一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

# ニューラルオーディオコーデックにおける雑音頑健性分析

## ～Zipf 則・Heaps 則に基づく言語統計構造と劣化音声の関係～

朴　　浚鎔[†*]　高道慎之介[††,†]　DavidM. Chan[†††]　神藤　駿介[†]　齋藤　佑樹[†]
猿渡　　洋[†]

† 東京大学大学院情報理工学系研究科
†† 慶応義塾大学理工学部
††† カリフォルニア大学バークレー校 Berkeley Artificial Intelligence Research Lab (BAIR)

**あらまし**　ニューラルオーディオコーデック（Neural Audio Codec; NAC）は音声を離散トークン列として表現し、擬似言語的な統計解析を可能にする。ただし、NAC トークンの統計構造が劣化音声下でどの程度崩れ、その崩壊が認識・知覚・音響品質の低下とどのように対応するかは体系的に検証されていなかった。本研究では、雑音付加や音声劣化条件下において、NAC トークンの言語統計構造がどの程度頑健に保たれるかを分析する。Zipf 則および Heaps 則を用いてトークン分布をモデル化し、クリーン音声からの分布的エラー率を定義することで、音声の可聴品質とは異なる「話し言葉レベルでの構造的崩れ」を捉える指標を検討する。その結果、雑音条件が混在する設定においてトークンの言語統計指標は既存の意味的・音響的指標との一貫した相関関係を示し、NAC トークンの語彙成長構造の崩壊が認識・知覚・音響歪みの広範な劣化を説明することができた。

**キーワード**　音声分析、ニューラルオーディオコーデック、音声の雑音頑健性、言語モデリング、音声モデリング

# Analysing the Noise Robustness of Neural Audio Codecs

## − Relationship between Language Statistical Structure and Degraded Speech Based on Zipf's Law and Heaps' Law −

Joonyong PARK[†*], Shinnosuke TAKAMICHI[††,†], David M. CHAN[†††],
Shunsuke KANDO[†], Yuki SAITO[†], and Hiroshi SARUWATARI[†]

† The University of Tokyo　Hongo 7–3–1, Bunkyo-ku, Tokyo, Japan
†† Keio University　Mita 2-15, Minato-ku, Tokyo, Japan
††† University of California, Berkeley　110 Sproul Hall, Berkeley, CA, USA

## 1. はじめに

音声の離散化は、自然言語処理分野で発展してきた強力な系列モデリング手法を音声モデルに適用可能にする。これらの手法の中でも、ニューラルオーディオコーデック（Neural Audio Codec; NAC）モデルは、近年、音声モデリングにおいて極めて有効な手法として注目を集めている。NAC モデルは、自動音声認識、音声合成、および音声言語理解といったタスクにおいて本質的となる音響的詳細を捉えた高解像度のトークン系列を生成可能であることが示されている [1], [2]

NAC モデルはもともと、高効率な波形圧縮と高忠実度な音声再構成を目的として設計されてきたが、近年では生成モデルや表現学習の枠組みにも積極的に統合されつつある。具体的には、事前学習済み NAC モデルから得られるトークン系列が、下流の音声処理タスクにおける中間表現として用いられている。したがって、NAC によって得られるトークンが、言語的・統計的構造を内包し得ることが示され、生成音声モデリングにおける言語的表現としての重要性から、その性質に関する基礎的な分析が近年進められてきた [3]。

しかしながら、NAC によって生成されるトークンが、雑音付

加や音声劣化といった実環境条件下において、言語統計的構造をどの程度安定に維持できるのかについては、未だ十分に検証されていない。実際の音声応用においては、入力音声はしばしば雑音や歪みを含み、こうした劣化がトークン系列の分布構造や言語的規則性に与える影響を理解することは極めて重要である。特に、Zipf 則や Heaps 則に代表される言語統計法則は、単なる分布の形状を超えて、冗長性・情報効率・構造的安定性といった言語の本質的特性を反映する指標である。したがって、劣化音声条件下においてこれらの統計則がどの程度保たれるかを分析することは、NAC トークンが「言語的構造を持つ表現」としてどれほど頑健であるかを評価する上で重要な視点となる。

