正書法および音韻の複雑さによる音声認識の精度への影響

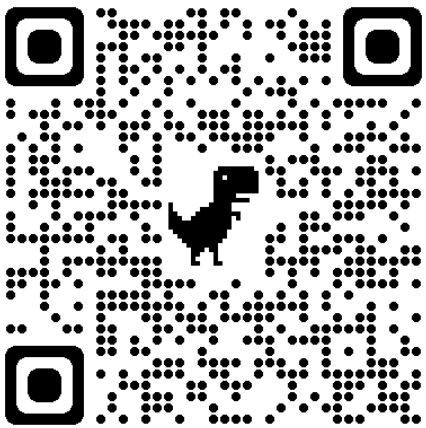
田口 智大
Chihiro Taguchi
ctaguchi@nd.edu
https://ctaguchi.github.io

Department of Computer Science and Engineering,
College of Engineering, University of Notre Dame
IN, USA

2024 年 3 月 12 日
言語処理学会第 30 回年次大会

はじめるまえに

※本発表では、2024 年 1 月の予稿提出後に行なった再実験の結果を含んでいます。最新の実験結果とコードは
https://github.com/ctaguchi/NLP2024ASRcomplexity



はじめに

問い:
言語のどのような複雑性が音声認識の精度に影響を与えるのか？

仮説

音声認識精度に影響を与えそうな言語的要因とは...？
人間の（第一・第二）言語習得に基づいた仮説を提起

1. 書記素の数
   - 文字種が多いと学ぶのが大変...
   - 日本語・中国語の漢字
2. 表語性（logographicity）
   - 発音との対応が不規則な書記体系だと学ぶのが大変...
   - 英語の正書法
3. 音素の数
   - 区別すべき発音の種類が多いと学ぶのが大変...
   - 英語の母音（約 15 個） vs. 日本語の母音（5 個）

音声認識モデルも、このような言語的複雑性に苦戦するのか？

先行研究: 多言語音声認識

Transformer アーキテクチャの発展・普及と共に、（多言語）音声認識の精度もめざましく進展

近年主流の（多言語）音声認識モデル

- Wav2Vec 2.0 (Baevski et al., 2020): 自己教師あり学習モデル
- Whisper (Radford et al., 2022): 弱教師あり学習、Encoder-Decoder モデル

Whisper は事前学習中の訓練データに対象言語が含まれているかどうかで性能の差があるため (Rouditchenko et al., 2023)、本研究では Wav2Vec 2.0 を使用する

