音声トークンの言語に関する分析

朴 浚鎔[1,a)]　高道 慎之介[2,1]　David M. Chan[3]　神藤 駿介[1]　齋藤 佑樹[1]　猿渡 洋[1]

概要：近年の音声処理研究は、HuBERT などの自己教師あり学習 (Self-Supervised Learning: SSL) モデルや、EnCodec などのニューラル音声コーデック (Neural Audio Codec: NAC) モデルによって得られる離散トークン表現によって大きく進展している。しかし、これらの表現が言語学的、統計学的な指標を通じて綿密に分析されることは少ない。本研究では、音声から得られるさまざまな離散トークンの統計的および言語的特性を比較分析する。統計的法則に基づく特徴の違いや共通性に加え、エントロピーや冗長性、文法構造、トポロジカル埋め込みの類似性についても明らかにすることを目的とする。さらに、トークン誤り率や音声品質といった指標を用いて、それぞれのトークン化手法が音声データの意味的・音響的情報をどの程度保持しているかを検証する。これらの実験は、個々の音声トークンが持つ統計的性質の理解につながり、音声モデルの設計に洞察を与えることができる。

Joonyong PARK[1,a)]　Shinnosuke TAKAMICHI[2,1]　David M. CHAN[3]　Shunsuke KANDO[1]
Yuki SAITO[1]　Hiroshi SARUWATARI[1]

1. はじめに

近年の音声処理研究は、自己教師あり学習 (Self-Supervised Learning: SSL) モデルや、ニューラル音声コーデック (Neural Audio Codec: NAC) モデルによって得られる離散トークン表現によって大きく進展している [1]。SSL に基づくアプローチは、主に音声認識や音声合成、音声言語理解といったタスクにおいて必要とされる音素的・意味的特徴を捉えることに重点を置いている [2]。一方で、NAC は音響的忠実度や圧縮効率を重視し、波形復元において重要となる詳細な音響情報の保持を目的として設計されている [2]。

しかしながら、これらのトークン化手法が生成する離散表現の言語的・統計的性質について、詳細に分析された研究は未だ少ない。即ち、SSL モデル由来のトークンと NAC のトークンの違いが統計的言語特性に基づいて体系的に比較・検討された事例は限られている。

本研究では、このギャップを埋めるべく、言語統計学の手法を活用した包括的な分析を行う。SSL および NAC によって得られる離散音声トークンの系列を比較することで、それらが持つ統計構造と言語的振る舞いの違いや共通点を明らかにすることを目的とする。本研究によって得られる知見は、音声トークン化の理解を深化させるだけでなく、音声合成や認識における新たな手法の設計指針としても有用である。

[1] 東京大学大学院情報理工学系研究科
[2] 慶応義塾大学理工学部
[3] カリフォルニア大学バークレー校 Berkeley Artificial Intelligence Research Lab (BAIR)
[a)] joonyong-park@g.ecc.u-tokyo.ac.jp

2. 先行研究

2.1 SSL によるトークン

HuBERT [3] や Wav2Vec 2.0 [4] に代表される SSL モデルは、音声信号から明示的な音素ラベルやテキスト書き起こしを必要とせず、離散音声表現を直接生成する手法として音声処理分野に大きな進歩をもたらした。これらの SSL モデル由来のトークンは、意味的・音素的構造の両方を捉える能力があることが示されており、音声認識、音声合成などのタスクにおいて顕著な性能向上を実現している [5]。

2.2 NAC によるトークン

一方、EnCodec [6] や SoundStream [7] などの NAC は、音声信号の効率的な圧縮と忠実な波形再構成を主な目的として設計されている。NAC によって得られるトークンは、極めて詳細な音響情報を符号化しており、精密な音声再現が求められる利用に適している。しかし、その音響的な精度とは裏腹に、NAC 系列の言語的な性質は十分に検討されていない。