本研究では、NAC トークンの言語統計的構造に対する雑音頑健性を体系的に分析する。具体的には、雑音レベルの増加に伴う NAC トークンの分布からの乖離を定量化し、分布偏差指標を用いてクリーン音声条件との比較を行う。さらに、これらの分布偏差が既存の音声評価指標とどのような関係を持つかを検証することで言語統計構造の劣化と音声性能低下との対応関係を明らかにする。その結果、雑音条件が混在する設定においてトークンの言語統計指標は既存の意味的・音響的指標との一貫した相関関係を示し、NAC トークンの語彙成長構造の崩壊が認識・知覚・音響歪みの広範な劣化を説明することができた。

## 2. 先 行 研 究

### 2.1 音声離散トークン

当初、自己教師あり学習（Self-Supervised Learning; SSL）モデルは、音声トークン化の発展において中心的な役割を果たしてきた。これらのモデルは、$k$-means クラスタリングなどの離散化手法を用いることで、音素ラベルやテキスト転写に依存することなく、生音声から直接離散トークンを学習する。HuBERT [19] や wav2vec 2.0 [20] は、その代表的な例であり、これらのモデルから得られるトークン系列が、音韻的情報と意味的情報の双方を捉えることが示され、多様な音声処理タスクにおいて性能向上をもたらしている [21]。

これらの基盤的研究を踏まえ、EnCodec [17] や SoundStream [22] に代表される NAC モデルは、音声の再合成を明確な目的として設計された。NAC モデルは、当初、ビットレートを低減することで効率的なデータ伝送を実現するために開発され、詳細な音響情報をコンパクトなトークン系列として符号化する。これらのモデルは、高効率な圧縮と高忠実度な波形再構成に最適化されており、音声処理において高い性能を示す。このような微細な音響情報の保持に重点を置く点が、SSL ベースの手法との大きな違いであり、高品質な音声出力が求められる生成音声タスクにおいて特に有効である。

さらに、NAC モデルは多様な下流タスクにおいても応用されている。例えば、テキスト音声合成においては、NAC トークンを用いることでテキスト入力から高品質な波形生成が可能となり、音声の忠実度および自然性の向上に寄与している [17], [22]。また、音声認識パイプラインにおいても、圧縮されたトークン表現を用いることで、計算コストを抑えつつ高精度な文字起こしを実現している [23]。さらに、音声分離タスクにおいても、圧縮トークン空間上で直接処理を行う手法が検討されており、計算効率の高い音源分離が可能であることが示されている [24]。

### 2.2 言語の統計法則

本研究では、計算言語学に着想を得た 2 つの主要な統計法則・モデル・指標を対象とする。

**Zipf の法則:** 言語学および情報理論の分野では、自然言語における単語や文字 $n$-gram の出現頻度分布が Zipf の法則に従うことが広く知られている [25]。これは、ごく少数の高頻度要素と、多数の低頻度要素から構成される分布である。例えば、英語文書において 3 番目に頻出する単語 “and” は、最頻出語 “the(注1).” のおよそ 3 分の 1 の頻度で出現する。この関係は、単語の頻度順位 $r$ とその頻度 $f(r)$ の間の以下のべき乗則で表される：

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

ここで、$a > 0$ はスケーリング定数、$\alpha > 0$ は分布の鋭さやスケーリング特性を表す指数である。理想的な Zipf 分布では、$\alpha \approx 1$ となることが多い。このような分布は、自然言語のみならず、動物のコミュニケーション体系や大規模言語モデルにおいても観測されており、$\alpha$ は冗長性と情報効率のバランスを示す重要な指標とされている [26], [27]。

本研究では、離散データに対するべき乗分布モデルを仮定し、最尤推定により $\alpha$ を推定する。さらに、推定された Zipf 分布と実データとの適合度を評価するため、データの経験累積分布関数と推定された理論の累積理論分布のとの差の最大値である Kolmogorov–Smirnov（KS）距離を用いる。

**Heaps の法則:** 語彙の一意性という観点から自然言語の性質を捉えるために、Heaps の法則に基づく分析も用いられる [28]。この法則は、文書中の語数 $m$ と語彙サイズ $V(m)$ の間に成り立つ劣線形関係を以下のように表す：

