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
Department of Computer Science and Engineering, University of Notre Dame
ctaguchi@nd.edu

概要

本研究は，諸言語の正書法や音韻体系の複雑性が，個別言語の音声認識の精度にどのように影響を与えるのかを検証する．音声認識の精度に悪影響を与えうる複雑性として，書記素の数，書記素と音素の対応の不規則性，音素の数，という三つの要因を仮説として提示する．実験の結果，明らかに悪影響となりうる要因は書記素の数のみであり，その他二つの要因は精度にあまり影響を与えないことが明らかとなった．

1 はじめに

人間が第二言語を学習するとき，その言語が複雑な正書法や複雑な音韻論的体系を持っている場合，言語学習は困難になりうる．例えば，漢字を知らない学習者が中国語や日本語を書記言語として学習する場合，数千もの漢字を学ぶために膨大な時間を費やすことになる．また，日本語の「生」という漢字は，十を超える発音の形式を有しており，どの発音形式が選択されるかは文脈に依存するため，日本語の非母語話者にとっては難点となる．他にも，五つの母音しか持たない日本語を母語とする英語学習者にとって，アメリカ英語の十五個の母音の区別を習得することは容易ではない [1]．近年の音声認識（Automatic Speech Recognition; ASR）モデルが非常に高い精度を達成していることを鑑みると，果たして音声認識モデルにとってもこのような言語的複雑性が学習の障壁となるのであろうか，という疑問が生じる．そこで，本研究では，音声認識の精度に影響しうる言語的複雑性として，以下の三つの仮説を提示する．

仮説 1（書記素の数）． ある言語における書記素（grapheme）[1)] の数が多ければ多いほど，音声認識の精度は低下する．上で取り上げた三つの例のうち，一つ目がこの仮説に相当する．

仮説 2（表語性）． ある言語における書記素と音素（phoneme）の形式の間に一対多の関係がみられるほど，音声認識の精度は低下する．上で取り上げた三つの例のうち，二つ目がこの仮説に相当する．なお，本研究では Sproat & Gutkin (2021)[2] に倣い，このような傾向のある言語を表語的（logographic）言語と呼び，反対に，書記素と音素の間に規則的な一対一の関係がみられる傾向のある言語を表音的（phonographic）言語と呼ぶ．つまり，この仮説を言い換えると，ある言語が表語的であればあるほど，音声認識の精度は低下する．

仮説 3（音素の数）． ある言語における音素の数が多ければ多いほど，音声認識の精度は低下する．この仮説は，上で挙げた例のうち，三つ目の例に動機付けられている．

これらの仮説を検証するため，本研究では，訓練済み多言語音声認識モデル Wav2Vec2-XLSR-53[3] を用いて，異なる正書法と異なる音韻体系を持った言語にファインチューニングし，各言語の認識精度を比較する．

2 先行研究

多言語音声認識． Transformer を利用したアーキテクチャの発展とともに，多言語自動音声認識の分野もめざましい発展を見せている．Wav2Vec 2.0[4] はその成果の一つであり，テキスト言語処理における BERT[5] のように，自己教師あり学習によって音声情報の表現を学習する．固定長の各音声フレームに対し，まず畳み込みニューラルネットワーク（CNN）を用いて特徴抽出を行い潜在表現 $\boldsymbol{z}$ を得た後，積量子化（product quantization）[6] を用いて離散化されたベクトル $\boldsymbol{q}$ を計算する．次に，既定の確率に基づいて $\boldsymbol{z}$ をマスクし，このマスクされた表現の量子化ベクトル $\boldsymbol{q}$ を予測することを自己教師あり学習の訓練目標とする．

1） 意味を区別しうる最小の書記単位．本研究では，Unicode 上で区別される文字を書記素とする．

Wav2Vec 2.0 は，事前学習済みモデルをファインチューニングすることで様々な音声認識タスクにおいて良い性能を発揮することが報告されている．中でも，本研究で用いる Wav2Vec2-XLSR-53[3] は，53 言語の計 5 万時間のラベル無し訓練データに基づいて事前学習され，少量のデータを用いてファインチューニングするだけで，事前訓練データに含まれていない未知の言語を含めた多様な言語の音声認識を学習することができる．

