音声モデルにおける Critical Period 仮説の検証

古賀 友里愛 [1]　神藤 駿介 [1]　宮尾 祐介 [1,2]
[1] 東京大学　[2] 国立情報学研究所大規模言語モデル研究開発センター
{ykrasp7isweet,skando,yusuke}@is.s.u-tokyo.ac.jp

概要

本研究では，自己教師あり学習音声モデル（SSL モデル）の第二言語（L2）獲得過程を，Critical Period（CP）仮説の観点から分析する．L2 獲得における CP 仮説とは，人間は L2 への接触開始時期が遅いほど，その習得が困難になるとするものである．CP に注目することは，SSL モデルの学習メカニズムや効率的な L2 学習手法に加え，人間の脳の言語学習の仕組みに関する新たな示唆をも与える可能性がある．実験の結果，SSL モデルでは L2 の音韻獲得における CP 仮説は成り立たなかったが，早期に L2 の学習を開始したモデルは L1 モノリンガルモデルや初めから 2 言語で学習したモデルとは異なる埋め込みを獲得していることが示唆された．

1　はじめに

近年，自己教師あり学習音声モデル（SSL モデル）は様々なタスクで高い精度を達成しており，その学習過程を人間と比較する研究が増加している [1, 2, 3]．こういった比較は，SSL モデルの学習の仕組みや効率的な学習手法，人間の脳における学習の仕組みに関して新たな示唆を与える可能性がある．先行研究では，第一言語（L1）における比較は多く行われている [1, 2] 一方，第二言語（L2）における研究は依然として限定的である．

本研究では，SSL モデルの L2 獲得過程を Critical Period（CP）仮説 [4, 5] の観点から分析する．CP 仮説とは，人間が特定の時期（CP）を過ぎると言語獲得が困難になるとする仮説であり，音韻，文法，意味といった様々な言語能力の観点から広く議論されている [6]．L2 獲得においては，図 1 のように，L2 への接触時期が遅いほどその習得が困難になり，CP の終了後に初めて L2 に触れると L2 を完全には習得できないとされる．本研究では，L2 の音韻獲得における CP 仮説に着目し，SSL モデルとして HuBERT [7] を用いた分析を行う．具体的には，

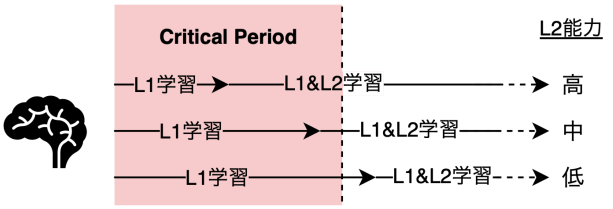

図 1　人間の L2 獲得における Critical Period（CP）仮説の概念図．本研究では HuBERT で CP 仮説を検証する．

L2 の学習開始時期を変化させて HuBERT を訓練し，「L2 学習期間が等しい場合，L2 の学習開始時期が遅いほど最終的な L2 能力が低下するか」を検証する．加えて HuBERT の L2 獲得過程を観察し，L2 学習開始時期の違いがモデルに与える影響を分析する．以降，CP は L2 の音韻獲得における CP を指す．

実験の結果，HuBERT では CP 仮説は成り立たなかったものの，早期に L2 学習を開始したモデルでは L1, L2 とも正解率が向上し，L2 に対する可塑性が示唆された．また，L2 の学習開始時期が早いモデルは，遅いモデルと異なり，L1 のみで学習したモデルや初めから 2 言語で学習したモデルとは異なる埋め込みを獲得していることが示された．

2　関連研究

L2 知覚における音声モデルと人間の比較　音声モデルと人間の L2 知覚を比較する研究はこれまでにも行われており．従来は RNN ベースのモデルに L1 や L2 を学習させ，音声弁別 [8] や単語の意味認識 [9] などのタスクにおける比較が行われてきた．近年では自己教師あり学習音声モデル（SSL モデル）が注目され，例えば Contrastive Predictive Coding（CPC）[10] を用いた研究では，人間の発達過程において L1 に存在しない音の弁別が困難になる現象が，モデルで再現されるかを検証している [11]．この研究では，子供に向けた音声データで L1 を学習した場合はこの現象は再現されず，クリーンな音声データを使用すると再現されることが示された．一方で，HuBERT [7] と wav2vec2.0 [12] では同様の現象