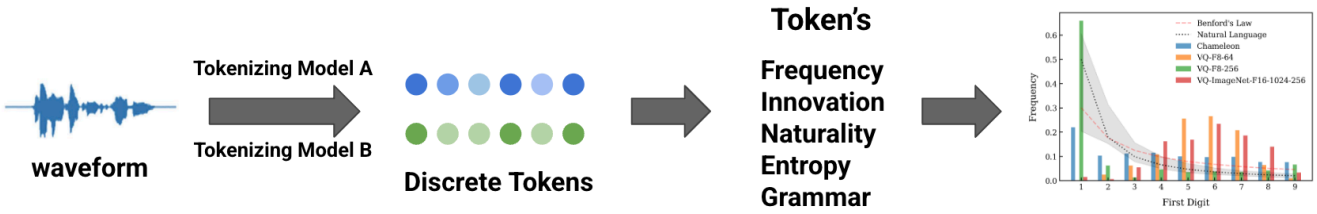

図 1: 本稿で実施する実験の概要

2.3 言語の統計的法則

本節では、本稿で離散音声トークンの分析を行う上で用いるいくつかの法則、モデル、指標について説明する。

Zipf の法則：言語学や情報理論の分野では、自然言語における単語や文字 n-gram の頻度分布が、少数の高頻度要素と多数の低頻度要素からなるという Zipf の法則 [8] に従うことが知られている。この法則は、単語の頻度順位 r とその頻度 $f(r)$ の間に次のべき乗則が成り立つことを示す。

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

ここで、a はスケール定数、η は分布の鋭さやスケール特性を表すパラメータである。η の値は分布の形状を決定し、理想的な Zipf 分布では $\eta \approx 1$ となる。この分布は多くの自然言語や動物のコミュニケーション、さらには大規模言語モデルにおいても確認されており、冗長性と情報効率性のバランスを示す指標として注目されている [9], [10]。

Heaps の法則：語彙拡張の動的特性を捉えるために、Heaps の法則 [11] に基づく分析も行われている。これは、入力テキストサイズ m に対して語彙サイズ $V(m)$ がサブリニアに増加する関係を示し、次式で表される。

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

自然言語では、既存の語彙を繰り返し利用しながら、漸進的に新しい語彙を追加する傾向があるため、$\beta < 1$ のサブリニア成長が観察される。

Yule-Simon 過程：Heaps の法則における語彙拡張特性を確率論的に説明するモデルとして、Yule-Simon 過程 [12] が知られている。Yule-Simon 過程では、新たなトークンが追加される確率 ρ と、既存トークンが再利用される確率 $1-\rho$ に基づき、k 番目のトークン w_k が追加される確率を次式で表す。

$$P(w_k = \text{new}) = \frac{\rho}{k + \rho - 1} \tag{3}$$

$$P(w_k = \text{existing } w) = \frac{\text{count}(w)}{k + \rho - 1} \tag{4}$$

ここで、ρ は新規トークン導入率、$\text{count}(w)$ は既存トークン w のこれまでの出現回数である。この確率過程に従うことで、語彙サイズ $V(k)$ は次のようなべき乗成長を示すことが知られている。

$$V(k) \propto k^{\frac{\rho}{1+\rho}}, \tag{5}$$

ここで、k は系列の長さ（トークン数）である。このように、データが増加するにつれて新規語彙の出現確率が減少し、既存語彙の再利用が優先されるという、自然言語の語彙拡張に見られる特性を説明するモデルとなっている。

エントロピーと冗長性：Shannon [13] によって提案されたエントロピーは系列内の情報多様性や予測困難性を示す指標として広く用いられており、言語や符号化効率の評価に活用されてきた。エントロピー H は次式で定義される。

$$H = -\sum_{i=1}^{V} p_i \log_2 p_i, \tag{6}$$

ここで、V は語彙サイズ、p_i は i 番目のトークンの出現確率である。さらに、エントロピーに基づくハフマン符号化を用いることで、系列のビット削減率や冗長性を定量化できる。ハフマン符号化による平均符号長 L は、トークン分布のエントロピーに対して次のように評価される。