なお，いくつかの言語での音声認識タスクの SOTA は Whisper[7] に取って代わられている．Whisper は，多言語のデータに基づいて弱教師あり学習によって訓練されたエンコーダー・デコーダー音声認識モデルであり，同じくファインチューニングによって個別言語の音声認識タスクの精度を向上させることができる．しかしながら，Whisper は元の訓練データに含まれている言語では Wav2Vec 2.0 の性能を上回るものの，事前学習データに含まれていない言語に対しては，Wav2Vec 2.0 が概してより良い精度を持つことが報告されている [8]．加えて，Whisper の事前学習では一部の訓練データにラベルが含まれているため，あらかじめ訓練言語の正書法での書き起こしを学習している．この点は，純粋に音声の表現のみを事前学習する Wav2Vec 2.0 系のモデルとは大きく異なる．したがって，本研究では，音声・音韻と正書法はそれぞれ異なる学習上の難点となることを仮定していることから，事前学習モデルとして Whisper ではなく Wav2Vec 2.0 を用いることとする．

表語的複雑性． いくつかの言語では，正書法とその発音の間に乖離がみられる．Sproat & Gutkin (2021)[2] は，このような言語を表語的言語と呼んでいる．このような言語では，複数の読み方を有する文字の発音を特定するために，前後の文脈に注意を払うことが必要となる．Sproat & Gutkin (2021) は，attention matrix にみられる attention の拡散を用いることで，正書法がどの程度文脈に依存しているのかを計算する手法を提案している．詳細は 4.1 を参照されたい．

3 手法

3.1 データセット

本実験では，全ての実験設定において，Common Voice 13.0[9] を用いる．各言語の音声データ量をなるべく均一にするため，各言語の音声データの長さが合計一万秒に達するまで訓練データを抽出する．また，学習中のメモリ不足を予防するため，15 秒を超える音声データは取り除いておく．

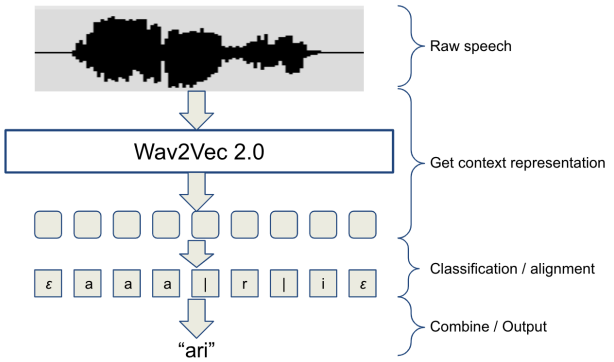

図 1 ファインチューニング時の推論例の図．

3.2 事前学習モデルと追加学習

本研究では，全ての実験において Wav2Vec2-XLSR-53[2] を事前学習モデルとして用いる．ファインチューニングでは，事前学習モデルをもとに対象言語のデータをもとに，コネクショニスト時系列分類法 (CTC) を用いた教師あり学習を行う．モデルは，音声の入力を固定長のフレームに分割し，それぞれの文脈表現（context representation）を計算し，文字全体の確率分布を得る．訓練目標は，式 (1) に定義される CTC 損失を最小化することである．ここで，$\boldsymbol{x}$ は入力音声，$\boldsymbol{y}$ は出力の書き起こし，そして $\mathcal{D}$ は訓練データである．

$$\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} -\log p(\boldsymbol{y}|\boldsymbol{x}) \tag{1}$$

確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ は次のように求められる．

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{A\in\mathcal{A}_{\boldsymbol{x},\boldsymbol{y}}} \prod_{t=1}^{T} p_t(a_t|\boldsymbol{x}) \tag{2}$$

ここで，$\mathcal{A}_{\boldsymbol{x},\boldsymbol{y}}$ は，入力音声 $\boldsymbol{x}$ を元に正しい書き起こし $\boldsymbol{y}$ を出力しうる全ての可能なアラインメントの集合，$A = a_1, \cdots, a_T$，そして T は入力音声 $\boldsymbol{x}$ のフレーム数である．推論の段階では，最も確率の高いアラインメントを選択する．すなわち，$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x})$．推論の過程の例を図 1 に図示する．実験では，全ての追加学習において同一のハイパーパラメータを用いる．特に，エポック数は 20，学習率は 0.0003 に統一する．

2) https://huggingface.co/facebook/wav2vec2-large-xlsr-53

4 実験設計

4.1 仮説検証のための設定

仮説 1（書記素の数） 仮説 1 を検証するために，同じ言語の書記素の数の異なる正書法を用いて，モデルを追加学習する．対象言語として用いるのは日本語であり，漢字仮名混じり，カタカナのみ，ローマ字のみの三つの正書法を用いる．元のデータセットの漢字仮名交じり文は，SudachiPy[10] を用いてカタカナのみおよびローマ字のみの文に変換する．