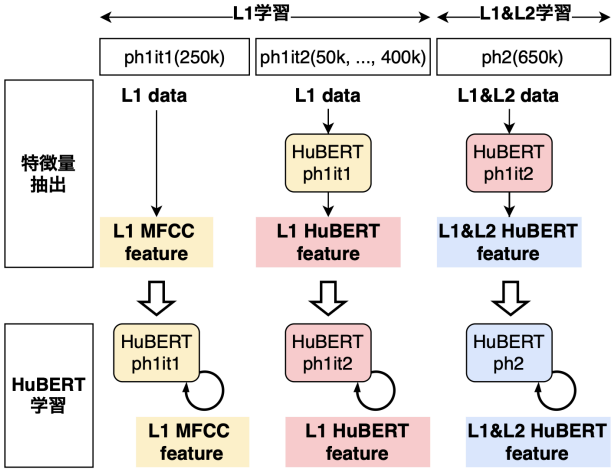

図 2　実験概要．ph1it1 → ph1it2 → ph2 の順に，ph1it2 のステップ数を変化させて学習することで CP を検証する．

は確認されていない [13]．SSL モデルを用いた研究ではモデルは L1 のみで学習されているが，本研究では HuBERT を L1 と L2 で学習し，CP という新たな観点から音声モデルと人間の L2 知覚を比較する．

Critical Period（CP）　L2 獲得における CP 仮説 [4, 5] とは，人間は L2 への接触開始時期が遅いほどその習得が困難になり，CP の終了後に初めて L2 に触れた場合，その完全な習得が不可能になるとする仮説である．テキスト言語モデルにおける CP 仮説の検証は既に行われている [3] が，音声モデルではまだ研究されていない．テキスト言語モデル，例えば RoBERTa [14] や GPT-2 [15] では，L2 の学習開始時期を遅らせることによる BLiMP や GLUE [16] の精度向上が確認されており，CP 仮説は成り立たないとされている [3]．また，CP より広く言語転移について分析した研究 [17] でも，L1 での事前学習後に L1 と L2 の翻訳ペアで XLM [18] を学習した場合，事前学習無しの場合に比べ，最終的な L2 の BLiMP [19] 精度が向上することが示されている．本研究では，音声モデルにおける CP 仮説を検証するため，L2 の学習開始時期をより幅広く変化させた実験を行う．音声情報はテキスト情報よりも人間の言語獲得への寄与が大きいため，音声モデルを用いることで人間により近い設定が実現可能となる．

3　実験

本研究では，L2 の学習開始時期を変化させて SSL モデルを学習し，「L2 学習期間が等しい場合，L2 の学習開始時期が遅いほど最終的な L2 能力が低下するか」を検証する．ここで，L1 は英語，L2 はフランス語とする．実験の概要を図 2 に示す．

表 1　データセット統計．Train, Valid 列は音声ファイル数，平均時間列は全音声ファイルでの平均時間．

言語	データセット	Train	Valid	合計時間	平均時間
EN	Providence	185,475	1,852	129h	2.48s
FR	Lyon	83,495	815	83.7h	3.57s

3.1　データセット

モデルを可能な限り人間に近い方法で学習するため，学習データは L1, L2 とも子供向け発話データベースの CHILDES [20] から取得する．L1 については，Voice Activity Detection（VAD）による音声区間の抽出や，子供による発話の除去といった前処理 [21] が施された Providence データ [22] を用いる．L2 については，Lyon データ [23] に同様の前処理を施し，モノラル音声に変換した上で使用する．話者のアノテーションが存在しない音声ファイルに関しては全ての発話を用いる．データセットの統計情報を表 1 に示す．Train, Valid データはランダムに 99:1 の比率で分割する．全ての音声データは 16kHz でサンプリングされており，サンプル数が 8,000 以下の短い音声ファイルは除去する．

3.2　モデル

本研究では，SSL モデルとして HuBERT [7] を採用し，FAIRSEQ [24] の実装を用いて実験を行う．HuBERT は主に CNN エンコーダと後続の BERT エンコーダで構成されており，BERT エンコーダは 12 層の Transformer 層からなる．学習手法は Masked Language Modeling（MLM）に基づいており，マスクされたフレームに対して k-means クラスタリングで生成された擬似ラベルを予測する．学習は 2 回のイテレーション（it1, it2）に分かれており，各イテレーションで異なる擬似ラベルを使用する．it1 では MFCC の特徴量のクラスタリングにより生成されたラベルを使用し，it2 では it1 で学習したモデルの 6 層目の Transformer から抽出した特徴量のクラスタリングにより生成されたラベルを用いる．

