated similarly to speech, given that their temporal structures do not convey the same linguistic meaning?*

While some versatile audio LMs [22] have demonstrated steady success, suggesting at a high level that similar modeling mechanisms can be applied across domains, to the best of our knowledge, this has not been systematically validated through a lower-level analysis of the statistical properties of tokens across domains. A better understanding of domain variations would provide insights into whether a domain-aware audio LM or audio tokenization should be given consideration. Conversely, if different domains complement each other, this synergy could lead to cross-domain generalization and potential performance improvements, similar to those observed in multilingual models [23].

To address these questions, we conduct a comprehensive analysis of audio token properties using frameworks widely employed in natural language processing, under controlled and standardized experimental conditions with ESPnet-Codec [24]. First, to elucidate fundamental statistical patterns across different domains, we examine rank-frequency distributions of audio tokens. While a previous study investigated Zipf’s law for semantic tokens based on a speech dataset and demonstrated that these tokens follow a power law [25], it remains unclear whether similar statistical patterns hold for audio tokens in different domains. Next, to assess the predictability of audio tokens, we compute the perplexity (PPL) at the codebook or layer levels. Since non-speech audio may not inherently exhibit the structured linguistic patterns observed in speech, this raises the question of the extent to which these differences exist and how unpredictable their patterns are. Subsequently, to investigate how token usage distributions contribute to the observed variations in PPL, we conduct a more detailed analysis of token usage patterns using the cosine similarity of term frequency-inverse document frequency (TF-IDF) representations.

Our major findings are as follows: (1) As inferred from Zipf’s law and PPL evaluations, statistical patterns and predictability indicate that domain differences are minor, with similar patterns emerging not only for speech but also for music and sound. (2) While the above analysis shows minor differences, a closer look at token usage reveals distributional differences across domains. (3) Based on these observations, the token usage vector computed using TF-IDF differs across domains, particularly in the 1st codebook of RVQ for acoustic tokens, suggesting that this classical method could potentially be applied for domain classification in an unsupervised manner.

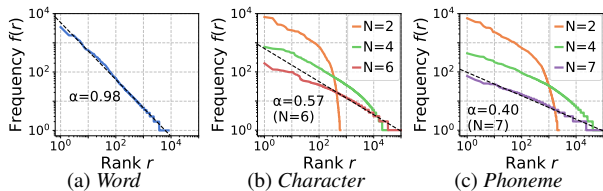


Figure 1: Rank vs. frequency of text symbols in the LibriTTS test-clean set [33]. N denotes the N -gram order. Each black dashed line shows a linear least-squares regression fit.

2. Related work

Some studies [26–29] have investigated the relationship between semantic tokens and phonetic or linguistic information. A previous study [30] demonstrated the information completeness and accessibility of semantic tokens before and after RVQ, showing that speaker and phonetic information is sufficiently present in both deep and shallow layers. Another study [31] evaluated EnCodec [11] with different numbers of codebooks in speech-processing tasks and demonstrated that tokenization behaves like a low-pass filter through the analysis of resynthesized speech. A separate study [32] used an N -gram LM with semantic tokens to extract acoustically similar subsets of pre-training speech data relative to a target speech corpus, effectively reducing SSL costs. Furthermore, through benchmark evaluations [4–6, 24], we can determine the effectiveness of audio tokens. In contrast, this paper focuses on the fundamental characteristics across different domains, rather than focusing on a single domain or higher-level performance evaluations.

3. Analysis methodology

In this paper, we apply fundamental methodologies—Zipf’s law, PPL, and TF-IDF-based cosine similarity—to analyze differences in audio token characteristics across domains.

Zipf’s law: A dataset is said to follow a power law if each item’s rank r and frequency $f(r)$ satisfy $f(r) \propto r^{-\alpha}$. It is commonly observed in natural language that a few words appear very frequently while many words appear rarely, leading to Zipf’s law with $\alpha \approx 1$, which is a special case of a power-law distribution. For example, the log-log plot of word rank-frequency distribution on the test-clean portion of LibriTTS [33] exhibits a linear relationship with $\alpha \approx 1$ (Fig. 1(a)). Similarly, we can also observe a power-law distribution for characters with $N = 6$ and for phonemes with $N = 7$, as shown in Figs. 1(b) and 1(c), respectively. Note that those values correspond to the average number of characters and phonemes per word in the dataset. These distributions suggest that individual characters, phonemes, or their N -grams with low N are subject to weaker or no syntactic and semantic constraints, resulting in less bias in their occurrences and a more uniform distribution. In contrast, N -grams appear with disproportionately high frequencies due to stronger constraints, leading to a heavy-tailed distribution. We use this empirical principle to investigate whether the fundamental statistical distributions in the music or audio domains resemble those observed in the text/speech domain [25], where the N -gram distribution of semantic tokens exhibits a power-law behavior. If a similar distribution emerges, token sequences (N -grams) may represent meaningful units governed by specific constraints, such as chords and notes in music [34]. This insight could help determine the appropriate context size for efficiently processing audio data across domains.