仮説 2（表語性） Sproat & Gutkin [2] は，attention を用いて表語性スコアを以下のように算出している．まず，attention 行列 A とマスク行列 M が与えられたとき，$M \circ A$ はそれらのアダマール積である．マスク行列 M は A と同じサイズの行列であり，k を対象の語の発音の長さ，m，n を対象の語のそれぞれ左端と右端の位置とすると，$0 \leq i < k$ かつ $m \leq j \leq n$ のとき，$A_{i,j} = 0$ である．このとき，語 S_w の attention spread は，次のように計算される．

$$S_w = \frac{\sum_{i,j}(M \circ A)_{i,j}}{\sum_{i,j} A_{i,j}} \tag{3}$$

これを応用して，式 (4) のように，コーパス全体における語の attention spread の平均を言語の表語性スコア S_{token} として用いることができる．ここで，N はコーパスの語数である．

$$S_{\text{token}} = \frac{\sum_w S_w}{N} \tag{4}$$

本実験では，異なる表語性スコアを持った対象言語としてスウェーデン語，ロシア語，フランス語，日本語を用いる．Sproat & Gutkin (2021)[2] によると，これらの言語の表語性スコアはそれぞれ 0.35, 0.46，0.57，0.97 である．

仮説 3（音素の数） 仮説 3 を検証するために，書記素の数が概ね同じだが音素の数が大きく異なる二対の言語ペアをもとに追加学習を行い，性能を比較する．本実験では，タタール語・アブハズ語ペア（音素数各 38 個・60 個）とポーランド語・リトアニア語ペア（音素数各 35 個・56 個）を用いる．なお，音素数のデータは Phoible 2.0[11] に基づく．

4.2 評価指標

ASR の性能を測る指標として，文字誤り率（Character Error Rate; CER）を用いる．S，D，I をそれぞれ置換数，削除数，挿入数とし，N を正解ラベルの文字数とすると，CER は次のように求められる．

$$\text{CER} = \frac{S + D + I}{N} \tag{5}$$

なお，単語誤り率（Word Error Rate; WER）も音声認識の評価指標として広く用いられているが，何が「語」であるかは言語間で大きなばらつきがあるため，言語間比較を主軸とする本研究では用いない．

	書記素数	CER (%)
日本語（漢字仮名交じり）	>2,000	33.43
日本語（カタカナのみ）	83	22.42
日本語（ローマ字のみ）	22	**17.71**

表 1 仮説 1 の実験結果．

	S_{token}	CER (%)
スウェーデン語	0.35	21.72
ロシア語	0.46	22.37
フランス語	0.57	20.50
日本語（漢字仮名交じり）	0.97	33.43

表 2 仮説 2 の実験結果．

5 結果

仮説 1（書記素の数） 仮説 1 の検証の実験結果は，表 1 にまとめられている．CER を比較すると，最も性能が悪かったのは漢字仮名交じり文であり，最も書記素数が少ないローマ字文が最も性能が良いことがわかる．ゆえに，この条件下では仮説 1 が確かめられた．

仮説 2（表語性） 仮説 2 の検証結果をまとめた表 2 によると，この比較で最も性能が悪かったのは日本語（漢字仮名交じり文）であるが，これはおそらく第 5 節で見た膨大な書記素の数に起因すると考えられる．その他の言語では，表語性スコア S_{token} と ASR の性能の間に相関は見受けられない．よって，仮説 2 は棄却される．

仮説 3（音素の数） 仮説 3 の検証結果を 3 にまとめる．各言語対の二言語を比較すると，音素数と ASR の精度の間には，正の相関が見られないばかりか，むしろ負の相関があるようにすら見受けられる．したがって，仮説 3 を棄却する．

これらの結果から，ASR の性能を左右する言語的複雑性の仮説として，書記素の数のみが立証された．

	書記素数	音素数	CER (%)
タタール語	39	38	27.10
アブハズ語	40	60	14.84
ポーランド語	32	35	16.41
リトアニア語	32	56	12.93

表 3　仮説 3 の実験結果.

