Analysing the Language of Neural Audio Codecs

Joonyong Park[1], Shinnosuke Takamichi[1,2], David M. Chan[3], Shunsuke Kando[1], Yuki Saito[1], Hiroshi Saruwatari[1]

[1]The University of Tokyo, [2]Keio University, [3]University of California, Berkeley.

joonyong-park@g.ecc.u-tokyo.ac.jp

Abstract—This study presents a comparative analysis of the statistical and linguistic properties of neural audio codecs (NACs). We investigate discrete speech tokens produced by various NAC models, examining their adherence to linguistic statistical laws such as Zipf's law and Heaps' law, as well as their entropy and redundancy. To assess how these token-level properties relate to semantic and acoustic preservation in synthesized speech, we evaluate intelligibility using error rates of automatic speech recognition, and quality using the UTMOS score. Our results reveal that NAC tokens, particularly 3-grams, exhibit language-like statistical patterns. Moreover, these properties, together with measures of information content, are found to correlate with improved performances in speech recognition and resynthesis tasks. These findings offer insights into the structure of NAC token sequences and inform the design of more effective generative speech models.

Index Terms—Self-Supervised Learning Model, Neural Audio Codec, Tokenization, Language modelling, Speech modelling.

I. INTRODUCTION

Speech discretisation allows models to take advantage of powerful sequence modelling techniques originally developed in natural language processing (NLP). Among such methods, neural audio codec (NAC) models have recently emerged as highly effective tools for modelling speech. They are capable of producing fine-grained token sequences capturing acoustic details essential for tasks such as automatic speech recognition (ASR), speech synthesis, and spoken language understanding [1], [2].

While NAC models have primarily been designed for efficient waveform compression and high-fidelity audio reconstruction, they are increasingly being integrated into generative and representation learning frameworks. Specifically, a speech token sequence derived from a pretrained NAC model serves as intermediate representations for downstream speech processing tasks. However, the linguistic and statistical properties of NAC-derived tokens remain underexplored. Determining whether these tokens exhibit language-like regularities is crucial, as it informs whether NLP-inspired or fundamentally new approaches are needed for generative speech modeling.

To address this open question, we conduct an in-depth analysis of NAC token sequences, comparing their statistical properties with those of natural languages. We focus on key linguistic regularities—including Zipf's law, Heaps' law, and entropy-based redundancy—to elucidate the extent to which NAC tokens adhere to, or deviate from, language-like statistical behaviours. Our analysis shows that these NAC-derived token sequences, particularly at the 3-gram level, closely follow natural language statistical laws, and this linguistic
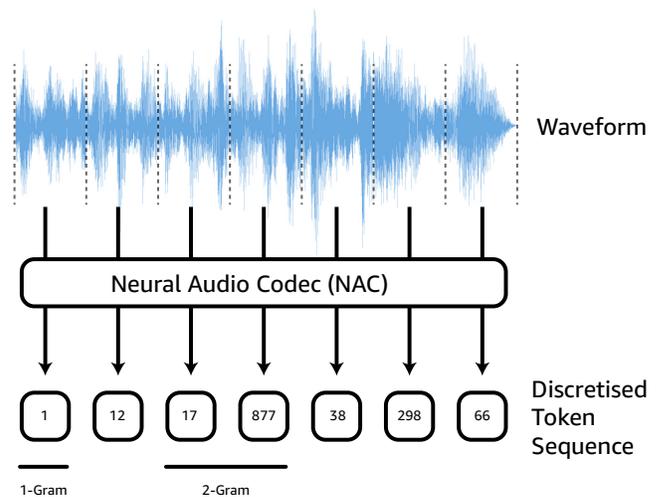

Fig. 1. Analysis of NAC tokens conducted in our paper. An NAC model segments input audio into short frames and assigns each a symbolic token from a learned codebook, forming "languages" of audio tokens.

regularity correlates with the improved objective evaluation metrics of resynthesised speech.

II. RELATED WORK

A. Discrete Tokens from Speech

Initially, self-supervised learning (SSL) models played a pivotal role in advancing speech tokenization. These models learn to produce discrete tokens directly from raw speech without relying on explicit phoneme labels or text transcriptions, by utilizing the discretisation methods such as k-means clustering. HuBERT [3] and wav2vec 2.0 [4] are well-known examples of these SSL models, whose token sequences have been shown to capture both phonetic and semantic information, thereby enhancing performance across a variety of speech processing tasks [5].