Perplexity: PPL is a measure of uncertainty in a probabilistic model, commonly used in language modeling. It quantifies

how well a probability distribution predicts a given sequence. Lower PPL indicates better predictive performance. Mathematically, given a model m , PPL $P(m)$ is related to cross-entropy (CE) $H(m)$ as: $H(m) = \log P(m)$. We use this measure to investigate how similarly predictable the music and audio domains are compared to the speech domain. Note that this study employs an N -gram LM with Kneser-Ney smoothing, trained and evaluated in terms of PPL using the KenLM toolkit [35].

Cosine similarity based on TF-IDF: TF-IDF is defined as the product of TF and IDF. TF measures how often a symbol appears in a dataset, while IDF downweights commonly occurring symbols across datasets. TF-IDF converts symbol sequences into feature vectors: Consequently, it enables similarity analysis across domains using cosine similarity between feature vectors. In this paper, we apply it to audio tokens to examine whether the same or different tokens are used across domains. Specifically, we compute TF-IDF across all three domains to mitigate the influence of frequently occurring tokens (e.g., silent intervals) and compare the cosine similarities between each pair of domains.

4. Experimental setup

4.1. Dataset

We used 960 hours of audio from three domains, *speech*, *music*, and *sound*, to train the N -gram LM and its subset to generate semantic tokens via k -means clustering. For *speech*, we used the training set of LibriSpeech [36]. For *music*, we used the training set of MUSDB18 [37], augmenting it with randomly selected samples from MTG-Jamendo [38]. Each music track in MUSDB18 and MTG-Jamendo was chunked into 10 seconds to align with the other datasets. For *sound*, we combined the entire balanced training set of AudioSet [39] with randomly selected excerpts from the unbalanced training set. For AudioSet, we excluded samples labeled as speech or music to maintain a clear distinction from the speech and music domains. We also used these datasets in three domains to compute the rank-frequency distribution for each domain.

For PPL and TF-IDF analyses, we randomly selected up to 1,000 samples per domain. Specifically, we used samples from the test-clean subset of LibriTTS [33] for *speech*, the test set of MUSDB18 [37] for *music*, and the evaluation set of AudioSet [39] for *sound*. These samples were also used to analyze token usage distribution, as discussed in Section 5.2.

4.2. Models for audio token

For semantic token extraction, we used HuBERT BASE [13] trained on LibriSpeech [36] sampled at 16 kHz.¹ To discretize features extracted using pre-trained HuBERT to form semantic tokens, we constructed a codebook of 1,024 codewords using k -means clustering with $k = 1024$. For clustering, we sampled 3.33% of the training data from each domain, resulting in approx. 96 hours in total—a size comparable to that used in HuBERT pre-training [13]. The sequential repetitions of semantic tokens were removed as in the previous research [3]. Other procedures for generating semantic tokens followed ESPnet [40].

We used the pre-trained models provided by ESPnet-Codec [24] for acoustic tokens to maintain consistency in training data and ensure fair comparisons. Specifically, we used EnCodec [11], and Descript Audio Codec (DAC) [12] as the NAC models², each trained on either LibriTTS [33] (only *speech*) or

¹We internally evaluated WavLM Base/Base+ [14], and obtained a similar trend to HuBERT, so we report only the results of HuBERT.

²We also evaluated SoundStream [10], showing the same trend.

AMUSE [24] (all three domains). Since this study aims to analyze the characteristics of these off-the-shelf models and fairly compare them with semantic tokens, all NAC models accept 16 kHz audio.³ All NAC models comprised 32 codebooks based on RVQ, each containing 1,024 entries of 512-dimensional vectors. Note that all the semantic and acoustic tokens were generated in frames of 25ms context with a 20ms shift.