先行研究: Wav2Vec 2.0 の事前学習

自己教師あり学習で（音声の）表現を学習をし、ファインチューニングで多様なタスクに対応するという点で、BERT と似ている

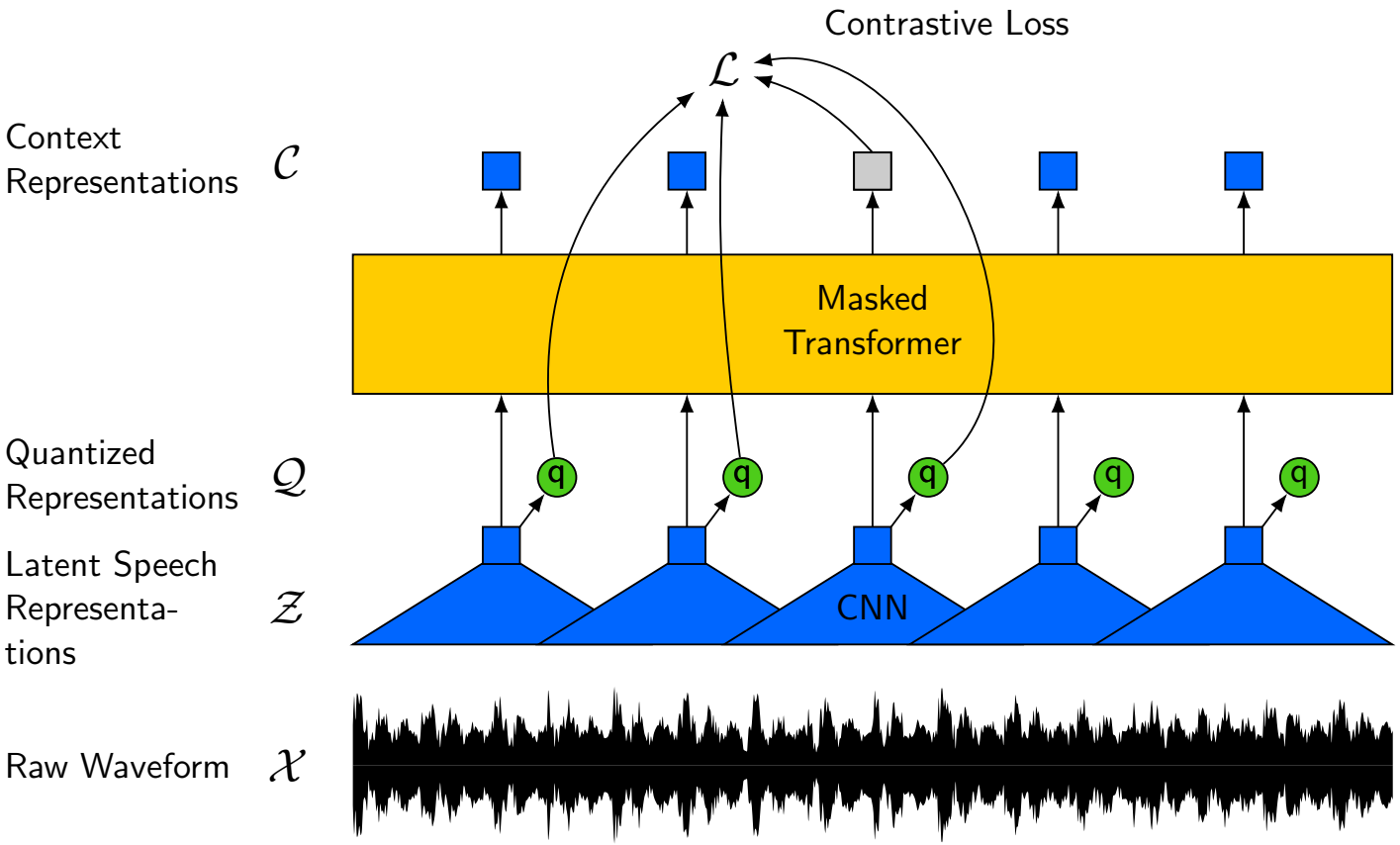


Figure: Wav2Vec 2.0 の事前学習

先行研究: 表語性

表語性（logographicity）: Sproat and Gutkin (2021) が提唱・定式化

- ある言語において、発音が同じである語が複数の異なる表記を持つとき、表語的であるという
  - 通常、異なる表記は文脈に依存する

日本語	/ほうこう/	方向、咆哮、芳香、砲口...
中国語	/shìshí/	事實、適時、嗜食、是時...
英語	/raɪt/	write, right, rite, Wright...

- 表語性の計測: attention 行列の拡散の度合いを計測
  - 音素 → 正書法の翻訳モデルを訓練し、decoder の最後の attention に注目
  - attention の数値が、ターゲットとなる語の外まで広く及んでいる場合、モデルは外の文脈を見ている → 表語性が高い

手法: データ

- 音声認識モデルの訓練データ: 全ての言語について、CommonVoice 16.1 (Ardila et al., 2020) を使用
- 表語性のデータ: Sproat and Gutkin (2021) で報告されている数値を使用
- 音素数のデータ: Phoible 2.0 (Moran and McCloy, 2019) で報告されている数値を使用