Building upon these foundations, NAC models, such as EnCodec [6] and SoundStream [7], have further constructed the process with the explicit goal of speech resynthesis. NAC models were initially developed to compress audio for efficient data transmission by lowering the bitrate, encoding detailed acoustic information into compact token sequences. These models are optimized for efficient compression and high-fidelity waveform reconstruction, demonstrating high effectiveness in processing the audio. This emphasis on preserving fine-grained acoustic detail distinguishes NACs from SSL-based approaches and makes them particularly well-suited for generative speech tasks that require high-quality audio output.

Moreover, NAC models have been successfully applied in a variety of downstream tasks. For example, in text-to-speech (TTS) synthesis, NAC tokens enable high-quality waveform generation from text-based inputs, contributing to improved audio fidelity and naturalness [6], [7]. NACs have also been integrated into ASR pipelines, where compressed token representations facilitate efficient and accurate transcription while reducing computational costs [8]. Furthermore, NAC models have been explored for speech separation tasks, operating directly in the compressed token space to achieve efficient source separation with lower computational requirements [9].

B. Statistical Laws of Language

In this study, we explore three primary statistical laws, models, and metrics inspired by NLP techniques.

Zipf's Law: In linguistics and information theory, it is well established that the frequency distribution of words or character n-grams in natural language tends to follow Zipf's law [10], characterised by a few high-frequency elements and many low-frequency elements. For example, the third most frequent word in an English document, "and," appears approximately one-third as often as the most frequent word, "the[1]." This relationship is given by the following power-law equation between the frequency rank r of a word and its frequency $f(r)$:

$$f(r) = a \cdot r^{-\eta}, \tag{1}$$

where $a > 0$ is a scaling constant and $\eta > 0$ represents the sharpness or scaling properties of the distribution. An ideal Zipfian distribution typically has $\eta \approx 1$. Such distributions have been often observed not only in natural languages but also in animal communication systems and large language models, making η a key indicator of the balance between redundancy and information efficiency [11], [12].

Heaps' Law: To capture the properties of a natural language in terms of the uniqueness of vocabulary, analyses based on Heaps' law [13] are also employed. This law describes a sublinear relationship between the number of words in document m and the vocabulary size $V(m)$, expressed as:

$$V(m) = K \cdot m^{\beta} \quad (0 < \beta < 1). \tag{2}$$

Here, $K > 0$ is a scaling factor that indicates the initial rate of vocabulary growth, while β determines the rate at which new vocabulary continues to be introduced as the text grows. A higher value of K suggests that the text introduces a relatively large number of unique words early on, whereas a lower value of K implies a more conservative initial vocabulary expansion. For values of β close to 1, the vocabulary grows almost linearly with the document size, implying a rich and diverse vocabulary set. For lower β values, the vocabulary growth is sublinear, reflecting a language that heavily reuses existing words and introduces new ones less frequently.

[1] https://www.cs.cmu.edu/~cburch/words/top.html

Entropy and Redundancy: Entropy, proposed by Shannon [14], is widely used as a measure of the diversity and unpredictability of information within a sequence and serves as a fundamental metric in evaluating linguistic systems and coding efficiency. The entropy H of a token sequence is defined as:

$$H = -\sum_{i=1}^{V} p_i \log_2 p_i, \tag{3}$$

where V is the vocabulary size and p_i is the probability of occurrence of the i-th token. Using entropy-based Huffman coding, it is possible to quantify the bit reduction rate and redundancy of a sequence. The average code length L achieved by Huffman coding can be evaluated with respect to the entropy as follows:

$$H \leq L < H + 1. \tag{4}$$

By computing the ratio between the actual average code length L and the entropy H, the compression efficiency or redundancy R can be quantified as:

$$R = \frac{L - H}{L}. \tag{5}$$

A smaller value of R indicates more efficient coding, reflecting lower redundancy and higher compressibility of the token.

C. Statistical Linguistic Analysis of Speech Tokens

There have been some previous attempts to analyse discrete speech tokens using this type of analysis. Takamichi et al. [15] analysed whether speech tokens generated by SSL models follow Zipf's law, as observed in natural language text. Their investigation revealed that these tokens exhibit power-law behaviour, which suggests that speech tokens possess certain linguistic structural properties. Moreover, Sicherman et al. [16] explored the interpretability and redundancy of SSL-based speech tokens and found a strong correlation between tokens and phonemes. They introduced a method for identifying redundancies that can degrade the performance of language models, and they proposed techniques to reduce such redundancies, leading to improvements in downstream tasks like speech synthesis and generative spoken language modelling. With respect to NAC-based tokens, Liu et al. [17] identified the challenge of inconsistency—different token sequences may be produced from the same speech input due to variations in the tokenization process. To address this, they introduced robust instruction-tuning strategies to enhance the consistency of tokens across different contexts, thereby improving the reliability of systems.

