音声トークナイズが音声言語モデルの性能に与える影響の調査

神藤駿介 [1]　宮尾祐介 [1,2]　高道慎之介 [3,1]
[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　[3] 慶應義塾大学
{skando,yusuke}@is.s.u-tokyo.ac.jp　shinnosuke_takamichi@keio.jp

概要

近年，音声言語処理の新たな手法として Textless NLP（テキスト資源に頼らず音声資源のみを用いて NLP システムを構築する試み）が注目を集めている．その基盤となる「音声言語モデル」は，音声シグナルを「トークナイズ」して離散表現に変換した上で言語モデルを学習し，音声言語処理の実現を目指すものである．本研究では，計 36 通りの音声トークナイズ手法を用いて音声言語モデルを学習して性能を比較し，音声トークナイズが音声言語モデルの性能に与える影響を調査した．実験の結果，音声シグナルを適切な粒度で分節すること，離散化時のクラスタ数を増やすことの有用性が示唆された．

1　はじめに

近年，音声言語処理のための新たな手法として Textless NLP[1)] という試みがある [1]．その目的は，テキスト資源に頼らず音声資源のみを用いてモデルを構築し音声言語処理タスクを実現することであり，対話 [2] や翻訳 [3]，構文解析 [4] などの研究がなされてきた．Textless NLP は，テキスト資源の少ない言語のモデリングを促進したり，音声に含まれるニュアンスや感情の情報を効果的に利用できるなど，様々な利点が期待できる．また，乳児は音声のみから言語に関する多くの知識を獲得できることを考えれば，Textless NLP は認知言語学や発達心理学への寄与もあり得る研究であると言える [5]．

本研究はその中でも**音声言語モデル**を研究対象とする．音声言語モデルとは，音声シグナルの連続表現の系列を，有限の語彙集合を持つ離散表現 (**Discrete Acoustic Unit; DAU**) の系列へと変換し，それを用いて言語モデルを学習するものとまとめられる（図 1）．本論文では音声の連続表現から DAU への系列変換を**音声トークナイズ**と呼ぶ．これまでも様々な音声トークナイズ手法が提案されているが

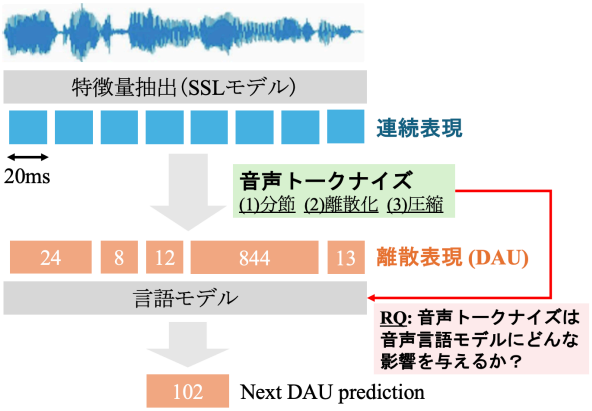

図 1　音声言語モデルの概略と本研究のリサーチクエスチョン．音声シグナルから連続表現を抽出し，それを「トークナイズ」することで音声言語モデルを学習する．

[6, 7, 8]，概してベースラインに対する性能比較に留まっており，各手法がどのような形で音声言語モデルの性能に寄与するかは明らかではない．

本研究はこの問いに答えるため，音声トークナイズ手法を 3 つのステップに分解して整理し，各ステップの設定を変えながら同一条件下で音声言語モデルを学習させ性能を比較する．まず，音声トークナイズを **(1) 分節 (2) 離散化 (3) 圧縮**の 3 つのステップに整理する．例えば (1) 分節とは，20ms 単位の表現となっている音声連続表現を音節などの単位に分節してから離散化するもので，DAU の系列長を削減する効果がある [6, 9, 10]．本研究では音素単位・単語単位の分節を適用して実験を行った．最終的に合わせて 36 通りのトークナイズ手法を適用して音声言語モデルを学習させ，音声言語の認識能力を測定する 5 つのベンチマークで評価を行った．

実験の結果，音素単位の分節は性能を向上させる一方で単語単位の分節は性能を劣化させること，離散化時のクラスタ数は多い方が良いことなどが示唆された．また，教師なし分割された音素や単語の分節単位ではなく，固定長の分節でも性能が劣化しないことが観察された．今後は特に分節設定に対する深い検証が必要である．