本実験では 2 言語を学習するため，L1 学習（ph1）と L1&L2 学習（ph2）の 2 段階で HuBERT の学習を行う（図 2）．各段階での学習手順を以下に示す．

3.2.1　ph1: L1 学習

HuBERT 本来の学習方法に倣い，it1 を 250k ステップ学習した後に it2 を 400k ステップ学習する．

L2 学習，つまり ph2 の開始時期を変化させるため，it2 の学習が 50k, 100k, ⋯, 400k ステップ終了した時点でそれぞれチェックポイントを保存し，そこから 3.2.2 節で説明する ph2 の学習を行う．学習データは Providence のみを用いる．

3.2.2 ph2: L1&L2 学習

ph1 で保存した 8 つのチェックポイントから，さらに L1 および L2 で学習を行う．1 回のイテレーションのみで 650k ステップの学習を実施する．クラスタリングの特徴量は ph1it2 と同様に，ph1it2 で学習したモデルの 6 層目の Transformer の特徴量を使用する．Train, Valid データは Providence と Lyon の Train, Valid データを連結したものとする．ph2 の学習後に得られた 8 つのモデルは，それぞれ L1-50k-bi, L1-100k-bi, ⋯, L1-400k-bi と表記する．

3.2.3 ベースライン

ベースラインモデルとして，L1 モノリンガルモデル（L1-mono）と L1&L2 バイリンガルモデル（L1-0k-bi）を用意する．それぞれ L1, L1&L2 で ph1 のみを学習したモデルであり，ここでは最も学習ステップ数の多いモデル（L1-400k-bi）と合計学習ステップ数を一致させるため，ph1it2 を 1,050k ステップ（400k + 650k）学習する．

3.3 評価

音声弁別の ABX テストを用いて言語獲得の度合いを評価する．ABX テストは，A, B, X の 3 つの音声を提示し，X が A と B のどちらに近いかを判定するテストである．A, B, X は全て長さ 3 の音素列の音声であり，A と B の音素列は中央の音素のみが異なる．X の音素列は A, B のいずれかと一致する（例：(A, B, X) = (dig, dog, dig)），X は A, B とは異なる話者による音声である．以降，A, B のうち X と一致するものを target, 一致しないものを other と表記する．本実験では，HuBERT から target, other, X の特徴量を抽出し，以下で定義する Δ [13] をまず計算する．

$$\Delta = DTW(M_{other}, M_X) - DTW(M_{target}, M_X)$$

ここで，DTW は cosine 類似度を用いて動的時間伸縮法により距離を計算する関数，M_x は HuBERT により抽出した x の特徴量を表す．つまり，Δ は正解と不正解の選択肢をどの程度区別できたかを表しており，正の場合は正解，負の場合は不正解となる．

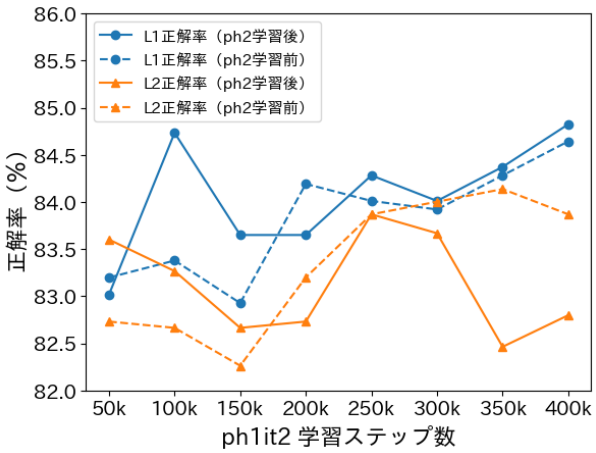

図 3 ph1it2（L1 学習）のステップ数を変化させた各モデルにおける，ph2（L1&L2 同時学習）前後の L1, L2 の ABX テストでの正解率．

本実験では，評価指標として Δ と正解率の 2 つを採用する．特徴量の抽出には，12 層の Transformer のうち，L1-mono の ABX テストにおいて L1 の正解率と Δ が学習の進行に伴い概ね上昇していた 2 層目を使用する．テストデータとしては，Perceptimatic [25] で提供されている ABX テストの中から，英語およびフランス語のデータを含む Zero Resource Speech Challenge 2017（ZeroSpeech）を用いる．