6　議論：第一言語音韻習得との関係

上述の実験結果は，多言語事前学習モデル Wav2Vec2-XLSR-53 を用いた場合，音韻の複雑性や正書法と音韻の対応の複雑性は音声認識の精度に負の影響を与えないものの，正書法の複雑性は影響を与えることを示した．本節では，この事前学習モデルの追加学習と人間の第一言語の音韻習得との関係について論じる．言語学の生成文法理論における普遍文法（Universal Grammar）は，人類が第一言語を習得することを可能にする計算体系の初期状態であり [12]，人間の幼児が極めて限られたインプットを元に第一言語の音韻・形態・統語といった構造を習得する事実を説明しうる理論的仮説として提唱されている [13]．本実験で見たような Wav2Vec 2.0 の多言語事前学習モデルと個別言語に特化したファインチューニングの学習は，普遍文法および第一言語音韻習得といくつかの興味深い類似点を有している．第一に，人間は第一言語の音韻構造の複雑性に関わらず第一言語の音韻を習得することができるが，Wav2Vec 2.0 の多言語事前学習モデルも追加学習においては弁別される音素の数による負の影響は見られなかった．第二に，人間の第一言語習得過程とモデルの追加学習過程の双方において，学習に用いられるインプットが少ないことである．人間の幼児は生後わずか一ヶ月ですでに第一言語の音素を区別し始めることが報告されているが [14]，本実験において，モデルもわずか約 2.8 時間のラベル付きデータで音素を区別していることが見られた．第三に，書記言語の習得に膨大な学習時間が必要である点である．人間による第一言語の書記体系の習得は後天的であり，日本の学校制度において少なくとも中学三年次まで漢字教育が行われているように，音声言語の習得と比較すると，より意識的な学習をより長期にわたって行う必要がある．同様に，Wav2Vec 2.0-XLSR-53 の性能も書記体系の複雑さに大きく影響されることが実験で示されたように，書記体系が複雑な言語に関しては比較的より多くのラベル付き訓練データが必要となることが考えられる．これらの共通点から，Wav2Vec 2.0 のようなアーキテクチャを用いた自己教師あり学習済み音声言語モデルは，普遍文法の音韻モジュールを，有限の言語データを元にパラメータとして近似的に表現したものであると考えることができる．しかしながら，実験結果はこの仮説を科学的に裏付けるものではない上に，普遍文法の実在性についても言語学において決着がついていない問題であるため，本節ではあくまでも可能性としての言及に留める．

7　おわりに

本研究では，多言語自動音声認識タスクにおいて，正書法上の複雑さや音韻構造の複雑さが書き起こし精度の低下の要因となるのかどうかを，自己教師あり学習済み多言語モデル Wav2Vec2-XLSR-53 を用いて検証した．実験結果は，言語における書記素の数が性能低下の要因につながることを示した反面，表語性の高さや音素の数といった複雑性による性能の違いは見られなかった．また，興味深いことに，このような学習の傾向は，人間の第一言語の音韻習得および書記習得とのいくつかの類似点を持っており，自己教師あり学習済み多言語音声モデルを普遍文法（あるいは言語獲得装置）の音韻モジュールをパラメータによって具体的に近似したものであると考えられる可能性を示した．

最後に，本研究の方法論的限界に言及する．まず，データとして用いた Common Voice はボランティアによって音声が読み上げられているため，読み上げ音声に誤りが含まれることが珍しくなく，必ずしも良質なデータであるとは言えない．また，本実験では十個の言語設定でしか比較していないため，実験の結果を確認するためにより多くの言語で追実験する必要がある．関連して，音韻的複雑さは弁別音素の数以外にも多くの要因があり，音素数の実験結果のみから音韻的複雑さが性能に影響を与えないとは言い切れない．最後に，評価指標として用いた CER は，音声認識の性能評価で広く用いられている指標の一つであるが，正書法の違いを考慮していないため，中国語のように一つの文字が複数の音素をまとめて表現する正書法を使用する言語に対して不利になっている可能性がある．これらの点は，本研究の問題設定のみならず，ASR の分野全体に広く関わる問題点であるため，今後の課題とする．

謝辞

本研究はアメリカ国立科学財団研究費 No. BCS-2109709 の助成を受けたものです．This material is based upon work supported by the National Science Foundation under Grant No. BCS-2109709.

参考文献

[1] Peter Ladefoged and Sandra Ferrari Disner. **Vowels and consonants**. Wiley Blackwell, 3 edition, 2012.

[2] Richard Sproat and Alexander Gutkin. The taxonomy of writing systems: How to measure how logographic a system is. **Computational Linguistics**, Vol. 47, No. 3, pp. 477–528, November 2021.

[3] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 33, No. 1, pp. 117–128, 2011.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[8] Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In **Proc. INTERSPEECH 2023**, pp. 2268–2272, 2023.

[9] 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)**, pp. 4211–4215, 2020.

[10] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Paris, France, may 2018. European Language Resources Association (ELRA).

[11] Steven Moran and Daniel McCloy. Phoible 2.0, 2019.

[12] Brett Miller, Neil Myler, and Bert Vaux. Phonology in Universal Grammar. In Ian Roberts, editor, **The Oxford Handbook of Universal Grammar**. Oxford University Press, 2016.

[13] Noam Chomsky. **Syntactic structures**. De Gruyter Mouton, Berlin, Boston, 1957.

[14] Peter D. Eimas, Einar R. Siqueland, Peter Jusczyk, and James Vigorito. Speech perception in infants. **Science**, Vol. 171, pp. 303–306, 1971.