$$H \leq L < H + 1 \tag{7}$$

これにより、エントロピー H に対する実際の符号長 L の比率を取ることで、圧縮効率（冗長性）R を次式で定量化できる。

$$R = \frac{L - H}{L} \tag{8}$$

ここで、R が小さいほど効率的な符号化が達成されていることを意味し、トークン列が持つ冗長性の少なさや情報圧縮性の高さを示す指標となる。

2.4 音声トークンの統計的言語分析

Takamichi らは、SSL により生成された音声トークンが自然言語のテキストで観測される Zipf の法則に従うかを分析し、これらのトークンが冪則に従うことを報告した [14]。これは、音声トークンが言語的構造をある程度持つことを示唆している。また、Sicherman らは、SSL ベースの音声トークンにおける冗長性と音素解釈性に着目し、トークン集合に内在する冗長性の削減が下流タスクの性能向上に寄与することを指摘している [15]。NAC に関しても、Liu らは、同一音声入力に対して異なるトークン系列が生成されるという問題を指摘し、トークンの一貫性と信頼性を向上させる手法を提案している [16]。

3. 実験設定

本研究では、音声トークンが自然言語と類似する統計的特性を持つかを検証するため、HuBERT および NAC（EnCodec [6]、HiFi-Codec [17]）から得られる音声トークンを対象に分析を行った。

HuBERT については、LibriSpeech [18] で事前学習された hubert-base[*1]モデルを使用し、言語的な表現として最適と指摘されている第 9 層 [19] の隠れ特徴量を抽出してトークン化を行った。さらに、LibriSpeech コーパスは、隠れ特徴量からトークン化のためのクラスタ ID を決定するための k-means クラスタリングモデルの学習にも用いた。

一方、NAC である EnCodec および HiFi-Codec については、AcademiCodec[*2]の実装を用い、LibriTTS [20]、VCTK [21]、AISHELL [22] といった複数のオープン音声コーパスで事前学習されたモデルを使用した。トークン化設定としては、bandwidth を 12 に固定し、コードブックサイズを 1024 となるように調整した。

いずれのモデルにおいても、取得されたトークン列は同一クラスタサイズを用い、固定長チャンク単位に分割して抽出した。さらに、連続する同一トークンの繰り返しが分析結果に影響を与えないよう、すべてのモデルで重複トークンの除去を実施した。

分析対象とする音声データは、英語 (LJSpeech)[*3]、日本語 (JSUT) [23]、韓国語 (CSS10)、中国語 (CSS10) [24] の 4 言語で、それぞれ 10 時間の単一話者音声データを 16 kHz に統一してサンプリングし使用した。これらのデータセットにはすべて手動書き起こしが付与されており、形態素解析などを通じて正規化された単語列を Ground Truth (GT) として用いた。GT との比較は、音声トークン系列と条件を揃えるため、単語単位ではなく文字単位の n-gram 系列にして分析を行う。

4. 実験結果

取得したトークン系列に対して、$n = \{1, 2, 3, 6, 9\}$ の n-gram 系列を生成し、統計的・構造的な分析を行った。具体的には、Zipf 則 [8] や Heaps 則 [11]、Yule-Simon モデル [12] 適合性、エントロピー [13] や冗長性の計算を通じて、各モデル・各言語におけるトークン系列の性質を評価した。

4.1 トークン頻度分布と Zipf 則の適合性

本実験では HuBERT、EnCodec、HiFi-Codec から抽出したトークンに対して、n-gram ごとの頻度分布を算出し、log-log 空間における順位頻度プロットを通じて Zipf 則 [8] への適合性を検証した。

音声トークンとの比較を正確に行うため、GT である自然言語テキストについては、単語単位ではなく文字単位の n-gram 系列を生成して評価を実施した。その結果、英語では 6-gram、韓国語では 3-gram、日本語・中国語では 2-gram の文字 n-gram において最も顕著な線形性が確認された。したがって、本研究ではこれらを自然言語側の参照分布として設定し、音声トークン系列との比較分析を行った。