1)　https://speechbot.github.io/

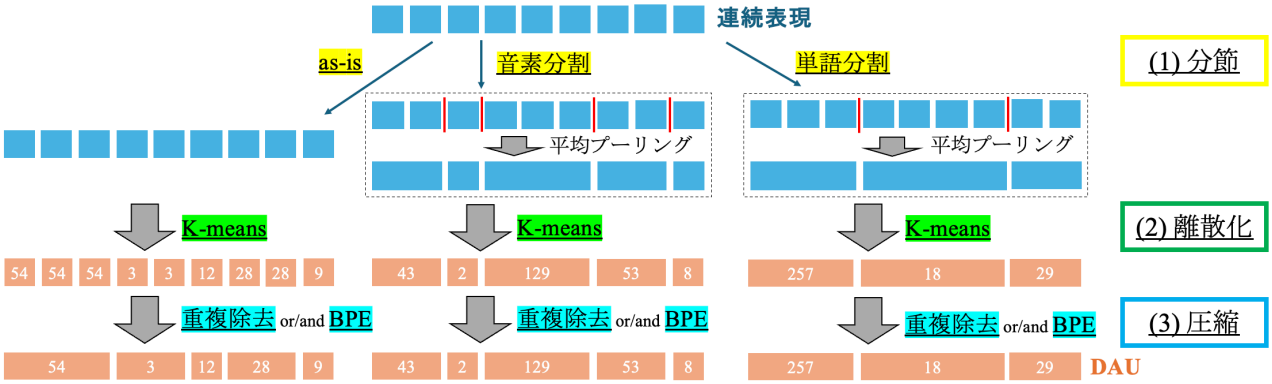

図 2　本研究で検証対象とする音声トークナイズ手法.

2　検証対象の音声トークナイズ手法

図 1 に示すように，音声トークナイズの目的は音声の連続表現の系列 $\boldsymbol{x} = x_1, x_2, \ldots, x_T$ を DAU の系列 $\boldsymbol{z} = z_1, z_2, \ldots, z_{T'}$ へと変換することである．音声言語モデルの多くの先行研究に則り，本研究では連続表現の抽出に代表的な音声自己教師あり学習モデル (SSL モデル) である HuBERT [11] を用いる．

HuBERT が出力する連続表現の解像度は 50Hz，すなわち，1 秒あたり 50 個の連続表現が存在する．ただし，各 x_i をそのまま離散表現に変換すると系列長が長くなり過ぎてしまい，言語モデルの学習が難しくなる．そのため，系列長を削減しながらトークナイズする手法が数多く提案されてきた [1, 6, 8]．加えて，離散表現のクラスタ数についても変更の余地がある．本節ではこれらの手法を図 2 に示す 3 つのステップ: **(1) 分節 (2) 離散化 (3) 圧縮** に整理し，検証対象とする音声トークナイズ手法を説明する．

2.1　(1) 分節

$\boldsymbol{x}$ の系列長を短くする手段として，何らかの分節に従って連続表現をまとめることが考えられる [6, 9, 10]．本研究では以下の三つの分節設定を試す．

- **as-is**: 分節を追加せずそのまま使う
- **音素分割**: 音素単位で分節し，平均プーリング
- **単語分割**: 単語単位で分節し，平均プーリング

音節単位で分割する手法も提案されているが [9, 10]，再現出来なかったため今後の課題とする．

2.2　(2) 離散化

離散化においては，音声言語モデルのベースラインである GSLM [1] をはじめとする多数の先行研究に則り K-means クラスタリングを行う．ここでは**クラスタ数** K を自由に定めることができるため，複数のクラスタ数（2 の冪乗）で検証を行う．

表 1　検証対象の音声トークナイズ手法と訓練トークン数．(1) は分節単位，(2) は K-means のクラスタ数，(3) は圧縮手法を示す．BPE の語彙数は $2^{14} = 16384$ とする．トークン数は各設定における最小値と最大値を示す．

設定	(1)	(2)	(3)	トークン数
(A)	as-is	2^7–2^{14}	重複除去	4.9B–7.1B
(B)	音素分割	2^7–2^{14}	重複除去	2.1B–2.2B
(C)	単語分割	2^7–2^{14}	重複除去	610M–630M
(D)	as-is	2^6–2^9	BPE	2.0B–2.9B
(E)	as-is	2^6–2^9	重複除去 & BPE	1.2B–1.9B