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

ここで、$K > 0$ は語彙成長の初期速度を示すスケーリング係数であり、$\beta$ は文書が拡張されるにつれて新しい語彙が導入される割合である。$K$ が大きい場合、初期段階で多くの固有語彙が導入されることを意味し、一方で $K$ が小さい場合は、語彙拡張が比較的保守的であることを示す。$\beta$ が 1 に近い場合、語彙サイズは文書長にほぼ線形に増加し、語彙的に豊かな言語を示唆する。一方、$\beta$ が小さい場合、語彙成長は劣線形となり、既存語彙の再利用が多く、新語の導入頻度が低い言語特性を反映する。

本研究では、語彙に該当するトークン gram 数 $m$ に対する異なり gram 数 $V(m)$ を算出して、その関係に対して対数空間上での最小二乗法から $\beta$ および $K$ を推定することで近似的に値を取得する。

### 2.3 音声トークンの統計言語分析

これまでに、離散化された音声トークンに対して、本研究と同様の統計的言語分析を試みた先行研究がいくつか存在する。Takamichi ら [29] は、SSL モデルによって生成された音声トークンが、自然言語テキストと同様に Zipf の法則に従うか否かを

(注1)：https://www.cs.cmu.edu/~cburch/words/top.html

表 1: NAC コーデック構成の比較。$n_d$ はトークン次元数，SR はサンプリング周波数を表す。

| Codec | Configuration | Training Data | $n_d$ | SR / kbps |
|---|---|---|---|---|
| SpeechTokenizer [4] | 16k | LibriSpeech [5] | 8 | 16k / 4 |
| AcademiCodec [6] | hifi_24k_320d | LibriTTS [7], VCTK [8], AISHELL [9] | 4 | 24k / 3 |
| AudioDec [10] | 24k_320d | Valentini [11] | 8 | 24k / 6.4 |
| DAC [12] | 24k | Common Voice [13], DNSC [14],Jamendo [15]AudioSet, FSD50K [16] | 32 | 24k / 24 |
| EnCodec [17] | 24k_24bps | Common Voice, DNSC, Jamendo, AudioSet, FSD50K | 32 | 24k / 24 |
| FunCodec [18] | en_libritts_16k_nq32ds320 | LibriTTS のサブセット | 32 | 16k / 16 |

表 2: DEMAND ノイズタイプの分類

| カテゴリー | DEMAND ノイズタイプ |
|---|---|
| L-BAB | DKITCHIN, DWASHING, NFIELD, NRIVER, OHALLWAY, OOFFICE, TCAR |
| M-BAB | DLIVING, NPARK, PSTATION, SPSQUARE, STRAFFIC |
| H-BAB | OMEETING, PCAFETER, PRESTO, TBUS, TMETRO |

分析した。その結果、音声トークンがべき乗則的な挙動を示すことが明らかとなり、音声トークンが一定の言語的構造特性を有している可能性が示唆された。

また、Sicherman ら [30] は、SSL ベースの音声トークンにおける解釈性および冗長性に着目し、トークンと音素との間に強い相関が存在することを示した。さらに、言語モデルの性能低下を招く冗長性を同定する手法を提案し、それらを低減することで、音声合成や生成型音声言語モデリングといった下流タスクの性能が向上することを報告している。

一方、NAC ベースのトークンに関しては、Liu ら [31] が、トークン化過程のばらつきにより、同一の音声入力から異なるトークン系列が生成され得るという「不整合性」の問題を指摘した。さらに、Park ら [3] は、NAC によって得られた離散音声トークン列に対して Zipf 則および Heaps 則に基づく統計的分析を行い、これらの統計指標が音声品質指標や認識性能と相関する可能性を示した。特に、NAC のトークンについては、3-gram でクラスター化するのが最も言語統計的指標が優れていることを示して、トークン分布の逸脱量を定量指標として定義した。

### 2.4 劣化音声におけるトークン安定性と生成性能

また、劣化音声条件下における離散音声トークンの頑健性と、言語情報の維持に関する研究も近年報告されている。Lu らは、ノイズを含む音声入力に対して、NAC 由来の離散トークンをデノイジングすることで、劣化音声条件下でも高品質な音声生成が可能であることを示した [32]。また、Song らは、セマンティック音声トークナイザが雑音に対して不安定であることを指摘し、トークン系列の安定性を向上させることで、下流の音声タスク性能が改善されることを報告している [33]。これらの研究は、劣化音声条件下におけるトークン表現の安定性が、言語・音響情報の保持および生成性能に密接に関係することを示唆している。