手法: ファインチューニング

- 事前学習モデル: Wav2Vec2.0-XLSR-53 (Conneau et al., 2020)
  - 53 言語、5.6 万時間のデータで事前学習
  - 約 3 億パラメータ（約 1.2GB）
- ファインチューニング: コネクショニスト時系列分類法（CTC）
  - 訓練目標: 入力音声 $\boldsymbol{x}$ から出力文 $\hat{\boldsymbol{y}}$ を予測し、ラベル文 $\boldsymbol{y}$ との CTC 誤差を最小化
- 全ての実験において、訓練データは 1 万秒、エポック数は 20、学習率は 0.0003 に統一
- 評価手法は Character Error Rate（CER; 文字誤り率）を使用

実験設計: 仮説１（書記素の数）

- 日本語の三種類の書記法について、同じデータを用いて精度を比較
  - 漢字仮名交じり文
  - カタカナのみ
  - ローマ字のみ
- カタカナ変換には SudachiPy (Takaoka et al., 2018)、ローマ字変換には pykakasi[1] を使用

[1] https://pykakasi.readthedocs.io

実験設計: 仮説２（表語性）

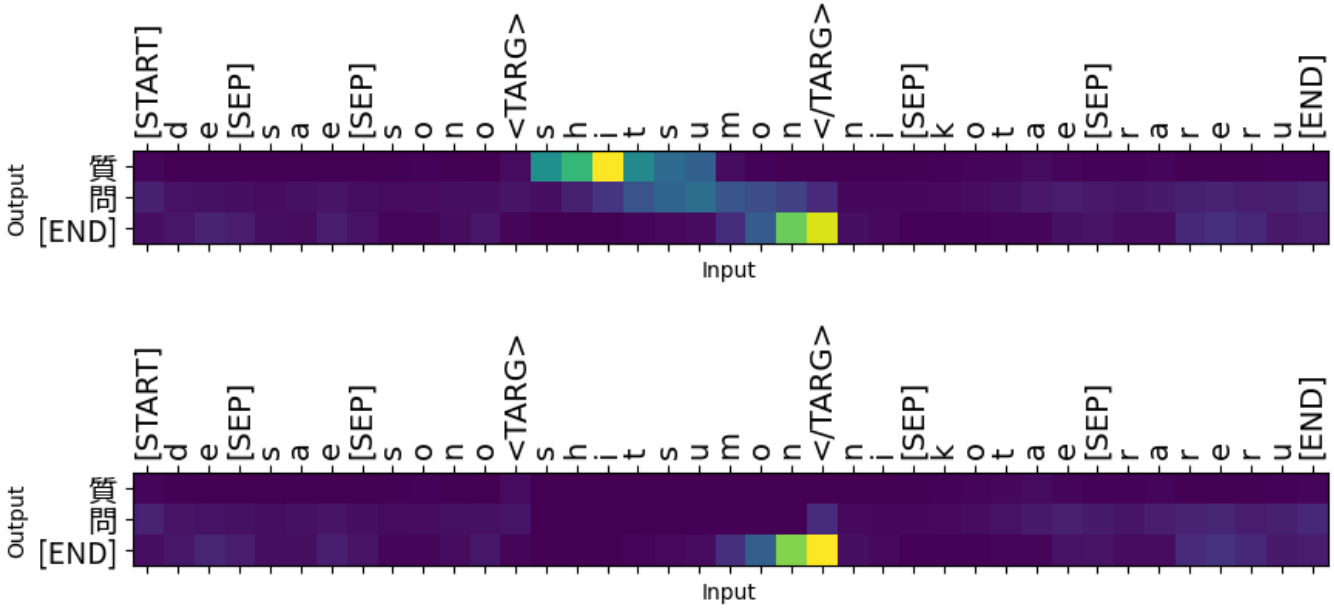


Figure: モデルの最終層の attention を取得し（上図)、ターゲットの語の部分をマスクする（下図)。各 attention 行列の要素の和を計算し、それぞれの行列の要素の和の比がその語の表語性スコアである。コーパス全体で平均を取った数値を、その言語の表語性スコア S_{token} と見なす。