自然言語および各トークンモデルの n-gram 別 η 値を計算した結果を図 2 に示す。この図における η の値は、図 **??** に示す log-log スケール上での n-gram 順位頻度分布に基づいて計算されたものである。図 **??**の黒線に見られるように、自然言語では log-log スケール上で直線的な順位頻度分布が観察された。即ち、自然言語の単語や文字 n-gram は、従来の研究 [14], [25] と同様に Zipf 則や冪乗則に従う典型的な分布を示したことが確認された。一方、図 **??**の青線に見られるように、HuBERT トークンは n が小さい場合にはこの傾向に比較的近い分布を示しているが、n が大きくなるにつれて η が小さくなる（即ち、トークン利用のばらつきが小さくなる）傾向が確認できる。このことから、HuBERT は音響的特徴を捉えつつも、より高次の n-gram においては語彙の圧縮や再利用性が限定的であることが示唆される。

図 **??**の緑線および紫線に示すように、EnCodec や HiFi-Codec といった NAC は、全体として HuBERT よりも Zipf 則から大きく逸脱する分布を示している。即ち、これらのモデルの分布が自然言語のような冪乗則とは異なる傾向を持つことを示している。特に HiFi-Codec は、$n > 2$ においてほとんど η がゼロ付近の値となっており、トークン出現頻度がほぼ均一であることが読み取れる。これは、これらのモデルが音響再現性やビット削減率を優先し、情報の再利用や言語的圧縮構造を重視しない設計となっていることを反映している。

4.2 トークン革新性

続いて、音声トークンの革新性を評価するために Heaps の法則 [11] および Yule-Simon モデル [12] に基づく分析を行った。

Heaps の法則に基づく分析では、各コーパスに対して、自然言語の文字 n-gram 系列および音声トークン系列を対象に、トークン列を先頭から順番に読み取りながら、累積トークン数に対するユニークトークン数の増加曲線を算出し比較を行った。自然言語については、Zipf 則の分析と同様に、単語単位ではなく文字 n-gram 系列を用いており、上記の分析から日本語・中国語では 2-gram、韓国語では 3-gram、英語では 6-gram の文字 n-gram をそれぞれ自然言語の基準系列として設定した。

[*1] github.com/facebookresearch/fairseq/examples/hubert
[*2] github.com/yangdongchao/AcademiCodec
[*3] keithito.com/LJ-Speech-Dataset/

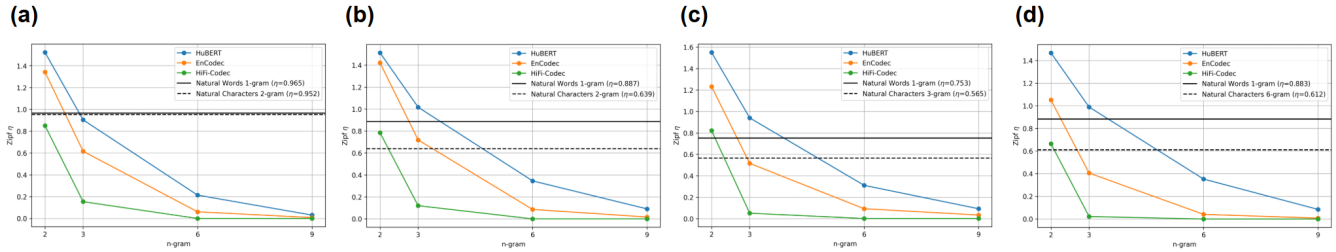

図 2: 音声トークンにおける Zipf 則への適合性比較。各言語（a: 日本語, b: 中国語, c: 韓国語, d: 英語）について、自然言語および各トークンモデルの n-gram 別 η 値を比較している。

