日本語音声基盤モデルは英語をどう聞くか？ — 音韻知覚の分析 —*

☆白井 透，Bagus Tris Atmaja, Sakriani Sakti (NAIST)

1 はじめに

自己教師あり学習音声モデル (SSL モデル) は，ラベルのない音声から表現を学習し，大量の音声データを用いた事前学習によって，音声認識や感情認識などのタスクで高い性能を示している。また，これらのモデルは音響的・韻律的特徴を自動的に獲得する。

一方で, こうして得られる表現がどの程度学習言語に依存しているのか，特に日本語音声で事前学習された SSL モデルが英語音声をどのように知覚するのかについては，十分に明らかになっていない。そこで本研究では，英語音声および日本語音声を用いて事前学習された SSL モデルの音素認識精度を比較することで，SSL モデルの表現における言語依存性を明らかにする。

2 関連研究

SSL モデルが獲得する表現の言語依存性に関する研究はこれまでも行われてきた。英語およびフランス語音声で学習した SSL モデルの埋め込みを用いて音素弁別実験を行い, 人間の音素知覚の傾向と比較した結果, SSL モデルの表現は母語による知覚バイアスを十分に再現できていないことが示された [1]。同様の傾向が, ヒンディー語音声で事前学習された SSL モデルを用いた分析においても報告されている [2]。

さらに, フランス語・中国語・ベトナム語音声で事前学習された SSL モデルの全ての層の埋め込みの分析も行われている [3]。この研究では, 音素および声調に関しては, 中間層の表現において学習言語に対応する音素・声調分類タスクでわずかな性能向上が見られ, 言語依存性が部分的に示されている。

3 実験的評価

3.1 実験条件

本稿では，SSL モデルから得られる埋め込み表現の音素弁別能力を評価するため，ABX テストを用いる。ABX テストは，3 つの音声刺激 A・B・X からなる triplet を提示し，刺激 X が A と B のいずれに近いかを判定する手法である。triplet において，X と同一の音素を含む刺激を target，異なる音素を含む刺激を other と定義する。target，other，および X に対応する SSL モデルの埋め込みをそれぞれ $M_{\text{target}}, M_{\text{other}}, M_X$ とし，次式で定義される Δ 値を計算する。

$$\Delta = \mathrm{DTW}(M_{\text{other}}, M_X) - \mathrm{DTW}(M_{\text{target}}, M_X) \quad (1)$$

ここで DTW は，フレーム単位のコサイン距離を動的時間伸縮法により集約した距離を表す。Δ 値が正のとき，音素を正しく弁別できていることを意味する。

本研究では，英語音声で事前学習された HuBERT および wav2vec 2.0 に加え，日本語音声で学習された kushinada と izanami の計 4 種類の base サイズの SSL モデルを用いて評価を行う。構築手法としては kushinada に HuBERT, izanami には wav2vec 2.0 が用いられている。先行研究において中間層の表現が音韻的情報の抽出に有効であることが示されている [4] ため，本実験では第 6 層の埋め込み表現を用いて評価を行う。

英語音素の ABX テストには, 73 話者による合計約 20 時間の読み上げ音声で構成されている英語データセット [5] を利用する。この実験では，英語に含まれる多様な音素対を対象とし，英語音素全般に対してどの程度の弁別能力を有するかを検証する。

人間の音素知覚との比較では，268 人の日本人大学生を対象とした音素テスト [6] の結果を用いる。モデルの正答率と人間の正答率の間に順位相関が存在するかを評価するため，スピアマンの順位相関係数を用いて相関を計算する。ここでは，15 種類の母音と 13 種類の子音を用いる。

次に，子音連結の弁別性を評価する。日本語話者は，子音結合間に存在しない母音を知覚する傾向があることが示されている [7]。本実験では，子音連結を含む単語の子音結合箇所を用いた ABX テストを実施する。日本人および米語話者による読み上げ英語音声から構成される UME-ERJ コーパス [8] を用い，A および X にはそれぞれ異なる米語話者の発話，B には日本語話者の発話を割り当てる。これにより，米語話者にとって自然な子音連結を基準の刺激とし，日本語話者の発話に見られる母音挿入を伴う発音との差異を，モデルがどの程度弁別できるかを評価する。英語音声には信号対雑音比の高い男女各 1 名の発話を使用する。音素レベルの時間的アラインメントを取得するため, Montreal Forced Aligner[9] を用いて, 音声と音素のアラインメントを行った。このアライメントに基づいて，子音連結を自動検出した。

*How Do Japanese-Based Speech Models Hear English? — An Analysis of Phoneme Perception —, by Toru SHIRAI, Bagus Tris ATMAJA, Sakriani SAKTI (Nara Institute of Science and Technology).