5. Experimental results

5.1. Do tokens follow Zipf’s law in non-speech domains?

First, to assess Zipf’s law in semantic tokens, we plot token frequency against rank in Fig. 2. We analyzed tokens derived from the output from HuBERT’s convolutional encoder as well as the 6th and 12th Transformer blocks. Tokens from the 6th Transformer block follow a power law as reported in the previous study [25], and we further found that tokens from the convolutional encoder and the 12th Transformer block exhibit a similar trend, as shown in Fig. 2(a). This suggests that even shallow layers can capture meaningful units in N-grams, which supports the previous findings that discrete units derived from lower-layer features achieved reasonable error rates in phoneme recognition [30]. We also found a similar trend for tokens extracted from music and sound, both of which exhibit a distinct power-law distribution for $N \geq 4$ as in Figs. 2(b) and 2(c). This indicates that short-time token chunks may serve as word-like meaningful units, e.g., chords or notes in the music domain [34].

We also performed a similar analysis using acoustic tokens obtained from the 1st codebook. As shown in Fig. 3, they followed a power-law distribution, or one approximating Zipf’s law, when $N \geq 4$ across all domains and models. This trend remains consistent even for much larger N-gram orders, such as $N = 24$, which can be attributed to the absence of de-duplication in acoustic tokens, unlike semantic tokens. Since acoustic tokens would directly represent the waveform on a frame-by-frame basis, they reflect the inherent stationarity and redundancy of audio signals, preserving the power-law trend even at large N . For example, a prior study using hand-crafted spectral-based tokens [41] found that the occurrence of each frequency bin follows Zipf’s law, suggesting that similar characteristics might also exist in the NAC model, albeit at a lower bitrate. In fact, we further analyze the input waveforms corresponding to high-rank N-gram sequences and found that these N-grams were primarily associated with silences between vocalizations⁴ in speech, basic drum sounds in music, and low-frequency hum in sound.

5.2. How predictable are tokens in non-speech domains?

Figs. 4 and 5 show the PPL values of semantic and acoustic tokens for each layer and codebook level, respectively. In both figures, the top-left panel illustrates the PPL values computed from the speech dataset, and the rest panels present the normalized values, defined as the difference of CE between N-grams and unigrams, to emphasize sequential predictability over mere frequency-based predictability. Consequently, if the PPL value of an N-gram LM is lower than that of a unigram LM, the normalized value is high and reaches zero when the PPL values are equal. According to the results, semantic tokens in all layers and acoustic tokens at the 5th and shallower codebook levels

³Publicly available at https://huggingface.co/espnet/{amuse, libritts}_{encodec, dac}_16k.

⁴Different token indices (e.g., 1022, 730, 200, or 297 for DAC-AMUSE) redundantly map to similar silence waveforms, making high-rank N-grams primarily associated with silence in the speech domain.