2.3　(3) 圧縮

ステップ (2) で得られる離散表現の系列をさらに圧縮することで，最終的な DAU 系列を得る．**重複除去**は，連続する離散表現を一つにまとめる処理を指す（図 2 に例を示す）．GSLM をはじめとする多数の先行研究で用いられているが，各 DAU の長さの情報が落ちるという欠点がある．これに対し **Acoustic BPE** [8] は，重複除去をせずに DAU 系列に対して Byte-pair Encoding を行うもので，DAU の長さの情報を保持しながら圧縮することができる．

検証対象のトークナイズ手法　検証対象の各設定を表 1 に示す．例えば，(A) 内の性能を比較することで as-is 設定におけるクラスタ数の影響を調査でき，(A) と (B) の性能を比較することで音素分割の影響を調査できる．BPE の設定の根拠は 4 節で後述する．また，表 1 には各設定における訓練トークン数も示している．K-means のクラスタ数に応じて圧縮によるトークン数の減少度合いが変化するため[2]，最小値と最大値を示している．

2)　重複除去について言えば，クラスタ数が少ないほど同じ離散表現が連続する可能性が高くなるため，最終的なトークン数が少なくなる．

表 2　評価用ベンチマークのタスクの例.

	ペアの例
sWUGGY	(✓ brick, ✗ blick)
sBLIMP	(✓ Dogs eat meat, ✗ Dogs eats meat)
prosaudit	✓ She went to jail [PAUSE] for murder. ✗ She went to jail for [PAUSE] murder.
sStoryCloze (sSC)	Ana was tanning on the beach. She dozed off in the warm sun. She woke three hours later. Her eyes widened as she looked in the mirror. ✓ Ana was extremely sunburnt. ✗ Ana was extremely pale.
tStoryCloze (tSC)	（最初の 4 文は sSC と同じ） ✓ Ana was extremely sunburnt. ✗ Michael hoped the new squirrel would fare.

3　実験設定

3.1　訓練データの構築

音声言語モデルの訓練データとして LibriLight [12]（60,000 時間の英語オーディーブック）を用いた．2 節で説明したトークナイズ手法に基づき，(1) では教師なし音素分割 [13]・教師なし単語分割 [14] を行って分節を定めた．(2) で用いる K-means モデルおよび (3) で用いる BPE モデルは共に LibriSpeech [15] の 100 時間のサブセットで訓練して構築した．訓練速度を向上させるため，訓練データを全て連結した上でトークン数を 2048 に切り揃える処理を施した．

3.2　モデルの訓練

言語モデルは OPT [16] を使用し，ハイパーパラメータは HuggingFace のデフォルト値を使用する．各設定のモデルをそれぞれ 1 epoch 訓練し，最後のチェックポイントを読み込んで評価を行う．

3.3　評価指標

音声言語モデルの言語理解能力を測定するため，表 2 に示す 5 つのベンチマークを用いる．いずれのタスクも共通して，音声のペアを与えて正しい音声により高い尤度を割り当てた場合に正解とみなす．

sWUGGY [17]　モデルが語彙の知識を持っているかを検証するベンチマークである．実在する単語と，それを一文字変えた単語のペアからなる．

sBLIMP [17]　モデルが文法的な知識を持っているかを検証するベンチマーク: BLIMP [18] を音声に変更したものである．文法的な文と非文のミニマルペアからなる．

prosaudit [19]　モデルが統語構造を把握しているかを検証するベンチマークである．統語構造の切れ目にポーズを入れた音声と，統語構造の内部にポーズを入れた音声のペアからなる．

sSC,tSC [20]　モデルが常識知識を持つかを検証するベンチマーク: StoryCloze (SC, [21]) を音声に変更したものである．5 つの文からなる物語において，最後の一文を書き換えて不正解の物語を作成する．Spoken SC (sSC) はオリジナルの SC を音声に変えたものである．Topic SC (tSC) は，最後の一文をデータセットからランダムに抽出して作成したものであり，より易しいタスクである．