III. EXPERIMENTAL SETUP

Despite the above analyses, the question of whether NAC tokens inherit linguistic properties, such as statistical regularities or syntactic structures, remains underexplored. We thus aim to fill this gap by systematically analysing NAC token sequences to investigate their adherence to known linguistic properties and to clarify the relationships between speech metrics and linguistic structure in speech tokenization.

	Codec configurations	Training data	n_d	SR	kbps
A	16k	LibriSpeech [24]	8	16k	4
B1	hifi_16k_320d	LibriTTS [25],	4	16k	2
B2	hifi_16k_320d_large_uni	VCTK [26],	4	16k	2
B3	hifi_24k_320d	AISHELL [27]	4	24k	3
C	24k_320d	Valentini [28]	8	24k	6.4
D	24k	Same with E1∼E5	32	24k	24
E1	24k_1.5bps	Common Voice [29],	2	24k	1.5
E2	24k_3bps	DNSC [30],	4	24k	3
E3	24k_6bps	Jamendo [31],	8	24k	6
E4	24k_12bps	AudioSet [32],	16	24k	12
E5	24k_24bps	FSD50K [33]	32	24k	24
F1	en_libritts_16k_gr1nq32ds320		32	16k	16
F2	en_libritts_16k_gr8nq32ds320	Subset of LibriTTS [25]	32	16k	16
F3	en_libritts_16k_nq32ds320		32	16k	16
F4	zh_en_16k_nq32ds320	25k hours collected data	32	16k	16

This study investigates whether speech tokens obtained from various NAC models exhibit statistical properties similar to those of natural language. For the NAC models, we adopt six open-source codec models, where each of these models was trained under distinct conditions, including differences in datasets and bitrates, resulting in a total of 15 unique codec configurations shown in Table I for comparative analysis [23].

The extraction of token was conducted by the implementation of Codec-SUPERB [34]. To ensure consistency across models, we employed a standardised procedure for token sequence extraction and preprocessing. The speech signal was fed into a pretrained NAC model to extract a speech token sequence using a uniform codebook size of 1,024 for each dimension. Each token represented a 20 ms segment of input speech. A deduplication process was also applied to eliminate consecutive repetitions of the same token, thereby preventing potential biases in the statistical analysis, particularly in n-gram distribution calculations.

Furthermore, to account for differences in dimension sizes n_d caused by varying codec configurations across NAC models, we implemented an adjustment procedure. We treated each dimension of the NAC output as a separate token sequence and then concatenated them along with the time axis to create a single flattened token sequence. However, because each dimension uses an independent codebook but shares the same label space (e.g., 0 to 1,023), naive flattening would cause label collisions that obscure the actual distribution patterns. To resolve this, we offset each dimension's token IDs by adding an index shift proportional to the dimension order (e.g., the first dimension uses IDs 0–1,023, the second uses 1,024–2,047, and so on). This effectively expands the overall codebook size while preserving the uniqueness of tokens

across dimensions. Additionally, we prepared extra tokens that represented "dimension start" and "dimension end" and concatenated them at the beginning and the end of token sequences for each dimension, respectively. This process aims to mitigate undesired effects caused by artificial token overlap across codebooks during analysis.

The analysis was performed on 10 hours of single-speaker speech data for English (LJSpeech[2]) and Mandarin (Chinese CSS10 [35]) in order to investigate the language dependency of token. We used the manual transcriptions annotated to the corpora, and normalised through morphological analysis to obtain ground-truth (GT) word sequences for comparison.

IV. EXPERIMENTAL RESULTS

A. Language-based Statistical Analyses

For the speech token sequences obtained in this study, we generated n-gram sequences with $n = \{2, 3, 4, 6\}$ and conducted a series of statistical and structural analyses. Specifically, we evaluated the properties of the token sequences for each model on Zipf's law [10], Heaps' law [13], and calculations of entropy [14] and redundancy. Each token sequences were compared by the natural language baseline, which is a word 1-gram.

1) Token Frequency Distribution and Fit to Zipf's Law: In this experiment, we computed the frequency distributions of token n-grams extracted from various NAC models and evaluated their fit to Zipf's law [10] by analysing their rank-frequency distributions on a log-log scale.