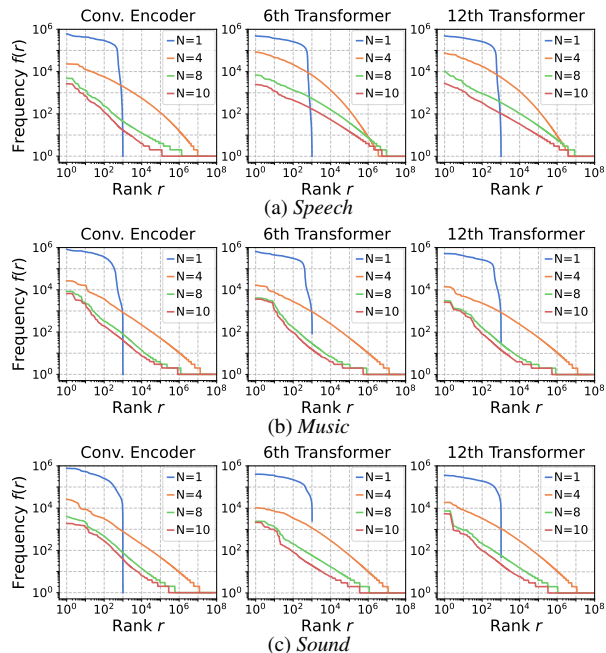


Figure 2: Rank vs. frequency of semantic tokens from HuBERT

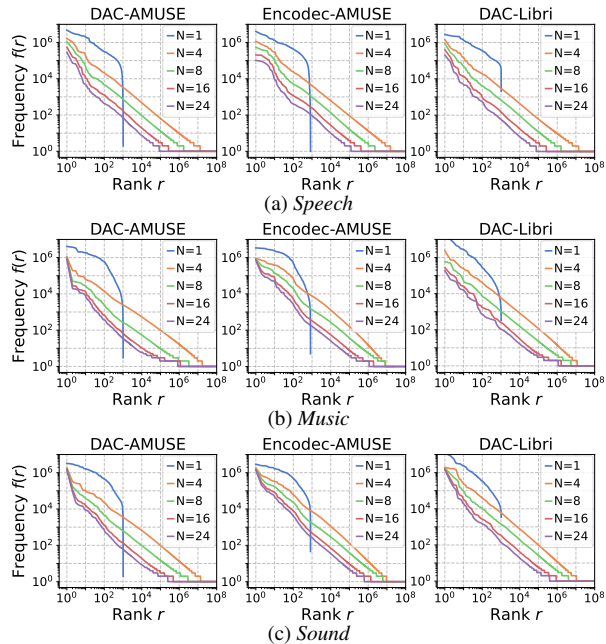


Figure 3: Rank vs. frequency of acoustic token of 1st codebook

exhibit lower PPL for N-grams than for unigrams. This trend is consistent across all three domains, indicating the existence of short-range dependencies and structured patterns in token sequences. This is expected in the speech domain, whereas we reveal that a similar trend appears in other domains, which may explain the effectiveness of multi-domain LLMs [22]. Notably, even semantic tokens (Fig. 4), which are exclusively trained on speech datasets, become more predictable when leveraging N-gram context than in the unigram setting, supporting the applicability of speech SSL models to the music domain [7, 15].

To explore the factors behind the PPL trends, we also present the token usage patterns in Figs. 6 and 7. They show token frequency as a function of token indices, where tokens are sorted by frequency within each layer or codebook. In

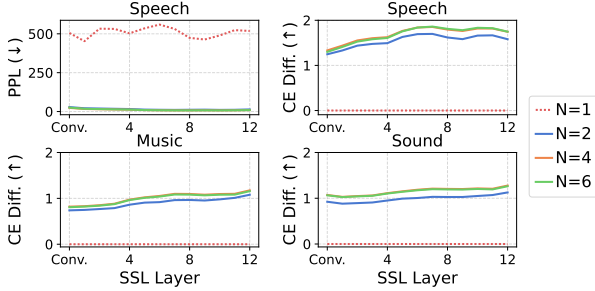


Figure 4: PPL of semantic token from HuBERT

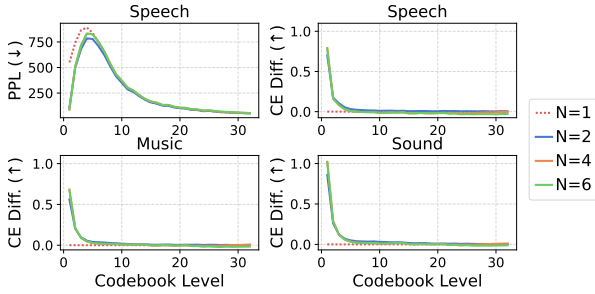


Figure 5: PPL of acoustic token from DAC-AMUSE

Fig. 6, the top 512 semantic token indices in the speech domain are highlighted in red, while the remaining indices are shown in blue. This figure indicates that the music and audio domains utilize not only domain-specific tokens but also tokens frequently used in the speech domain, increasing the difficulty of sequence prediction and resulting in lower normalized CE values in Fig. 4. Note that the usage distribution shows little variation across layers, resulting in plots with a consistent pattern and minimal dispersion.

For acoustic tokens, Fig. 7 illustrates the token usage patterns, where the lines progressively thin from the 1st-level codebook to the final level. To enhance the visibility of early-level transitions, a power-law normalization is applied to the color gradient. Each column represents a different NAC model, while each row corresponds to a different domain. As shown in Fig. 7, token usage changes as the codebook levels of RVQ increase: it starts with moderate diversity, reaches its peak, and finally converges to a small set of specific tokens across domains. This pattern aligns with the emergence of a PPL peak around the 5th codebook level and a gradual decrease in PPL differences between unigrams and N-grams as the codebook levels approach their final stages. We also confirm that DAC exhibits more diverse token usage than Encodec, consistent with the findings of the original DAC paper [12], which introduced a technique to enhance token usage. In AMUSE-trained models, the 1st codebook demonstrates distinct usage patterns, while usage in subsequent codebooks becomes nearly identical across domains. This suggests that the 1st codebook serves as the primary source of domain differences. Conversely, in models trained on LibriTTS, evaluation in non-speech domains would result in larger residual components from earlier codebooks, leading to more diverse utilization in later codebooks.⁵

5.3. How do frequency patterns differ across domains?

As discussed in Section 5.2, we observe domain differences in token usage for semantic and acoustic tokens, with the lat-

⁵Note that the internal evaluation of Encodec at 24 kHz showed greater token diversity than at 16 kHz, suggesting higher sampling rates require more complex codebook representations.