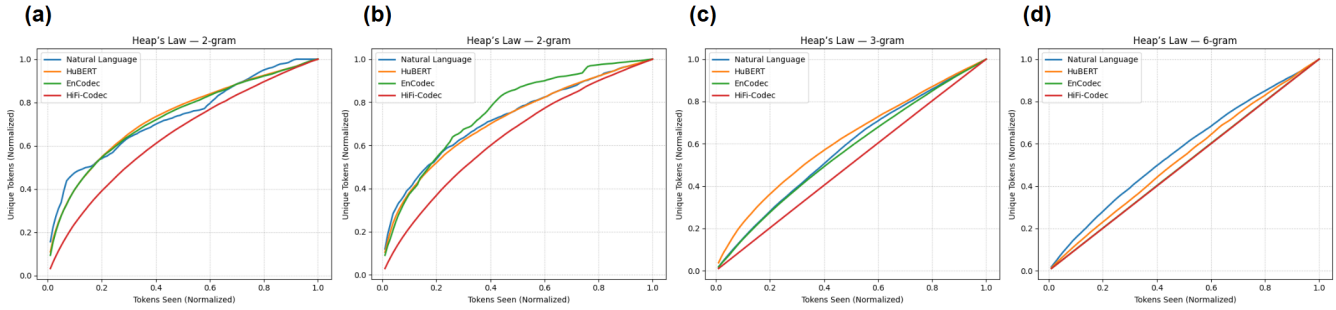

図 3: 音声トークンにおける Heaps 則への適合性比較。各言語（a: 日本語, b: 中国語, c: 韓国語, d: 英語）について、自然言語および各トークンモデル（HuBERT, EnCodec, HiFi-Codec）の n-gram 別ユニークトークン成長曲線を示す。

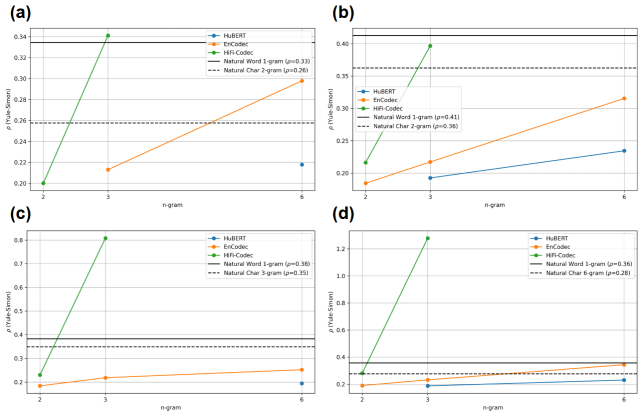

図 4: 各言語（a: 日本語, b: 中国語, c: 韓国語, d: 英語）におけるモデルトークン系列の n-gram 分布と Yule-Simon モデルフィットの関係を示す。グラフに表示されていないデータは確率質量関数の収束失敗を意味する。

図 3 に Heap 則への適合性に関する分析結果を示す。自然言語は期待どおり、入力テキストの総単語量に対して語彙サイズがサブリニア（$\beta < 1$）に増加する典型的な Heaps の法則に従う挙動を示した。一方、HuBERT および NAC 系列のトークンもすべて、入力テキストの総単語量の増加に伴い語彙サイズが増加する傾向を示したが、その成長速度にはモデル間で違いが見られた。特に HuBERT は、低次 n-gram では急速に語彙を拡張するものの、n が大きくなるにつれて語彙成長が緩やかになり、NAC 系列よりも遅い速度でリニア（$\beta \approx 1$）に近づく傾向を示した。これに対し、NAC では、より早い段階から語彙サイズがほぼリニアに増加し、入力ごとに新規トークンが高確率で導入される特徴が観察された。これは、NAC 系列が情報の再利用よりも新規情報の導入を優先し、音響的表現の多様性を重視する設計であることを示唆している。

さらに、Yule-Simon モデル [12] との適合性についても評価を行った。Yule-Simon モデル [12] に基づく適合性評価では、各言語コーパスおよび音声トークン系列に対して、Zipf 則の分析と同様に n-gram ごとの出現頻度分布を算出した。具体的には、各 n-gram 系列から得られたトークン頻度データから、出現頻度の順位分布に対して Yule-Simon 確率質量関数に基づく曲線フィッティングを適用し、最適な ρ 値を算出した。