## 3. 実験設定

以上の関連研究を踏まえつつも、NAC トークンが雑音付加や音声劣化といった条件下において、言語統計的構造をどの程度安定に維持できるかについては、依然として体系的な検証が不足している。そこで本研究では、複数の NAC モデルから得られる NAC トークンが、劣化音声条件下においてどのように統計的性質を変化させるかを検証する。

表 3: 劣化音声を NAC により再合成した際のクリーン自然音声に対する値の変化率の平均（%）。クリーン自然音声よりも良い結果を示した値を太字処理する。

| Codec | WER↓ | UTMOS↑ | MCD↓ |
|---|---|---|---|
| AcademiCodec | 79.06 | −11.91 | **−0.75** |
| AudioDec | 138.54 | −19.01 | 5.48 |
| DAC | **−32.36** | −0.53 | 1.01 |
| EnCodec | **−26.08** | −5.40 | 3.72 |
| FunCodec | 36.38 | −6.04 | 0.68 |
| SpeechTokenizer | 267.03 | −7.43 | 0.87 |

### 3.1 NAC とデータセットの設定

対象とする NAC モデルとして、学習データセット、ビットレート、および次元構成の異なる公開されている 6 種類のオープンソース音声コーデックを採用して比較対象とした。各モデル構成の詳細は Table 1 に示す [34]。

評価には、約 10 時間の LJSpeech(注2) 単一話者英語音声コーパスを用いた。なお、クリーン音声に対しノイズを付与するため、ホワイトノイズおよび複数環境で収録された雑音を含む DEMAND データセット [35] を用いた。この時、DEMAND からの雑音タイプを雑音中に含まれる発話量に基づいて L-BAB、M-BAB、H-BAB の 3 カテゴリに分類した。L-BAB は、発話をほとんど含まない、もしくは全く含まない雑音、M-BAB は、信号長の半分未満に発話成分を含む雑音、H-BAB は、それ以外もラジオ、テレビなどの背景音声を多く含む雑音群である。各カテゴリに含まれる具体的な雑音タイプを表 2 に示す。

以上の DEMAND ノイズ条件およびホワイトノイズを用い、各 SNR 条件 {−10 dB, −5 dB, 0 dB, 5 dB, 10 dB} 下でクリーン音声と加算することで、背景雑音量の異なる劣化音声コーパスを構築した。

### 3.2 NAC トークンと再合成音声の設定

クリーン音声及び構築した劣化音声コーパスを各 NAC モデルに入力し、Codec-SUPERB [36] の実装を用いて NAC トークン系列と、そのトークンを音声波形でデコードし、研究対象である各モデルに対して再合成された音声出力をそれぞれ生成した。