For real-world data, when the tail is too heavy or too light, it is difficult to properly predict the shape of the data using general linear regression in Zipf's model [36]. Thus, to robustly estimate the parameters of Zipf's law under such conditions, we employed the `powerlaw` library [37], which fits the rank-frequency data using the maximum likelihood estimate:

$$\alpha = 1 + \frac{n}{\sum_{i=1}^{n} \log(x_i/x_{\min})},$$

where x_i is the frequency of the i-th ranked token in the empirical distribution and $x_{\min}$ works as a threshold, which is lower bound of the power-law behaviour in the rank-frequency distribution. Here, $\alpha = \eta + 1$. Values of α approaching 2 indicate a close adherence to Zipf's law. Also, the goodness-of-fit was evaluated using the Kolmogorov–Smirnov (KS) distance, where larger suggest deviation from linearity [38].

Figure 2 shows the results of the Zipf's law analysis. In this figure, each small dot represents one NAC model configuration, while the large dots represent the mean position of each n-gram group, calculated by averaging α and KS values. Distances from each n-gram mean to the GT were calculated using Z-score normalisation and displayed in the figure legend. The token frequency and distribution including α and KS-distance values can be seen in Figure 3. For the mean NAC token, the entire token outputs from the model used in this study are stacked to represent the distribution of
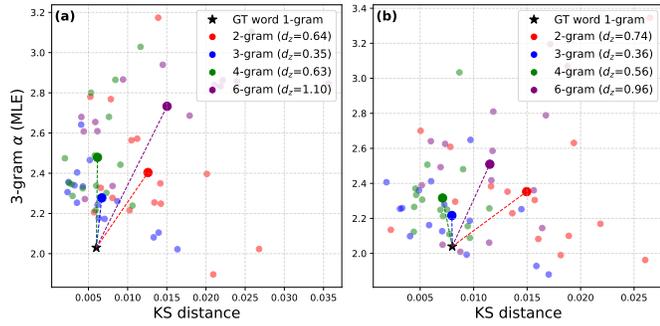
[2]https://keithito.com/LJ-Speech-Dataset/

Fig. 2. Relationship between KS distance and Zipf's law exponent (α) for NAC token sequences compared to natural language reference (word 1-gram) for (a) English and (b) Chinese.
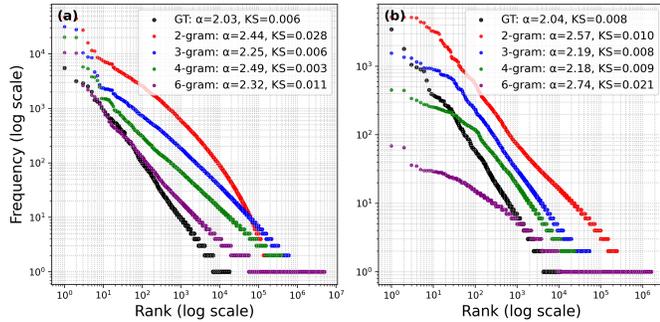

Fig. 4. Relationship between Heap's law components k and β for NAC token sequences compared to natural language reference (word 1-gram) for (a) English and (b) Chinese.

Fig. 3. Comparison of n-gram distributions of the mean NAC token sequences with natural language reference (word 1-gram) for (a) English and (b) Chinese.

Fig. 5. Comparison of n-gram token vocabulary growth of the mean NAC token sequences with natural language reference (word 1-gram) for (a) English and (b) Chinese.

the entire token sequence in a single dimension, then divided by the total number of dimensions in the model, resulting in a normalized token with a codebook size of 1,024.

It shows that 3-gram token sequences consistently provided the best fit across the NAC models, demonstrating the most linear behaviour and thus the strongest alignment with the Zipfian distribution. This indicates that, at the 3-gram level, NAC tokens exhibit a frequency distribution where high-frequency tokens are used relatively frequently while low-frequency tokens remain present, albeit at lower frequencies. This pattern holds for both English and Chinese data, as showing the linear distribution graph in Figure 3.

2) Token Novelty and Vocabulary Growth: To evaluate the novelty of speech tokens, we conducted analyses based on Heaps' law [13]. We analysed the natural language and the speech token sequences considering the token order, tracking the increase in unique tokens relative to the total number of tokens observed.

In Figure 4, we estimated both the scaling factor k and the growth exponent β for each model and n-gram setting. It is shown that 3-gram token sequences consistently display high similarities with those of natural language reference in both k and β values across both English and Chinese. Additionally, the variability of cluster assignment across different models was found to be substantial, indicating that some models tend to group tokens in ways that exhibit higher volatility, as reflected in the spread of k and β estimates.