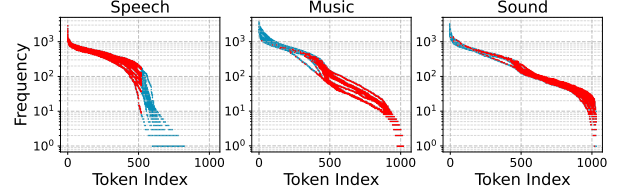


Figure 6: Semantic token usage at each layer. Red indicates the top 512 tokens in the speech domain; blue, all others.

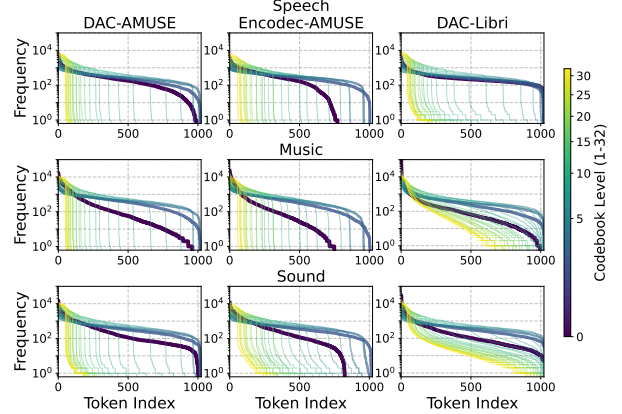


Figure 7: Acoustic token usage at each codebook level

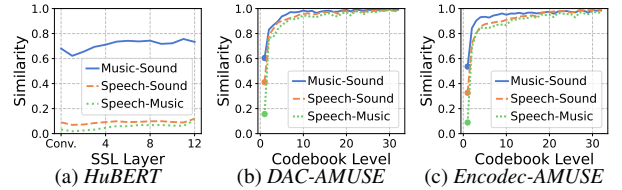


Figure 8: Cosine similarity of TF-IDF

ter showing notable variations, particularly at the 1st codebook level. To quantitatively capture these differences, we present a visualization of the cosine similarity of TF-IDF vectors at each layer or codebook level in Fig. 8. As shown in Fig. 8(a), the similarity for the semantic tokens is low between speech and other domains but high between music and audio across all layers, consistent with token usage in Fig. 6. In Fig. 8(b) and 8(c), the similarities show significant variation across domains at the 1st codebook level, but little to none at higher levels, aligning with Fig. 7 and indicating that residual components are minimally related to domains. These results imply that the 1st codebook level should be carefully handled when using tokens, such as in VALL-E [16]. Furthermore, this metric could be applied to fully unsupervised domain clustering or information retrieval in future research.

6. Conclusion

This paper provides an analysis for a deeper understanding of the nature of audio tokens across different domains. The results demonstrate that similar statistical and predictable sequence patterns are observed not only in speech but also in music and audio, while codeword usage varies across domains. We hope that these results will contribute to the development of more efficient audio tokens and the advancement of versatile audio language models.