(注2)：https://keithito.com/LJ-Speech-Dataset/

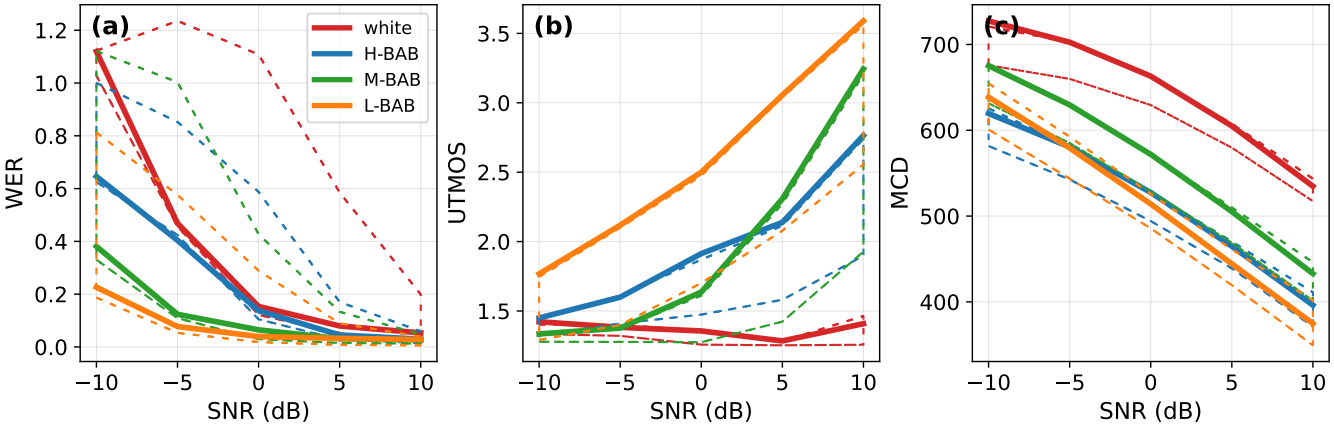


図 1: 各ノイズ条件および SNR における劣化音声結果の WER、UTMOS、MCD の推移。太線は雑音により劣化した自然音声（NAC による再合成なし）であり、上下の点線は、再合成条件下におけるコーデック間の最大値と最小値の範囲を示す。

トークン処理に関して、すべてのモデルにおいて、単一次元あたりのコードブックサイズは 1,024 に統一され、各トークンは 20 ms の音声区間に対応する。ここで、統計分析、特に $n$-gram 分布に対する影響を抑制するため、 連続する同一トークンの繰り返しを除去する重複削減（deduplication）処理を適用した。さらに、NAC モデル間で異なる次元数 $n_d$ を持つことによる影響を考慮し、 各次元を独立したトークン系列として扱った上で、 時間軸方向に沿って連結するフラット化処理を行った。この際、各次元が同一のラベル空間（0〜1,023）を共有することによる ラベル衝突を防ぐため、次元番号に応じたオフセットを各トークン ID に付与した。加えて、各次元系列の先頭および末尾には、 それぞれ「dimension start」および「dimension end」を表す特別トークンを挿入し、 人工的なトークン混在が統計分布に与える影響を低減した。

## 4. 実 験 結 果

### 4.1 劣化音声に対する NAC の再合成性能

まず、劣化音声に対する NAC の再合成性能を調べるため、再合成された劣化音声に対して音声認識モデル（whisper-large-v3 [37]）を使用して文字起こしを出力した後、クリーン自然音声の文字起こしに対する単語エラー率（Word Error Rate; WER）・疑似自然性 MOS 値（UTMOS [38]）・音響歪み（Mel-Cepstral Distortion; MCD）の指標を用いて分析する。

#### 4.1.1 ノイズ条件別の全体的傾向

まず、ノイズ条件による変動傾向を把握する。図 1 に、各ノイズ条件および SNR における WER、UTMOS、MCD の変化を示す。図より、すべての指標に共通して明確な SNR 依存性が確認され、SNR が低下するにつれて WER および MCD は増加し、UTMOS は低下する傾向を示す。特に ホワイトノイズ条件は最も困難であり、低 SNR では、ほぼすべてのコーデックにおいて WER が飽和し、言語内容の保持が著しく困難となった。一方、babbling ノイズ条件では、全ての条件でホワイトノイズ条件より一貫して良い指標を示し、発話内容の混じった雑音でもホワイトノイズに比べて音声的・言語的構造が部分的に保持されやすいことが示唆される。

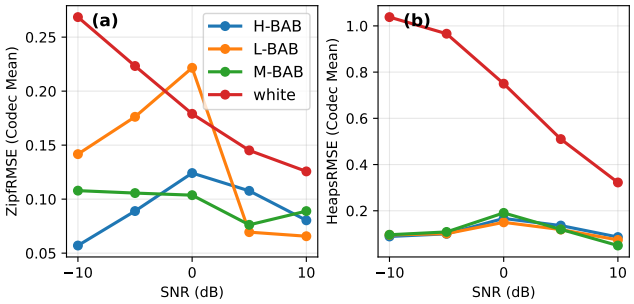


図 2: 各ノイズ条件および SNR において、6 種類の NAC からの ZipfsRMSE/HeapsRMSE 結果の平均値

#### 4.1.2 コーデック別雑音頑健性の特徴

次に、コーデック条件による変動傾向を把握する。まず、図 1 の最大–最小の誤差範囲から、各 NAC から再合成された劣化音声に対する評価値はコーデック間で大きな差を見せ、ほとんどが雑音により劣化した自然音声の結果（図中の太線）より低い性能を示すことを確認できる。