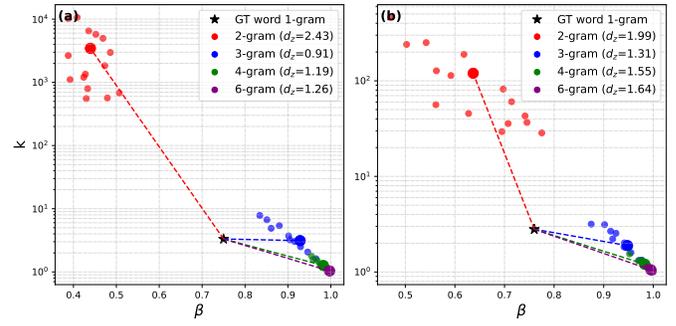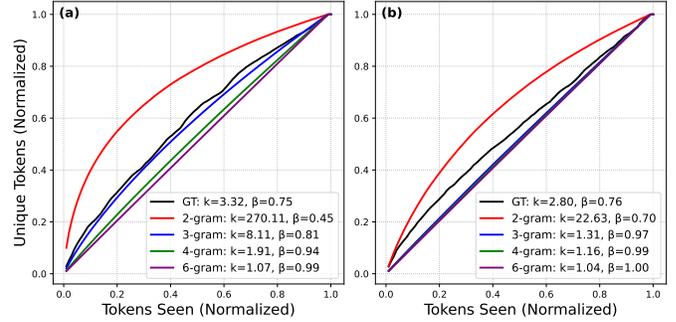The results indicated that token sequences with higher k

values tended to exhibit sublinear growth, indicating repeated reuse of common tokens and reduced diversity, as shown in Figure 5. Conversely, sequences with lower k values displayed linear or near-linear growth, reflecting a more uniform introduction of new tokens. Furthermore, we observed that as n increased, the complexity and diversity of the data also increased, leading to less repetitive patterns and larger vocabulary sizes.

3) Analysis of Token Entropy and Redundancy: Finally, we evaluated the entropy and compressibility of token sequences generated by NAC models and compared them with natural language sequences. Figure 6 shows the relationship between n-gram and the bit reduction rate. At the 1-gram level, natural language sequences exhibited a significantly higher bit reduction rate compared to NAC tokens. In contrast, NAC tokens displayed a relatively lower bit reduction rate at this level, suggesting a limited reuse of acoustic token patterns.

The bit reduction rate tend to drop sharply from 1-gram to 2-gram for natural language, while rather increased for token sequences. Interestingly, at the 2-gram and 3-gram levels, NAC token sequences showed higher bit reduction rates than natural language. This indicates that, at these intermediate n-gram levels, NAC token sequences contained acoustic patterns that frequently repeated, leading to increased redundancy that could be efficiently compressed using Huffman coding. However, at higher n-gram levels (4-gram and above), the bit reduction rate of NAC token sequences decreased sharply and fell below that of natural language. This suggests that NAC token sequences
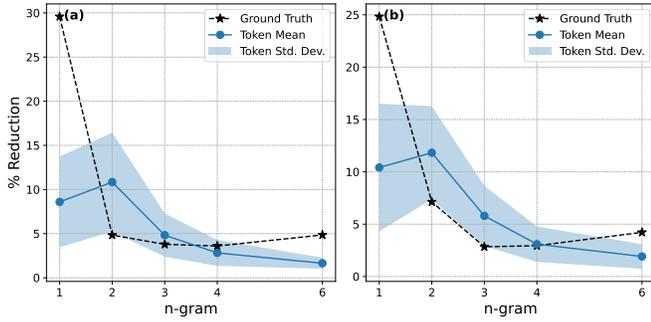
Fig. 6. Comparison of token bit reduction rates across different n-gram configurations for NAC model tokens versus natural language reference (word n-grams) for (a) English and (b) Chinese. The results for 1-gram NAC tokens are also shown.

become highly sparse at these higher n-gram levels, as tokens no longer exhibit repetitive structures but rather fragment into near-unique combinations. These observations indicate that NAC token shows a high compression ratio in lower n-grams, but the effect decreases as n becomes higher when modelling with more than 2-gram, in contrast to the natural language baseline. This difference indicates that NAC models may prioritize acoustic fidelity over statistical redundancy, resulting in less compressible token representations as n gets bigger.

Also, it can be observed that between 3-gram and 4-gram configurations, natural language and tokens exhibit similar bit reduction rates. This suggests that the following n-gram distribution shows similar token distributions between natural language words and tokens.

B. Comparison with Speech Performance Benchmarks

This section investigates the relationship between the statistical properties of NAC tokens and speech performance benchmarks. We specifically examine whether tokens exhibiting more "language-like" statistics correlate with improved intelligibility (WER/CER) and naturalness (UTMOS).