7. References

- [1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, “Self-supervised speech representation learning: A review,” *JSTSP*, 2022.
- [2] 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:2402.13236*, 2024.
- [3] K. Lakhota, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, “On generative spoken language modeling from raw audio,” *TACL*, 2021.
- [4] X. Chang, J. Shi, J. Tian, Y. Wu, Y. Tang, Y. Wu, S. Watanabe, Y. Adi, X. Chen, and Q. Jin, “The Interspeech 2024 challenge on speech processing using discrete units,” in *Interspeech*, 2024.
- [5] P. Mousavi, L. Della Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, “DASB-Discrete Audio and Speech Benchmark,” *arXiv:2406.14294*, 2024.
- [6] H. Wu, X. Chen, Y.-C. Lin, K. Chang, J. Du, K.-H. Lu, A. H. Liu, H.-L. Chung, Y.-K. Wu, D. Yang *et al.*, “Codec-Superb @ SLT 2024: A lightweight benchmark for neural audio codec models,” in *SLT*, 2024.
- [7] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, L. Cagliero, P. Garza, and S. M. Siniscalchi, “Benchmarking representations for speech, music, and acoustic events,” in *ICASSP Workshops*, 2024.
- [8] Y. Su, J. Bai, Q. Xu, K. Xu, and Y. Dou, “Audio-language models for audio-centric tasks: A survey,” *arXiv:2501.15177*, 2025.
- [9] 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,” *TASLP*, 2023.
- [10] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An end-to-end neural audio codec,” *TASLP*, 2022.
- [11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *TMLR*, 2023.
- [12] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” in *NeurIPS*, 2023.
- [13] 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,” *TASLP*, 2021.
- [14] 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,” *JSTSP*, 2022.
- [15] Y. Ma, R. Yuan, Y. Li, G. Zhang, X. Chen, H. Yin, C. Lin, E. Benetos, A. Ragni, N. Gyenge *et al.*, “On the effectiveness of speech self-supervised learning for music,” in *ISMIR*, 2023.
- [16] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *TASLP*, 2025.
- [17] X. Wang, M. Thakker, Z. Chen, N. Kanda, S. E. Eskimez, S. Chen, M. Tang, S. Liu, J. Li, and T. Yoshioka, “SpeechX: Neural codec language model as a versatile speech Transformer,” *TASLP*, 2024.
- [18] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, “MusicLM: Generating music from text,” *arXiv:2301.11325*, 2023.
- [19] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” *NeurIPS*, 2024.
- [20] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, “AudioGen: Textually guided audio generation,” in *ICLR*, 2023.
- [21] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, “DiffSound: Discrete diffusion model for text-to-sound generation,” *TASLP*, 2023.
- [22] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, “UniAudio: An audio foundation model toward universal audio generation,” in *ICLR*, 2024.
- [23] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” in *Interspeech*, 2022.
- [24] J. Shi, J. Tian, Y. Wu, J.-W. Jung, J. Q. Yip, Y. Masuyama, W. Chen, Y. Wu, Y. Tang, M. Baali *et al.*, “ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech,” in *SLT*, 2024.
- [25] S. Takamichi, H. Maeda, J. Park, D. Saito, and H. Saruwatari, “Do learned speech symbols follow Zipf’s law?” in *ICASSP*, 2024.
- [26] T. A. Nguyen, B. Sagot, and E. Dupoux, “Are discrete units necessary for spoken language modeling?” *JSTSP*, 2022.
- [27] D. Wells, H. Tang, and K. Richmond, “Phonetic analysis of self-supervised representations of english speech,” in *Interspeech*, 2022.
- [28] A. Sicherman and Y. Adi, “Analysing discrete self supervised speech representation for spoken language modeling,” in *ICASSP*, 2023.
- [29] B. M. Abdullah, M. M. Shaik, B. Möbius, and D. Klakow, “An information-theoretic analysis of self-supervised discrete representations of speech,” in *Interspeech*, 2023.
- [30] S.-L. Yeh and H. Tang, “Estimating the completeness of discrete speech units,” in *SLT*, 2024.
- [31] K. C. Puvvada, N. Rao Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, “Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition,” in *ICASSP*, 2024.
- [32] Z. Lu, Y. Wang, Y. Zhang, W. Han, Z. Chen, and P. Haghani, “Unsupervised data selection via discrete speech representation for ASR,” in *Interspeech*, 2022.
- [33] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Interspeech*, 2019.
- [34] J. I. Perotti and O. V. Billoni, “On the emergence of Zipf’s law in music,” *Phys. A: Stat. Mech. Appl.*, 2020.
- [35] K. Heafield, “KenLM: Faster and smaller language model queries,” in *WMT*, 2011.
- [36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *ICASSP*, 2015.
- [37] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” 2017.
- [38] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The MTG-Jamendo dataset for automatic music tagging,” in *ICML*, 2019.
- [39] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio Set: An ontology and human-labeled dataset for audio events,” in *ICASSP*, 2017.
- [40] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Interspeech*, 2018.
- [41] M. Haro, J. Serra, P. Herrera, and A. Corral, “Zipf’s law in short-time timbral codings of speech, music, and environmental sound signals,” *Plos One*, 2012.