図 4 に結果を表し、黒線で自然言語の基準線を示した。これに対し、HuBERT は低次 n-gram においてこのモデルと乖離する傾向が見られたが、高次 n-gram では適合性が改善され、一定の言語的性質を維持していることが確認された。一方、EnCodec は HuBERT より高い ρ 値を示し多くのケースで乖離なく適当な結果を示した。HiFi-Codec の値は 3-gram を超えるにつれて急激に増加し、高次 n-gram でモデルと乖離する特性を示した。

この結果から、音声トークン系列の Yule-Simon 適合性評価から、HuBERT は n が大きくなるほど自然言語に近づく一方、NAC 系列（特に HiFi-Codec）は n が小さくなるにつれ ρ が、自然言語に近づくことが明らかとなった。これは、HuBERT が言語的な再利用性を一定程度保持するのに対し、NAC 系列は音響的多様性を優先している設計傾向を示唆していると言える。

4.3 トークンエントロピーと冗長性の分析

最後に、各モデルから得られるトークン系列のエントロ

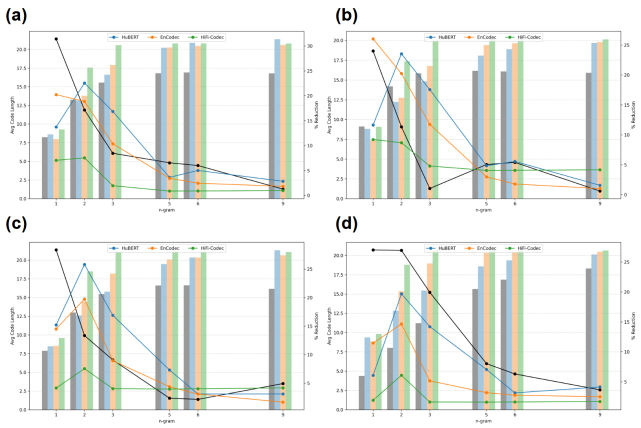

図 5: (a) 日本語、(b) 中国語、(c) 韓国語、(d) 英語における音声トークン系列の平均符号長（Bar）とビット削減率（折れ線）の比較結果。黒色は文字単位自然言語の参照値を表している。

ピーおよび圧縮性について検証を行った。具体的には、各言語コーパスおよび音声トークン系列に対して計算したエントロピーをもとに、各 n-gram 系列に対してハフマン符号化を適用し、各トークンの出現頻度に基づいた最適な可変長ビット列を生成し、その平均符号長を算出した。さらに、固定長符号化との比較により、ビット削減率を求めた。

図 5 に結果を示す。自然言語は、低次 n-gram において高いビット削減率を示し、語彙の再利用やパターンの繰り返しによる効率的な表現が可能であることが確認された。しかし、n が大きくなるにつれて圧縮効果は徐々に減少し、高次 n-gram ではほとんど圧縮が効かなくなる傾向が見られた。これは、より長い n-gram になるほど語彙の多様性が高まり、同一パターンの繰り返しが減少するためと考えられる。

HuBERT においても、自然言語と同様にある程度の圧縮性が観察された。特に低次 n-gram では比較的良好な圧縮効果を示し、一定の冗長性を保持していることが分かった。しかし、n が大きくなるにつれてビット削減率は低下し、情報がより均一に分散していく傾向が確認された。これは、HuBERT が音響情報を符号化しつつ、言語的な再利用性を部分的に維持している設計であることを示唆している。

一方、NAC 系列である EnCodec は HuBERT よりも低い圧縮性をみせ、特に HiFi-Codec は、全体的に極めて低い圧縮効果しか得られなかった。HiFi-Codec のトークンは、n の大きさに関わらずほとんど冗長性を持たず、トークン系列がほぼ圧縮不可能な状態であることが確認された。この結果は、波形復元や音質保持を優先する設計方針に起因すると考えられ、言語的な情報構造よりも音響的忠実度を重視していることを反映している。