To evaluate this relationship, we used the Codec-SUPERB [34] implementation to decode the NAC tokens back into speech waveforms, generating resynthesised speech outputs for each codec model under study. For the intelligibility task, we then applied `whisper-medium` [39] ASR model, to transcribe the synthetic speech. By comparing these ASR outputs with the GT transcripts provided in the corpus, we calculated word error rate (WER) for English and character error rate (CER) for Chinese using the Levenshtein distance. This provided a quantitative measure of how closely the resynthesised speech aligns with the original linguistic content. Also, we assessed the naturalness of the synthetic speech using UTMOS [40], speech quality assessment metric that estimates the perceived naturalness of speech based on neural network architectures trained on large-scale human-annotated datasets. By applying UTMOS to the resynthesised speech samples, we were able to derive pseudo-naturalness scores that reflect how naturally the speech would be perceived by listeners. These benchmarks collectively allow us to analyse whether the statistical regularities observed in token sequences
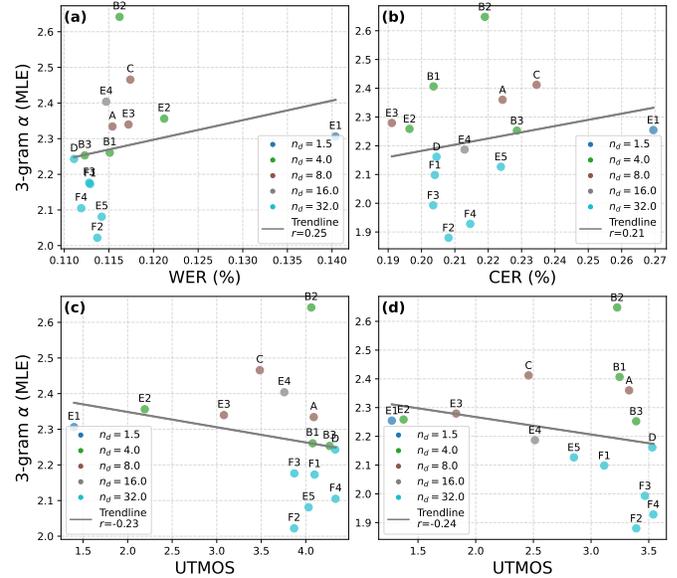

Fig. 7. Relationship between the 3-gram Zipf's law exponent (α) and (a) English WER, (b) Chinese CER, (c) English and (d) Chinese UTMOS for speech token sequences. Each point represents each codec configurations, color-coded by dimension size n_d. The trendline was obtained by Pearson correlation coefficient r.

meaningfully impact the performance of speech processing systems in real-world tasks.

1) Zipf's Law: We utilized 3-gram token sequences against the benchmarks to investigate the relationship between token frequency and the performances in downstream tasks.

Figure 7 shows scatter plots of α values derived from 3-gram token sequences against the benchmark performances. The results reveal a consistent trend: as α decreases, where approaching the Zipf's law's theoretical value of 2, error rates tend to decrease while UTMOS scores increase. In addition, it is showed that model configurations with large dimensions displayed lower error rates and higher UTMOS, while also demonstrating lower α. This suggests that token sequences exhibiting Zipfian behaviour with high-frequency token usage correlate with improved speech recognition accuracy and perceived naturalness.

Interestingly, we observed that this trend was most clearly observed with 3-gram token sequences, while analyses with 2-gram and 4-gram configurations did not show the same consistent relationship between α and benchmark performances.[3] This observation suggests that 3-gram sequences may strike an optimal balance between capturing local linguistic patterns and maintaining token diversity, which in turn aligns well with both statistical regularity and acoustic information preservation. In contrast, 2-grams may be too short to adequately represent contextual dependencies, while 4-grams may introduce excessive sparsity and volatility in token usage.

2) Heaps' Law: We analysed the correspondence between the Heaps' law exponent β and k—estimated from 3-gram token sequences—and speech benchmarks. Figure 4 shows