Table 1　事前学習設定と ABX テストの結果

モデル	事前学習		ABX 正答率 (%)	
	言語	音声データ量 (h)	英語音素弁別	子音連結弁別
HuBERT	英語	960	89.6	81.2
wav2vec 2.0	英語	960	87.8	78.5
kushinada	日本語	62,215	88.2	79.8
izanami	日本語	5,315	85.0	80.2

Table 2　モデルの弁別結果と人間知覚との順位相関

モデル	子音	母音
kushinada	-0.379	-0.565
izanami	-0.374	-0.635

3.2　実験結果

Table 1 に，各 SSL モデルにおける英語音素弁別および子音連結弁別の ABX テストの結果を示す。また，Table 2 には，英語音素に関するモデルの弁別結果と人間の知覚実験との順位相関を示す。

まず，英語音素の ABX テストにおいては，英語音声で事前学習された HuBERT が 89.6% と最も高い性能を達成した。日本語音声のみで事前学習された kushinada も 88.2% と英語学習モデルに近い性能を示しており，日本語音声に基づく埋め込み表現であっても，英語音素の弁別能力を有することが確認された。一方，izanami は他のモデルと比較してやや低い正答率にとどまっており，学習データ量が性能に影響している可能性が考えられる。

さらに，英語音素に関する人間の知覚実験との順位相関を分析した結果，日本語音声で学習された kushinada および izanami のいずれにおいても負の相関が観測された。この結果により，SSL モデルの音素知覚は，人間の音素知覚とは異なる傾向を持つことが示された。

次に，子音連結に対する ABX テストでは，すべてのモデルにおいて英語音素弁別と比較して正答率が低下する傾向が見られた。英語音素弁別と同様に，英語音声で事前学習された HuBERT が 81.2% と最も高い性能を示した。HuBERT を基盤として構築されたモデル間の性能の大小関係は，英語音素弁別テストの結果と概ね一致している。一方，wav2vec 2.0 を基盤とするモデルでは，英語音素弁別時とは異なる傾向が見られ，この差異が生じた要因の調査が今後の課題である。

4　おわりに

本研究では，日本語音声で学習された音声基盤モデルが英語音声をどのように知覚しているのかを明らかにするため，ABX テストを行った。その結果，日本語音声で学習されたモデルであっても英語音素の弁別に対して一定の汎化能力を示す一方で，人間の音素知覚とは異なる傾向を持つことが確認された。また，日本語話者に特有の母音挿入を伴う子音連結に対しては，英語音素単体の場合と比べて弁別が困難となる傾向が見られた。

謝辞　本研究の一部は，JSPS 科研費 JP21H05054, JP23K21681, および JP24K02967 の助成を受けて実施したものである。

参考文献

[1] J. Millet et al.,“Do self-supervised speech models develop human-like perception biases?,”Proc. of ACL, 2022.

[2] J. Rodriguez et al.,“Self-supervised speech representations display some human-like cross-linguistic perceptual abilities,”Proc. of CoNLL, 2024.

[3] M. Gubian et al.,“Analyzing the relationships between pretraining language, phonetic, tonal, and speaker information in self-supervised speech models,”arXiv:2506.10855, 2025.

[4] A. Pasad et al.,“Layer-Wise Analysis of a Self-Supervised Speech Representation Model,”Proc. of ASRU, 2021.

[5] T. A. Nguyen et al.,“The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling,”arXiv:2011.11588, 2020.

[6] 大塚朝美.,”日本人大学生を対象とした英語の単音及び単語間の音のつながりの知覚と自己モニターを取り入れた教授法に関する研究”，博士論文，2020.

[7] E. Dupoux et al.,”Epenthetic vowels in Japanese: A perceptual illusion?”, Journal of Experimental Psychology: Human Perception and Performance, 1999.

[8] 中川聖一，”特定領域研究「メディア教育利用」日本人学生による読み上げ英語音声データベース（UME-ERJ）”，国立情報学研究所 音声資源コンソーシアム，2007.

[9] M. McAuliffe et al.,“Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,”Proc. of Interspeech, 2017.