さらに注目すべき点として、図 5 の線グラフより自然言語とは対照的に音声トークンでは、1-gram から 2-gram に移行した際に急激にビット削減率が低下する傾向が観察された。特に、1-gram ではある程度のビット削減が可能であったものの、2-gram では圧縮効果が大幅に減少し、その後も n の増加とともにほとんど圧縮が効かなくなる様子が確認された。

この現象は、n-gram の構造に関連していると考えられる。1-gram では、各トークンが独立して扱われるため、頻出トークンとそれ以外のトークンとの間に出現頻度の偏りが存在し、ハフマン符号化による効率的なビット削減が可能となる。しかし、2-gram 以降では、連続するトークンペアやシーケンスの種類が急激に増加し、個々の n-gram の出現頻度が均一化していく。その結果、出現頻度に基づく符号化効率が低下し、圧縮効果が著しく減少するものと考えられる。

また、符号長の増加とも密接に関連している。n が大きくなるほど、必要なビット長が増加し、固定長符号との差が縮小するため、圧縮による削減効果が薄れていく。特に音声トークン系列では、n-gram 化に伴い平均符号長が大幅に伸びる一方で、頻出 n-gram が限定的であるため、可変長符号による恩恵がほとんど得られない。このことが、2-gram 以降のビット削減率低下の主な要因であると考えられる。

5. おわりに

本研究では、音声信号から得られる離散トークン系列が自然言語のような言語的・統計的特性を持つかどうかを、多面的な指標に基づき検証した。その結果、HuBERT トークンは n-gram の小さい範囲において自然言語と類似した統計的挙動を示し、言語的圧縮性や再利用性を部分的に保持していることが明らかとなった。一方、NAC 系列のトークンは、言語的な構造よりも音響的忠実度を優先する設計思想に基づき、Zipf 則や Heaps 則から逸脱し、高いエントロピーと低いビット削減率を示した。このことは、音声トークナイザの設計や選択では、タスクに応じて「言語的情報処理」か「音響的高忠実度処理」のいずれを優先するかを考慮すべきであることを示唆している。特に、言語理解や発話生成などのタスクでは、言語的統計性を保持するトークン化手法が有効であり、高忠実度な音声復元や高音質合成タスクでは、情報を豊富に保持するトークン化手法が適している可能性がある。

一方、本研究にはいくつかの限界も存在する。まず、分析に用いた音声データセットは単一話者・単一条件に限定されており、多話者・雑音環境・長時間音声など、より現実的な条件での一般化は今後の課題である。特に、現在の実験は主に英語だけで学習されているオープンソースモデルに依存しており、これに対する言語依存性問題も存在する。また、本研究ではトークン系列の統計分析に留まり、実際の音声認識や音声合成といった下流タスクへの影響評

価までは踏み込んでいない。

今後の研究では、より多様な音声データやトークナイザ設定を対象とした検証を進めるとともに、トークナイザ設計自体の最適化や、下流タスクへの実用的な応用評価を行う必要がある。また、本研究で得られた統計的知見を活用し、言語性と音響性を両立させる新たなトークン化手法や学習戦略の開発に取り組むことが、今後の音声処理技術の更なる発展に寄与すると期待される。

謝辞： 本研究は, JST, Moonshot R&D 助成金番号 JPMJPS2011 の支援を受けたものである.

参考文献

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

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

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

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

[12] J. C. Willis and G. U. Yule, "Some statistics of evolution and geographical distribution in plants and animals, and their significance," *Nature*, vol. 109, no. 2728, pp. 177–179, 1922.

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

[14] 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, 2024, pp. 12 526–12 530.

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

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

[17] 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*, vol. abs/2305.02765, 2023.

[18] 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.

[19] V. Vielzeuf, "Investigating the 'autoencoder behavior' in speech self-supervised models: a focus on hubert's pretraining," *arXiv*, vol. abs/2405.08402, 2024.

[20] 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.

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

[22] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus," in *INTERSPEECH*, Brno, Czech Republic, 2021, pp. 2756–2760.

[23] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

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

[25] S. Yu, C. Xu, and H. Liu, "Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation," *arXiv preprint arXiv:1807.01855*, 2018.