深層学習で獲得される音声シンボルは自然言語シンボルと同様に Zipf 則に従うか？*

前田 紘希，○高道 慎之介，朴 浚鎔，猿渡 洋（東京大学）

1 はじめに

Zipf 則（ジップ則, Zipf's law）とは，データ集合に含まれる要素の出現頻度に関する経験則である．ある要素の出現頻度が当該データ集合の中で k 番目に大きいとき，その頻度は 1 番目に大きい要素の頻度の $1/k$ であるという法則である [1]．この法則は様々なデータ集合において成り立つことが知られており，自然言語シンボルも同様にこの法則に従う．文章内の自然言語シンボル（単語あるいは文字 n-gram）の出現頻度を集計し，シンボルの頻度順位を r 位，その出現頻度を f_r とすると，

$$f_r = ar^{-\eta} \tag{1}$$

が成り立つ．ここで，a, η はモデルパラメータである．$\eta \approx 1$ のとき当該文章は Zipf 則に，それ以外の時に冪乗則に従うと定義される．すなわち，Zipf 則は冪乗則の一種である．冪乗則に従うとき，順位–頻度の両対数グラフは直線状となる．

Zipf 則に従う事実あるいは Zipf 則からの乖離を見出すことで，当該データ集合の特性を解析できる．自然言語の場合には，以下の例が知られている [2]．

子供の発達：2 歳から 4 歳の発話した単語について頻度を計算するとグラフは直線にならず上に凸となる．すなわち子供の発話は頻度の大きい単語に偏る [3]．

言語間差異：異なる文字表記システムを持つ言語の間では異なる頻度分布になる．

他方，音声の新たな表現として，深層学習で獲得される音声シンボルがある．この表現は，従来の特徴量表現（例えば，スペクトル包絡）より縮約された表現であり [4,5]．音価や言語的意味の表現と見做すこともできる [4–6]．そのため，「音声に含まれる音韻や言語的意味を人間の定義に基づいて表したもの」が自然言語シンボルとすれば，「音声に含まれる音韻や言語的意味をデータ駆動で表したもの」が音声シンボルであると対比できる．以上の整理を踏まえると，**「自然言語シンボルにおいて成り立つ Zipf 則は，音声シンボルにおいても成り立つか」**の疑問が生じる．これを検証することで，前述した自然言語における解析を音声において実施できる可能性と，言語音以外の音声や音声以外の音についても同様の解析方法を展開できる可能性を見出す．

そこで本研究では，音声シンボル獲得の一種である generative spoken language model (GSLM) [4] を用いて検証を行う．文と音声の対データから成る音声コーパス，形態素解析，GSLM を用いて，文と音声それぞれについてシンボル列を獲得する．音声シンボル単体における検証と，自然言語シンボルとの比較により，傾向を実験的に調査する．

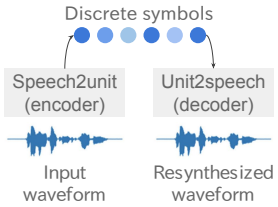

Fig. 1 Generative spoken language model.

2 GSLM による音声シンボル獲得

Figure 1 に示す GSLM [4] を概説する．GSLM は speech2unit, unit language model, unit2speech から成る，音声シンボルを介した分析再合成系である．このうち本稿では speech2unit モジュールのみを用いる．HuBERT [7] などの自己教師あり学習 (self-supervised learning: SSL) モデルエンコーダによる特徴表現と，k-means clustering によって，音声波形を音声シンボル系列に変換する．

GSLM は音素継続長より短い時間間隔で分析するため，同一のシンボルをしばしば連続して推定する．その冗長な表現を避けるため，本研究では連続する同一シンボル列を単一のシンボルに置き換える．

speech2unit は，言語に強く依存するため，対象音声と同じ言語の音声で学習されるべきである．現時点で，英語と日本語についてのモジュールが公開されているため [8]，本研究ではこの 2 言語を対象とする．

3 自然言語シンボルと音声シンボルを対照させた Zipf 則の検証

自然言語シンボルと対照して音声シンボルの Zipf 則を検証する方法論を述べる．この実験は，文と音声の対からなるコーパスを利用する．

文については自然言語シンボルを用いる．自然言語シンボルは単語あるいは文字 n-gram とする．単語には形態素解析により求める単語原型，文字 n-gram には連続する n 個の文字列を用いる．音声については音声シンボルを用いる．Section 2 の方法を用いて音声シンボルを獲得し，音声シンボル n-gram，すなわち連続する n 個の文字シンボル列を扱う．自然言語シンボル列と音声シンボル列の長さの比（直感的に記述すれば，1 個の文字シンボルが何個の音声シンボルに平均的に対応するか）を求め，検証に利用する．

4 実験的評価

4.1 実験条件

学習済み GSLM および音声分析条件には文献 [8] と同じものを用いた．音声シンボルの異なり数は 200

*Do learned speech symbols follow the Zipf's law as well as natural language symbols?, by Hiroki Meada, Shinnosuke Takamichi, Joonyong Park, and Hiroshi Saruwatari (UTokyo).