詳しい分析のため、各指標について、ノイズ種類（babbling / ホワイト）および SNR 条件ごとに再合成前の劣化した自然音声コーパスの評価値に対する NAC 再合成音声の評価値の相対変化率（%）を コーデック別に平均化した結果を表 3 に示す。まず、WER の観点では、一部の DAC や Encodec などの高ビットレートコーデックにおいて、劣化した自然音声と比較しても誤り率が低減する場合が観測され、NAC による音声の再合成が結果的に言語情報の正則化として機能する可能性が示唆された。一方、AcademiCodec や SpeechTokenizer などの低ビットレートのコーデックでは、強いノイズ条件下で言語情報の保持が困難となり、WER が大きく悪化する傾向を示した。

次に、UTMOS に関しては、多くの NAC からの再合成音声において劣化した自然音声よりもスコアが低下し、再合成による自然性の損失が一般的であることが確認された。ただし、その低下幅はコーデックによって異なり、一部のコーデックでは

比較的安定した自然性を維持する傾向も観測された。最後に、MCD の観点では、WER や UTMOS の観点で悪い性能を示した NAC が、劣化した自然音声よりも低い歪みを示す場合があり、音響的忠実度と言語的エラー率は必ずしも一致しないことが明確となった。

以上の結果から、NAC による劣化音声の再合成は、ノイズ条件およびコーデック設計に応じて、言語情報の保持、自然性、音響的忠実度の間に異なるトレードオフを生じさせることが示された。このことは、雑音頑健性のある音声生成および音声言語モデリングにおいて、単一の評価指標に依存することの限界を示唆している。

### 4.2 ノイズ条件下における統計言語指標の挙動分析

本研究では、このような傾向が各 NAC のトークンでも見られるかについて調べるためにノイズが付加された音声を複数の NAC により符号化し、得られたトークン系列が Zipf の法則および Heaps の法則に従うかという観点から調査した。

具体的には、$n$-gram（本研究では $n = 3$）単位で Zipf/Heaps の分布パラメータを算出し、クリーン音声条件のパラメータを基準として分布乖離を ZipfRMSE および HeapsRMSE として定量化した。具体的には、劣化音声条件（noise）とクリーン音声条件（clean）との間で、以下のように RMSE を定義する：

$$\mathrm{ZipfRMSE}_{c,t,n} = \sqrt{\frac{1}{2}\left[\left(\alpha^{\mathrm{noise}}_{c,t,n} - \alpha^{\mathrm{clean}}_{c,n}\right)^2 + \left(\mathrm{KS}^{\mathrm{noise}}_{c,t,n} - \mathrm{KS}^{\mathrm{clean}}_{c,n}\right)^2\right]} \tag{3}$$

$$\mathrm{HeapsRMSE}_{c,t,n} = \sqrt{\frac{1}{2}\left[\left(\beta^{\mathrm{noise}}_{c,t,n} - \beta^{\mathrm{clean}}_{c,n}\right)^2 + \left(K^{\mathrm{noise}}_{c,t,n} - K^{\mathrm{clean}}_{c,n}\right)^2\right]} \tag{4}$$

ここで、$c$ はコーデックの種類、$t$ は劣化音声生成条件、$n$ は $n$-gram 次数を表す。

実験結果を図 2 に示す。結果より、ホワイトノイズ条件では ZipfRMSE および HeapsRMSE の双方が SNR の増加に伴って一貫して減少する明確な傾向を示した。特に HeapsRMSE は低 SNR 条件下で顕著に増大し、高 SNR では急激に低下する挙動を示しており、語彙成長構造が ホワイトノイズによって大きく破壊されることが確認された。一方で、Babble noise 条件では、全体として RMSE の絶対値は ホワイトノイズよりも小さいものの、0 dB 付近で最大値を取る非単調な挙動が観測された。これは、音声成分と雑音成分のエネルギーが拮抗する条件において、トークン分布が最も不安定になる可能性を示唆している。