---

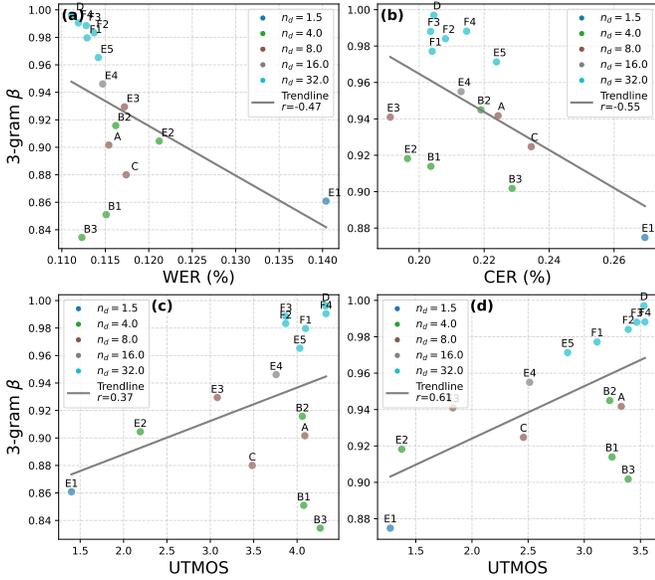[3]The exact values and graphs could not be included in the paper due to page limitations.

Fig. 8. Relationship between the 3-gram Heap's law exponent (β) and (a) English WER, (b) Chinese CER, (c) English and (d) Chinese UTMOS for speech token sequences. Each point represents each codec configurations, color-coded by dimension size n_d. The trendline was obtained by Pearson correlation coefficient r.
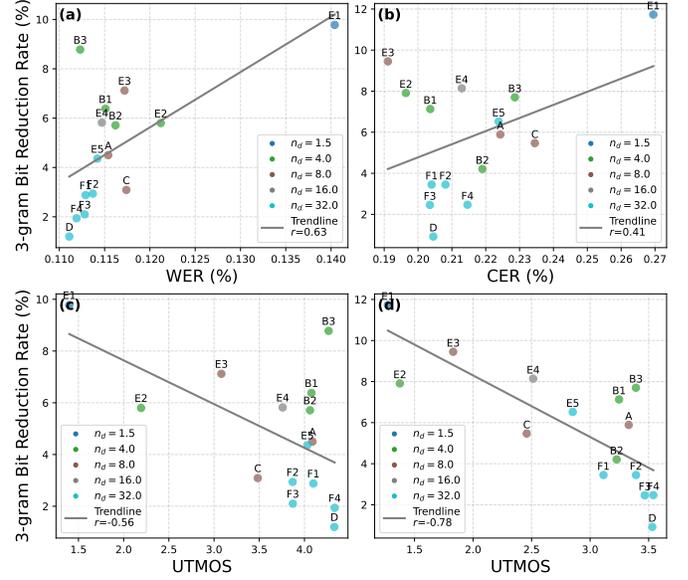

Fig. 9. Relationship between the 3-gram token bit reduction rate and (a) English WER, (b) Chinese CER, (c) English and (d) Chinese UTMOS for speech token sequences. Each point represents each codec configurations, color-coded by dimension size n_d. The trendline was obtained by Pearson correlation coefficient.

scatter plots of β values against the benchmark performances. The results reveal a consistent trend: as β approaches 1 (indicating a more linear growth of the vocabulary size with respect to the number of observed tokens), error rates (WER and CER) decrease, and UTMOS scores increase. It is also showed that model with a configurations of large dimensions displayed higher β. This finding suggests that token sequences exhibiting near-linear vocabulary growth—indicative of consistent token usage—correlate with improved speech recognition accuracy and perceived naturalness. This trend was robust across different n-gram configurations, with 2-grams to 6-grams consistently demonstrating that higher β values align with better benchmark performances, except for a few outliers. This reinforces the conclusion that token sequences capturing stable vocabulary expansion contribute positively to both recognition and synthesis.

Additionally, the scaling factor k in the Heaps' law formulation was also found to play an important role[3]. As k approached 1, meaning that the model effectively captures each token's presence without excessive redundancy, error rates decreased and UTMOS scores improved. These results collectively highlight that tokens adhering to linguistic patterns consistent with natural language.

3) Entropy: We analysed the bit reduction rate derived from 3-gram token sequences and its relationship with speech quality benchmarks. Figure 9 presents scatter plots showing these relationships. The analysis reveals a consistent trend: as the bit reduction rate approaches zero—indicating that the token sequences are less compressible—error rates decrease and UTMOS scores increase. It is also showed that model with a configurations of large dimensions displayed lower bit reduction rate. This suggests that token sequences with lower

redundancy and higher diversity are more closely aligned with the underlying structure of natural language and correspondingly support better speech quality. This pattern was also consistently observed across n-gram configurations from 2-grams to 6-grams[3], indicating that lower bit reduction rate across n-gram reliably correlate with improved recognition accuracy and higher perceived naturalness.