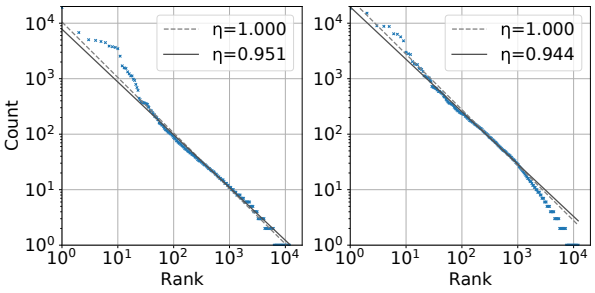

Fig. 2　単語の頻度．左：日本語，右：英語．

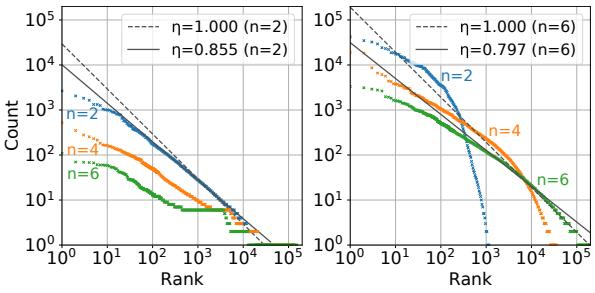

Fig. 3　文字 n-gram の頻度．左：日本語，右：英語．

とした．形態素解析には MeCab[1](ja) と NLTK[2](en) を用いた．コーパスには JSUT の約 7,600 文 (ja) と LJSpeech の約 13,000 文 (en) を用いた．Zipf 則は粗くしか成立せず，特に高頻度要素と低頻度要素は両対数グラフ上の直線からしばしば乖離する．そこで本稿では，出現頻度の上位 0.1%から 10%までのみを直線のモデルパラメータ推定に用いた．このパラメータは，両対数における最小二乗法により推定した．描写時には，画像ファイルサイズを抑えるため，低頻度要素について間引きした．

1 単語あたりの平均文字数は 1.6 (ja)，5.1 (en)，1 文字あたりの平均音声シンボル数は 5.7 (ja)，1.9 (en)，1 単語当たりの平均音声シンボル数は 8.9 (ja)，9.0 (en) であった．n-gram の n として，これらの数字を繰り上げた値を用いる．

4.2　実験結果

4.2.1　単語の場合

Figure 2 に単語頻度と η の値を示す．頻度分布はおよそ直線状であり，η の値は日英ともに 1.0 に近い．そのため，本実験で用いるコーパスの単語は Zipf 則に従うと見做して良い．

4.2.2　文字 n-gram の場合

Figure 3 に文字 n-gram 頻度と η の値を示す．回帰直線は，1 単語あたりの平均文字数に近い n について計算している．日本語に関して，$n = 2$ においておよそ直線状と見做せるが η の値は 1.0 から乖離している．故に，日本語の文字 2-gram は冪乗則に従うと見做される．一方，英語における同条件 ($n = 6$) では，直線上ではなく上に凸の形状となっている．また，他の n についても同様の形状となっている．本稿は音声シンボルに関する研究であるため，文字 n-gram に関する議論は以上に留める．

[1] https://taku910.github.io/mecab/
[2] https://www.nltk.org/

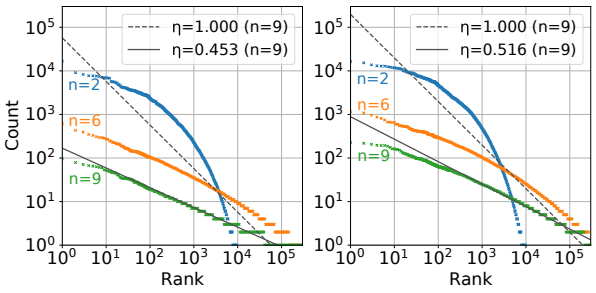

Fig. 4　音声シンボル n-gram の頻度．左：日本語，右：英語．

4.2.3　音声シンボル n-gram の場合

Figure 4 に音声シンボル n-gram 頻度と η の値を示す．回帰直線は，1 単語あたりの平均音声シンボル数に近い n について計算している．

日本語の場合，$n = 9$ において直線状の分布が認められる．故に，日本語の音声シンボル 9-gram は冪乗則に従うと見做される．そのため，日本語においては，単語が Zipf 則に従うならば，単語に対応する系列長を持つ音声シンボル n-gram が冪乗則に従う仮説が生まれる．

一方，英語の場合はいかなる n においても直線状とは認められず，上に凸の形状となっている．各 n について，日本語と英語を比較すると，全ての n について英語のほうが強く上に凸の形状となっている．このことから，日本語に比べ英語の発話が頻度の大きい単語に偏っている可能性が示唆される．ただし，本実験において用いられるコーパスは，音声合成のために設計された点で共通するものの，文ドメインが異なるため，その影響が含まれることに注意したい．

5　まとめ

本研究では，GSLM で獲得された音声シンボルが自然言語シンボルと同様に Zipf 則に従うかを調査した．

謝辞： 本研究は科研費 22H03639，23H03418，JST 創発的研究支援事業 JP23KJ0828，ムーンショット JPMJPS2011 の助成を受けたものです．

参考文献

[1] G. K. Zipf, “Human behaviour and the principle of least effort,” 1949.

[2] 田中久美子, **言語とフラクタル**． 東京大学出版会, 2021.

[3] L. Elena et al., “Two-year-old children’s production of multiword utterances: A usage-based analysis,” *Cognitive Linguistics*, vol. 20, no. 3, 01 2009. [Online]. Available: https://cir.nii.ac.jp/crid/1364233271162092672

[4] K. Lakhotia et al., “On generative spoken language modeling from raw audio,” *Transactions of the ACL*, vol. 9, pp. 1336–1354, 2021.

[5] Z. Borsos et al., “AudioLM: A language modeling approach to audio generation,” arXiv 2209.03143, 2022.

[6] J. Park et al., “How generative spoken language modeling encodes noisy speech: Investigation from phonetics to syntactics,” in *Proceedings of INTERSPEECH*, 2023.

[7] W.-N. Hsu et al., “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.

[8] 朴. 浚鎔 et al., “Generative spoken language model を用いた劣化雑音音声の分析と他言語への適用,” 2023.