### 4.3 WER/UTMOS/MCD と統計言語指標との関係分析

最後に、上記で分析した ZipfRMSE と HeapsRMSE の値を既存の指標である WER, UTMOS, 及び MCD との相関を分析した。

Pearson 相関係数の結果を表 4 および表 5 に示す。全ノイズ種類・SNR・コーデックを統合した条件において、HeapsRMSE は WER、UTMOS、MCD のすべてと一貫した相関傾向を示した。具体的には、WER および MCD とは正の相関、UTMOS とは負の相関が観測されており、語彙成長構造の崩壊が、認識誤りの増加、知覚的自然性の低下、および音響歪みの増大と同時に進行する傾向を持つことが示唆される。

表 4: 各 NAC コーデックにおける認識・知覚・音響指標（WER / UTMOS / MCD）と ZipfRMSE との Pearson 相関係数。

| Codec | WER↓ | UTMOS↑ | MCD↓ |
|---|---|---|---|
| AcademiCodec | 0.39 | −0.31 | 0.49 |
| AudioDec | 0.39 | −0.38 | 0.54 |
| DAC | −0.17 | 0.13 | 0.03 |
| EnCodec | 0.57 | −0.13 | 0.32 |
| FunCodec | 0.58 | −0.54 | 0.59 |
| SpeechTokenizer | −0.09 | −0.09 | −0.07 |
| **Overall** | 0.17 | −0.13 | 0.27 |

表 5: 各 NAC コーデックにおける認識・知覚・音響指標（WER / UTMOS / MCD）と HeapsRMSE との Pearson 相関係数。

| Codec | WER↓ | UTMOS↑ | MCD↓ |
|---|---|---|---|
| AcademiCodec | 0.48 | −0.37 | 0.57 |
| AudioDec | 0.45 | −0.21 | 0.46 |
| DAC | 0.24 | −0.02 | 0.24 |
| EnCodec | 0.55 | −0.55 | 0.69 |
| FunCodec | 0.67 | −0.61 | 0.68 |
| SpeechTokenizer | 0.11 | −0.11 | 0.23 |
| **Overall** | 0.39 | −0.25 | 0.33 |

これらの相関係数の大きさは、Guilford の経験則において低〜中程度の相関に相当し、特に Encodec、Funcodec のようなコーデックにおいて HeapsRMSE が NAC 再合成音声における複数の劣化側面を横断的に反映する構造的指標として一定の説明力を持つ可能性を示している。すなわち、HeapsRMSE は、個別の性能指標を直接予測するものではないものの、音声全体の劣化傾向を要約的に捉える指標として有効であると考えられる。

一方で、ZipfRMSE は同条件において MCD との間にのみ部分的な相関が観測された一方、WER や UTMOS との間には一貫した対応関係は確認されなかった。この結果は、ZipfRMSE が全体的な性能劣化を包括的に説明する指標というよりも、トークン頻度分布における局所的な歪みや分布形状の変化を反映する指標であることを示唆している。

さらに、WER、UTMOS、MCD における劣化の程度はコーデックごとに異なる傾向を示しており、雑音頑健性が入力 SNR やノイズ種類のみで一意に決定されるものではなく、トークン化方式や codebook 構造といったコーデック固有の設計要因に強く依存することが確認された。

## 5. おわりに

本研究では、雑音付加により劣化した音声を複数の NAC で符号化し、得られたトークン系列の言語統計的構造変化を Zipf/Heaps 則に基づいて分析した。その結果、ホワイトノイズでは両 RMSE が SNR の上昇に伴って単調に低下し、雑音エネルギーがトークン分布構造を連続的に崩壊させることが確認された。一方、babble 系雑音では 0 dB 付近で RMSE が最大となる非単調挙動が観測され、音声成分と雑音成分が拮抗する条件でトークン分布が最も不安定化する可能性が示唆された。さらに、全ノイズ種類・SNR・コーデックを統合した相関分析により、HeapsRMSE は WER・MCD と正、UTMOS と一貫した相

関関係を示し、NAC 再合成音声の「全体的劣化度」を説明する構造指標として有効であることを示した。

今後の課題として、(i) RMSE 定義におけるパラメータ正規化や重み付けの導入による指標の安定化、(ii) 言語・話者・発話スタイルを拡張した汎化検証、(iii) 分布指標を用いたコーデック設計（tokenization / codebook）改善指針の確立、が挙げられる。

**謝辞：**本研究は、JST、Moonshot R&D 助成金番号 JPMJPS2011 の支援と、NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP25006）の結果得られたものです。

## 文　献

[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 ", IEEE Transactions on Pattern Analysis & Machine Intelligence, 01, pp. 1–20 (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, **31**, pp. 2523–2533 (2023).