V. CONCLUSION

In this study, we explored in what cases speech token sequences from NAC models exhibit statistical properties similar to those observed in natural language. Overall, NAC tokens consistently aligned closely with natural language distributions at the 3-gram level, and trends in the Zipf's law exponent (α), Heaps' law parameters (β and k), and bit reduction rate all indicated that token sequences with higher linguistic regularity and lower redundancy contributed to improved speech representation quality. Together, these results indicate that certain linguistic statistical properties of NAC tokens, particularly at specific n-gram granularities, play an important role in bridging the gap between acoustic fidelity and linguistic structure in speech processing.

Beyond the contributions here, we believe that it is insightful to further explore the case of real-world, multi-speaker audio in noisy environments, and NACs designed for multilingual domains. We hope such explorations can yield further insights into how we can design token representations that balance linguistic regularity and acoustic detail, supporting high-quality speech generation and accurate language understanding.

REFERENCES

[1] Y. Guo, Z. Li, H. Wang, B. Li, C. Shao, H. Zhang, C. Du, X. Chen, S. Liu, and K. Yu, "Recent advances in discrete speech tokens: A review," *arXiv*, vol. arXiv:2502.06490, 2025.

[2] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[5] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[6] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

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

[8] K. Dhawan, N. R. Koluguri, A. Jukić, R. Langman, J. Balam, and B. Ginsburg, "Codec-ASR: Training performant automatic speech recognition systems with discrete speech representations," in *Proc. INTERSPEECH*, Kos, Greece, Sep. 2024, pp. 2574–2578.

[9] J. Q. Yip, S. Zhao, D. Ng, E. S. Chng, and B. Ma, "Towards audio codec-based speech separation," in *Proc. INTERSPEECH*, Kos, Greece, Sep. 2024, pp. 2190–2194.

[10] G. K. Zipf, *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.

[11] B. Mandelbrot, "Contribution à la théorie mathématique des jeux de communication," in *Annales de l'ISUP*, vol. 2, 1953, pp. 3–124.

[12] A. Gelbukh and G. Sidorov, "Zipf and heaps laws' coefficients depend on language," in *Proc. CICLing*, Mexico City, Mexico, 2001, pp. 332–335.

[13] H. S. Heaps, *Information retrieval: Computational and theoretical aspects.* Academic Press, Inc., 1978.

[14] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[15] S. Takamichi, H. Maeda, J. Park, D. Saito, and H. Saruwatari, "Do learned speech symbols follow Zipf's law?" in *Proc. ICASSP*, Seoul, South Korea, Apr. 2024, pp. 12 526–12 530.

[16] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *Proc. ICASSP*, Rhodes, Greece, Jun. 2023.

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2024.

[18] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech large language models," in *Proc. ICLR*, Vienna, Austria, 2024.

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

[20] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, Jun. 2023.

[21] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. NeurIPS*, New Orleans, U.S.A., Dec. 2023.

[22] Z. Du, S. Zhang, K. Hu, and S. Zheng, "FunCodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," in *Proc. ICASSP*, Seoul, South Korea, Apr. 2024.

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

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

[25] 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 *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1526–1530.

[26] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[27] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, Shanghai, China, Oct. 2017.

[28] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]," 2017, university of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR).

[29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proc. LREC*, Marseille, France, May 2020, pp. 4218–4222.

[30] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 Deep Noise Suppression Challenge," in *Proc. ICASSP*, Singapore, May 2022.

[31] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Proc. ICML*, Long Beach, U.S.A., Jul. 2019.

[32] 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 *Proc. ICASSP*, New Orleans, U.S.A., Mar. 2017, pp. 776–780.

[33] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[34] H. Wu, H.-L. Chung, Y.-C. Lin, Y.-K. Wu, X. Chen, Y.-C. Pai, H.-H. Wang, K.-W. Chang, A. Liu, and H.-y. Lee, "Codec-SUPERB: An in-depth analysis of sound codec models," in *Findings of ACL*, Bangkok, Thailand, Aug. 2024, pp. 10 330–10 348.

[35] K. Park and T. Mulc, "CSS10: A collection of single speaker speech datasets for 10 languages," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1566–1570.

[36] D. M. Chan, R. Corona, J. Park, C. J. Cho, Y. Bai, and T. Darrell, "Analyzing the language of visual tokens," *arXiv preprint arXiv:2411.05001*, 2024.

[37] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[38] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[39] A. Radford, K. Jong Wook, X. Tao, B. Greg, M. Christine, and S. Ilya, "Robust speech recognition via large-scale weak supervision." https://cdn.openai.com/papers/whisper.pdf, 2022.

[40] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 4521–4525.