スウェーデン語、ロシア語、フランス語、日本語を用いて検証

実験設計: 仮説３（音素の数）

	書記素数	音素数
タタール語	43	43
アブハズ語	41	66
ポーランド語	40	36
リトアニア語	39	52

Table: 実験対象とする二つの言語ペア。音素数は Phoible 2.0 (Moran and McCloy, 2019) より。音素数の報告が二つ以上存在する場合、平均値を用いる。

書記素数が概ね同じで、音素数が大きく異なる二つの言語ペアの結果を比較する

結果: 仮説 1（書記素の数）

	書記素数	CER (%)
日本語（漢字仮名交じり）	1,702	78.63
日本語（カタカナのみ）	92	25.37
日本語（ローマ字のみ）	27	**17.25**

Table: 仮説 1 の実験結果.

書記素の数が多ければ多いほど、モデルは正しい書き起こしに苦戦している

結果: 仮説 2（表語性）

	S_{token}	CER (%)
スウェーデン語	0.35	23.17
ロシア語	0.46	16.47
フランス語	0.57	22.05
日本語（漢字仮名交じり）	0.97	78.63

Table: 仮説 2 の実験結果．表語性スコア S_{token} は Sproat and Gutkin (2021) の報告による。

日本語（漢字仮名交じり）の精度のみ著しく悪いが、それ以外の言語で表語性スコアと音声認識精度の間に相関は認められない

結果: 仮説 3（音素の数）

	書記素数	音素数	CER (%)
タタール語	43	43	24.18
アブハズ語	41	66	**16.66**
ポーランド語	40	36	**14.88**
リトアニア語	39	52	22.06

Table: 仮説 3 の実験結果.

音素数の多さが CER に悪影響を及ぼしているとは言えない

（予稿の時点での）結論？

設定した仮説のうち、音声認識精度に影響を明確に与えているものは**書記素の数**（仮説 1）のみ

本当に...？

再実験

（以後の実験結果は予稿提出後に行われたものであり、予稿には含まれていません。結果と実験に用いたコードは GitHub (https://github.com/ctaguchi/NLP2024ASRcomplexity) にまとめています。）

- Sproat and Gutkin (2021) の表語性スコアの結果が疑わしいため、表語性スコアの計測を再現実験
- より一般的な結果を得たいため、実験対象の言語を拡大

再実験: 手法

- より多様な言語・文字種
  - 韓国語: ハングル (한글)、ハングル字母 (ㅎ ㅏㄴ ㄱㅡㄹ)
  - 中国語: 漢字、注音符号 (ㄏㄢˋ ㄗˋ)、ピンイン (hànzì)
  - タイ語: タイ文字
  - アラビア語: アラビア文字
  - 英語、フランス語、イタリア語、チェコ語、スウェーデン語、オランダ語、ドイツ語: ラテン文字
- 加えて、表音文字を使用する以下の言語についても実験 (S_{token} データなし)
  - リトアニア語、ポーランド語、バスク語、インドネシア語、カビル語、スワヒリ語、ハンガリー語、ロシア語、タタール語、アブハズ語、ジョージア語、アルメニア語、ヒンディー語

再実験: 結果

言語	書記体系	CER (%)↓	書記素数	S_{token}	音素数
日本語	漢字仮名交じり	78.63	1702	44.27	27.00
	カタカナ	25.37	92		
	ローマ字	17.25	27		
韓国語	ハングル	30.12	965	25.67	39.50
	ハングル字母	18.25	62		
中国語	漢字	75.79	2155	41.59	42.50
	注音符号	13.65	49		
	ピンイン	12.49	56		
タイ語	タイ文字	24.00	67	20.55	40.67
アラビア語	アラビア文字	42.72	53	21.57	37.00
英語	ラテン文字	4.36	27	19.17	41.22
フランス語	ラテン文字	22.05	69	20.37	36.75
イタリア語	ラテン文字	16.63	48	21.28	43.33
チェコ語	ラテン文字	19.19	46	20.57	39.00
スウェーデン語	ラテン文字	23.17	34	19.81	35.00
オランダ語	ラテン文字	14.20	36	19.67	49.38
ドイツ語	ラテン文字	9.43	48	18.03	40.00