4 結果と考察

4.1 Critical Period（CP）

図 3 に，ph1it2 のステップ数を変化させた各モデルにおける，ph2 前後の L1 と L2 の ABX 正解率を示す．CP 仮説が成立する場合，L2 の学習開始時期が遅い（ph1 の学習ステップ数が多い）ほど，ph2 学習後の L2 正解率が低くなるはずである．しかし HuBERT では，早期に L2 の学習を始めるとある程度 ph2 学習後の L2 正解率が高くなるものの，正解率のピークは ph1 を 250k ステップ学習したあたりに存在するため，CP 仮説は成立していない．一方，ph2 学習による改善幅を見ると，ph1 を 100k, 150k 学習したモデルでは L1，L2 ともに改善が見られるのに対し，300k 以上学習したモデルでは L1 の精度は少し改善するが，L2 の精度は劣化している．よって，L2 の学習開始時期が早いモデルは遅いモデルに比べて L2 に対する可塑性があると考えられる．ph1 で 50k 学習したモデルでは L1 の正解率が改善していないが，この原因は今後分析予定である．

ただ，本実験では正解率の変動が少なく，また

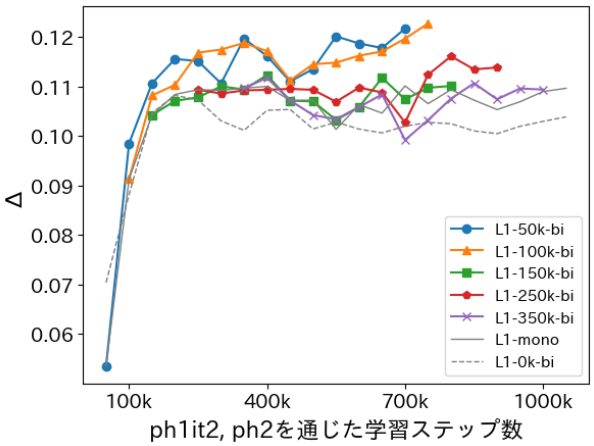

図 4　ph1it2（L1 学習）のステップ数を変化させた各モデルの，L2 の ABX テストにおける ph1it2, ph2 を通じた Δ の推移．L1-mono, L1-0k-bi はベースライン．

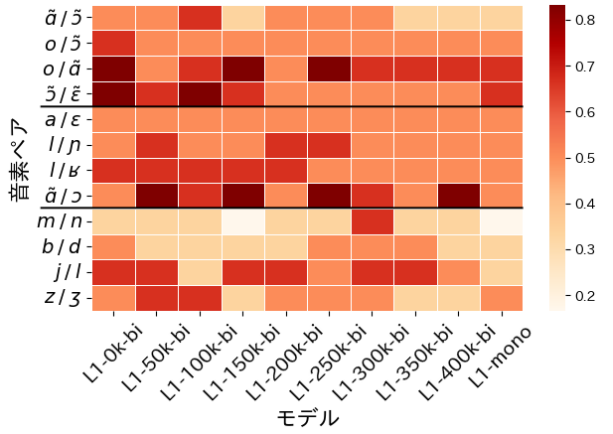

図 5　L2 の音素ペアごとの各モデルの正解率．各音素ペアについて中心の音素のみを記載．上 4 行が L2 特有の音素ペア，中央 4 行が L2 特有の音素と L1 と L2 に共通の音素からなるペア，下 4 行が L1 と L2 に共通の音素ペア．

ph2 学習前の，L2 に全く触れていないモデルが，高い L2 精度を達成している．原因として，ABX テストが HuBERT にとって簡単であるか，英語とフランス語の類似度が高い可能性が考えられる．これらの原因を取り除くことが今後の課題である．

4.2　L2 における Δ 推移

図 4 に，ph1 のステップ数を変化させた各モデルの，ph1it2, ph2 を通じた L2 の Δ 推移を示す．まず L1-50k-bi, L1-100k-bi では，Δ が L1-mono, L1-0k-bi と比べて大きく，また L2 の学習に伴い増加する傾向にある．一方で，L1-150k-bi, L1-250k-bi, L1-350k-bi では L2 を学習しても Δ が殆ど増加せず，最終的な値は L1-mono と同程度である．つまり，L2 学習を早くに開始したモデルは，遅くに開始したモデルよりも全体的に L2 音声の弁別性能が優れていることが分かる．この結果は 4.1 節で述べた結果とも一致する．一方，Δ は埋め込みの距離の差ともみなせるため，L2 学習を早くに開始したモデルは，遅くに開始したモデルとは異なる L2 埋め込みを獲得しており，またそれは L1 モノリンガルモデルや初めから 2 言語で学習したモデルとも異なることが示された．