4　実験結果と議論

5 つのベンチマークのうち，sBLIMP と sSC についてはいずれのモデルもチャンスレートを僅かに上回る結果となった[3)]．そのため，両者の結果は付録 A に譲り，残りの 3 つのベンチマークの結果を示す．

結果を図 3 に示す．参考のため，音声言語モデルの主要なベースラインである GSLM [1] の結果を灰色の点線で示している．これは表 1 における (A) の設定（グラフの青線 •）と同様の設定であり，クラスタ数 2^7–2^9 のモデルで同等の性能を再現できている．以下，各音声トークナイズ手法の比較を行う．

4.1　(1) 分節の影響

図 3 の上 3 つのグラフより，分節設定ごとの性能は高い順に音素，as-is，単語となっていることが分かる．特に，単語分節においてはいずれのモデル・評価指標においてもチャンスレートに近い結果となっており，粗すぎる分節は性能を損なうことが示唆された．音素のような適切な粒度で分節することが性能の向上に寄与すると考えられる．

4.2　(2) K-means クラスタ数の影響

図 3 の上 3 つのグラフより，クラスタ数が性能に与える影響はそれほど一貫していないことが分かる．prosaudit においてはクラスタ数が上がるにつれて性能も上昇傾向にある一方，tSC の as-is 設定だけはクラスタ数が増えるごとに性能が減少している．全体として，音素に分節した上でクラスタ数を増やした場合に全てのタスクで最良のスコアを記録している．ここでは，クラスタ数が多いほど性能向上に寄与すると結論づけることとする．

3)　両タスクの難しさは多数の研究で報告されている [1, 9, 20]

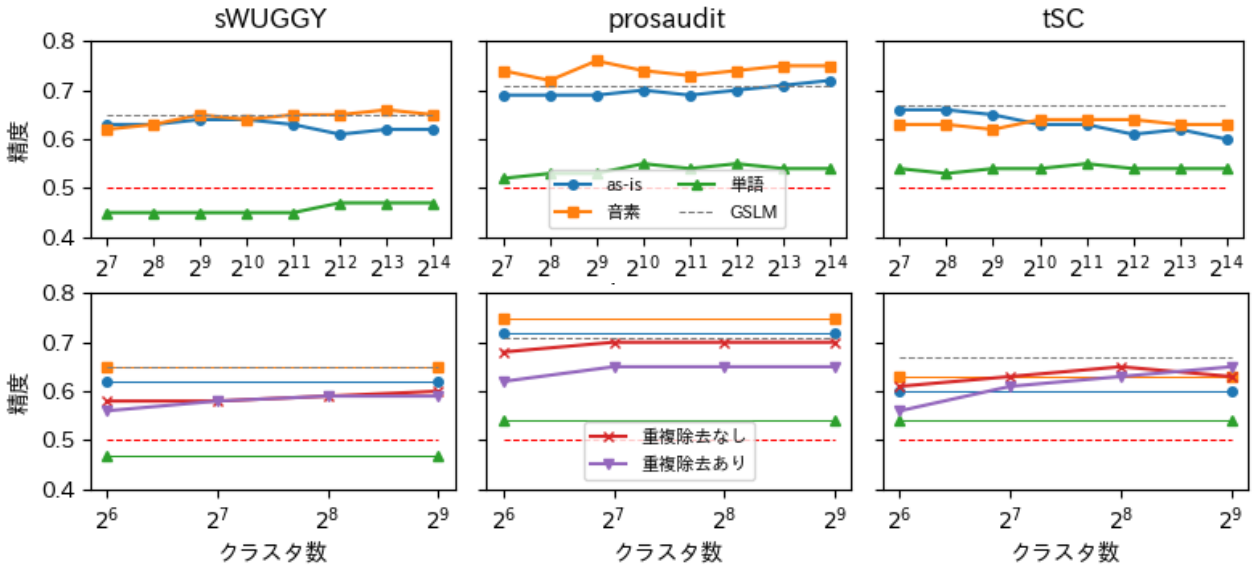

図 3　実験結果．上 3 つのグラフの各折れ線は表 1 の設定 (A)(B)(C) の結果を表す．下 3 つのグラフの各折れ線は表 1 の設定 (D)(E) の結果を表し，青 (●)・オレンジ (■)・緑 (▲) の直線は同一語彙数 (2^{14} = 16384) における as-is・音素分節・単語分節の結果を表す．赤い点線はチャンスレート (0.5) を表す．