Table: 再実験の結果のまとめ（抜粋）

再実験: 結果（書記素数と CER）

言語	書記体系	CER (%)↓	書記素数	S_{token}	音素数
日本語	漢字仮名交じり	**78.63**	**1702**	44.27	27.00
	カタカナ	**25.37**	**92**		
	ローマ字	**17.25**	**27**		
韓国語	ハングル	**30.12**	**965**	25.67	39.50
	ハングル字母	**18.25**	**62**		
中国語	漢字	**75.79**	**2155**	41.59	42.50
	注音符号	**13.65**	**49**		
	ピンイン	**12.49**	**56**		

Table: 再実験の結果のまとめ（抜粋）。大まかな傾向として、やはり書記素数が多いほど CER が悪化している。

再実験: 結果（相関）

	CER	書記素数	S_{token}	音素数
CER	1.00	0.89	0.95	−0.50
書記素数		1.00	0.96	−0.30
S_{token}			1.00	−0.44
音素数				1.00
R^2		0.80	0.89	0.25

Table: 再実験の結果に基づいた相関行列。CER と最も相関があるのは表語性スコア S_{token} であり、書記素数がそれに次ぐ。一方で、CER と音素数の間には特に相関は見られない。最下行は CER を目的変数とした重回帰分析の決定係数（R^2）。

結論

- 当初の実験結果（先行研究で報告された表語性スコアを使用）では、書記素数のみが音声認識の精度に影響を与える要因であった
- ところが、表語性スコアの計算の再現実装・追実験を行なった結果、**表語性・書記素数**の順で音声認識の精度に影響を与えていることがわかった
- 一方で、音素数はいずれの実験でも音声認識の精度に影響は与えなかった

おわりに: この知見の嬉しいこと

- **言語習得との関係**:
  - 人間: 音素の学習は普遍的・先天的（第一言語獲得）、識字の学習は後天的・人為的
  - モデル: 音素の学習は言語不問、識字は各言語の正書法の複雑性に大きく依存する
  - → 音韻習得の計算的モデリングの可能性？
- **少数言語音声認識に向けた知見**:
  - 低資源言語が複雑な正書法を持っている場合、音韻的な転写に近い出力に変えると精度が上がるかもしれない
  - ロロ文字（彝文字）、チェロキー文字、カナダ先住民文字など

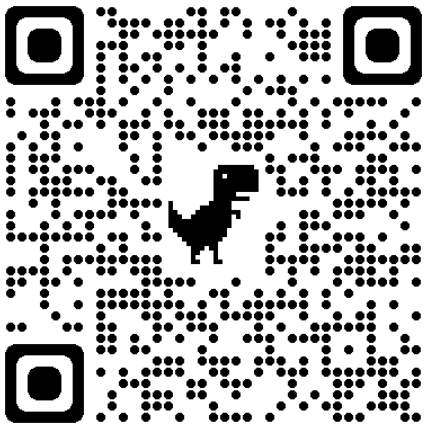
ありがとうございました

ご質問など: ctaguchi@nd.edu

https://ctaguchi.github.io

実験結果・コード:

https://github.com/ctaguchi/NLP2024ASRcomplexity



参照文献

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

S. Moran and D. McCloy. Phoible 2.0, 2019.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

A. Rouditchenko, S. Khurana, S. Thomas, R. Feris, L. Karlinsky, H. Kuehne, D. Harwath, B. Kingsbury, and J. Glass. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272, 2023. doi: 10.21437/Interspeech.2023-1061.

R. Sproat and A. Gutkin. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528, Nov. 2021. doi: 10.1162/coli_a_00409. URL https://aclanthology.org/2021.cl-3.16.

K. Takaoka, S. Hisamoto, N. Kawahara, M. Sakamoto, Y. Uchida, and Y. Matsumoto. Sudachi: a japanese tokenizer for business. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.