4.3　音素ペアごとの分析

図 5 に，L2 の各音素ペアの正解率を各モデルについて示す．全体の音素ペア数は，L2 特有の音素ペアが 26，L2 特有の音素と L1 と L2 に共通の音素からなるペアが 84，L1 と L2 に共通の音素ペアが 139 であった．音素ペアは各分類ごとに L1 モノリンガルモデルの正解率が低いものを抜粋しており，正解率は各ペアで 12 サンプルの正解率の平均である．

まず L2 特有の音素ペアでは，早期に L2 学習を開始すると精度が向上するペアも存在する一方で，最初から 2 言語で学習しなければ L1-mono から精度が改善しないペアも存在する．また，L2 を学習したにもかかわらず L1-mono よりも正解率が低下するモデルも多く，この原因については今後分析予定である．次に L2 特有の音素と L1 と L2 に共通の音素からなるペアでは，遅くに L2 の学習を開始したモデルの正解率は L1-mono と同程度だが，早期に L2 学習を開始すると精度が改善する傾向にある．また，L1-0k-bi が L1-mono から改善しないペアが多いことも見てとれる．最後に L1 と L2 に共通の音素ペアでは，L1-50k-bi, L1-250k-bi〜L1-350k-bi の正解率が L1-mono から改善する傾向にある．よって，L1 と共通の音素ペアは，早期に L2 学習を開始せずとも弁別が可能となることが分かる．

5　おわりに

本研究では，HuBERT の L2 獲得過程を Critical Period（CP）仮説の観点から分析した．L2 の学習開始時期を変化させて HuBERT の学習を行った結果，本実験設定では CP 仮説は成立しないことが確認されたが，早期に L2 学習を開始したモデルには L2 に対する可塑性が見られた．また，L2 の学習開始時期が早いモデルは，遅いモデルや L1 モノリンガルモデル，初めから 2 言語で学習したモデルとは異なる埋め込みを獲得していることが示唆された．今後は，他のタスクでの検証や多言語への拡張を行う予定である．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです.

参考文献

[1] Lotte Weerts, Stuart Rosen, Claudia Clopath, and Dan F. M. Goodman. The psychometrics of automatic speech recognition. **bioRxiv**, 2022.

[2] Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. **Proceedings of the National Academy of Sciences**, Vol. 118, No. 7, p. e2001844118, 2021.

[3] Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating critical period effects in language acquisition through neural language models, 2024.

[4] Wilder Penfield and Lamar Roberts. **Speech and Brain Mechanisms**. Princeton University Press, Princeton, 1959.

[5] Eric H. Lenneberg. The biological foundations of language. **Hospital Practice**, Vol. 2, No. 12, pp. 59–67, 1967.

[6] David Singleton. The critical period hypothesis: A coat of many colours. **International Review of Applied Linguistics in Language Teaching**, Vol. 43, No. 4, pp. 269–285, 2005.

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[8] Yevgen Matusevych, Herman Kamper, Thomas Schatz, Naomi Feldman, and Sharon Goldwater. A phonetic model of non-native spoken word processing. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1480–1490, Online, April 2021. Association for Computational Linguistics.

[9] Iuliia Zaitova, Badr Abdullah, and Dietrich Klakow. Mapping phonology to semantics: A computational model of cross-lingual spoken-word recognition. In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri, editors, **Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects**, pp. 54–63, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[11] Marvin Lavechin, Maureen de Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. Modeling early phonetic acquisition from child-centered audio data. **Cognition**, Vol. 245, p. 105734, 2024.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 12449–12460. Curran Associates, Inc., 2020.

[13] Juliette Millet and Ewan Dunbar. Do self-supervised speech models develop human-like perception biases? In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7591–7605, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[16] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[17] Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. Second language acquisition of neural language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13557–13572, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[19] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 07 2020.

[20] Brian MacWhinney. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. **Computational Linguistics**, Vol. 26, No. 4, pp. 657–657, 12 2000.

[21] Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. In **INTERSPEECH 2023**, pp. 4588–4592, 2023.

[22] Benjamin Börschinger, Mark Johnson, and Katherine Demuth. A joint model of word segmentation and phonological variation for English word-final /t/-deletion. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1508–1516, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[23] KATHERINE DEMUTH and ANNIE TREMBLAY. Prosodically-conditioned variability in children' s production of french determiners. **Journal of Child Language**, Vol. 35, No. 1, p. 99–127, 2008.

[24] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] Juliette Millet and Ewan Dunbar. Perceptimatic: A human speech perception benchmark for unsupervised subword modelling. In **Interspeech 2020**, pp. 4881–4885, 2020.