[3] J. Park, S. Takamichi, D. M. Chan, S. Kando, Y. Saito and H. Saruwatari: "Analysing the language of neural audio codecs", Proc. ASRU (2025).

[4] X. Zhang, D. Zhang, S. Li, Y. Zhou and X. Qiu: "SpeechTokenizer: Unified speech tokenizer for speech large language models", Proc. ICLR (2024).

[5] V. Panayotov, G. Chen, D. Povey and S. Khudanpur: "Librispeech: An ASR corpus based on public domain audio books", Proc. ICASSP, pp. 5206–5210 (2015).

[6] 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).

[7] 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", Proc. INTERSPEECH, pp. 1526–1530 (2019).

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

[9] H. Bu, J. Du, X. Na, B. Wu and H. Zheng: "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline", Proc. O-COCOSDA (2017).

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

[11] C. Valentini-Botinhao: "Noisy speech database for training speech enhancement algorithms and TTS models" (2017). University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR).

[12] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar and K. Kumar: "High-fidelity audio compression with improved RVQGAN", Proc. NeurIPS (2023).

[13] 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", Proc. LREC, pp. 4218–4222 (2020).

[14] 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", Proc. ICASSP (2022).

[15] D. Bogdanov, M. Won, P. Tovstogan, A. Porter and X. Serra: "The MTG-Jamendo dataset for automatic music tagging", Proc. ICML (2019).

[16] 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, **30**, pp. 829–852 (2021).

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

[18] Z. Du, S. Zhang, K. Hu and S. Zheng: "FunCodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec", Proc. ICASSP (2024).

[19] 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, **29**, pp. 3451–3460 (2021).

[20] A. Baevski, Y. Zhou, A. Mohamed and M. Auli: "wav2vec 2.0: A framework for self-supervised learning of speech representations", Proc. NeurIPS (2020).

[21] 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, **16**, 6, pp. 1179–1210 (2022).

[22] 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, **30**, pp. 495–507 (2021).

[23] 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", Proc. INTERSPEECH, pp. 2574–2578 (2024).

[24] J. Q. Yip, S. Zhao, D. Ng, E. S. Chng and B. Ma: "Towards audio codec-based speech separation", Proc. INTERSPEECH, pp. 2190–2194 (2024).

[25] G. K. Zipf: "Human behavior and the principle of least effort.", Addison-Wesley Press (1949).

[26] B. Mandelbrot: "Contribution à la théorie mathématique des jeux de communication", Annales de l'ISUP, Vol. 2, pp. 3–124 (1953).

[27] A. Gelbukh and G. Sidorov: "Zipf and heaps laws' coefficients depend on language", Proc. CICLing, pp. 332–335 (2001).

[28] H. S. Heaps: "Information retrieval: Computational and theoretical aspects", Academic Press, Inc. (1978).

[29] S. Takamichi, H. Maeda, J. Park, D. Saito and H. Saruwatari: "Do learned speech symbols follow Zipf' s law?", Proc. ICASSP, pp. 12526–12530 (2024).

[30] A. Sicherman and Y. Adi: "Analysing discrete self supervised speech representation for spoken language modeling", Proc. ICASSP (2023).

[31] H. Liu, C. Li, Q. Wu and Y. J. Lee: "Visual instruction tuning", Proc. NeurIPS (2024).

[32] Y.-X. Lu, H.-P. Du, F. Liu, Y. Ai and Z.-H. Ling: "Improving noise robustness of llm-based zero-shot tts via discrete acoustic token denoising", Proc. INTERSPEECH (2025).

[33] Anonymous: "Stabletoken: A noise-robust semantic speech tokenizer for resilient speechLLMs", The Fourteenth International Conference on Learning Representations (2026).

[34] 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).

[35] J. Thiemann, N. Ito and E. Vincent: "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings" (2013).

[36] 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", Findings of ACL, pp. 10330–10348 (2024).

[37] 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).

[38] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi and H. Saruwatari: "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022", Proc. INTERSPEECH, pp. 4521–4525 (2022).