表 3　固定分節 (fixed)・ランダム分節 (sampled) 設定の結果．(B)(C) 設定の結果はクラスタ数 2^{14} のものを抜粋．音素・単語分節設定それぞれの最高スコアを太字で示した．

設定	sWUGGY ↑	prosaudit ↑	tSC ↑
音素-表 1 (B)	**0.65**	0.75	**0.63**
音素-fixed	**0.65**	**0.76**	0.62
音素-sampled	0.59	0.61	0.56
単語-表 1 (C)	0.47	0.54	0.54
単語-fixed	0.46	**0.63**	**0.55**
単語-sampled	**0.50**	0.54	0.50

4.3　(3) 圧縮の影響

圧縮による影響を調査するため，重複除去を行わず・行ってから Acoustic BPE をかける設定（表 1 の (D)(E)）で実験を行った．4.2 節の議論から語彙数は大きく設定することとし，2^{14} = 16384 に固定した．

図 3 の下 3 つのグラフより，重複除去の有無に関わらず，BPE をかけたモデルの性能は sWUGGY と prosaudit においては音素分節設定のモデル（オレンジ線 ■）を下回る結果となった．tSC の性能も僅かに上回る程度である．表 1 より，音素分節のモデル (B) と重複除去を行わずに BPE をかけるモデル (D) は比較的トークン数が近いことを考えると，BPE は性能向上には寄与しづらいと考えられる．また，重複除去の有無で比較すると，重複除去なしの方が全体として性能が良いことが分かる．重複除去を行わないことで各 DAU の長さの情報が保持され，これがモデルの性能向上に寄与していると考えられる．

4.4　分析: 固定分節・ランダム分節の影響

ここまでの議論で，分節が音声言語モデルの性能に与える影響が大きいことが分かったが，教師なし分割にかかる計算リソース・時間は大きく無視できるものではない．そこで，より軽量な分節手法として「固定分節（分節の平均長で分節[4]）」と「ランダム分節（分節の分布からサンプリング）」を適用して実験を行った（クラスタ数は 2^{14} とした）．結果を表 3 に示す．ランダム分節は精度を損なう一方，固定分節は精度を損なわないどころか，単語分節設定では prosaudit の性能が特に良く，実は固定分節によるトークナイズで十分である可能性が示唆された．

5　おわりに

本研究では，音声トークナイズ手法が音声言語モデルの性能に与える影響を調査した．中でも分節の影響が大きいことが示唆され，今後は音節などの他の単位，並びに固定分節における深い検証が必要である．また，本研究では音声言語の認識タスクに特化して評価を行ったが，どれだけ流暢かつ意味的に一貫した音声を合成できるかも重要であり，今後は音声合成の性能に与える影響の調査も必須である．音声資源のみを用いて音声言語処理の実現を目指す Textless NLP は野心的ではあるが，本研究で培った知見を足がかりに更なる発展が期待される．

4)　音素・単語分節の平均長はそれぞれ 81ms,299ms であった．

謝辞

本研究は，JST ACT-X (JPMJAX24C9) および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ある．

参考文献

[1] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. **TACL**, Vol. 9, pp. 1336–1354, 2021.

[2] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. **TACL**, Vol. 11, pp. 250–266, 2023.

[3] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In **NAACL-HLT**, pp. 860–872, July 2022.

[4] Shunsuke Kando, Yusuke Miyao, Jason Naradowsky, and Shinnosuke Takamichi. Textless dependency parsing by labeled sequence prediction. In **Interspeech 2024**, pp. 1340–1344, 2024.

[5] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. **Cognition**, Vol. 173, pp. 43–59.

[6] Robin Algayres, Yossi Adi, Tu Nguyen, Jade Copet, Gabriel Synnaeve, Benoît Sagot, and Emmanuel Dupoux. Generative Spoken Language Model based on continuous word-sized audio tokens. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **EMNLP**, pp. 3008–3028, February 2023.

[7] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A Language Modeling Approach to Audio Generation, July 2023.

[8] Feiyu Shen, Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. Acoustic BPE for Speech Generation with Discrete Tokens, January 2024.

[9] Alan Baade, Puyuan Peng, and David Harwath. SyllableLM: Learning Coarse Semantic Units for Speech Language Models. https://arxiv.org/abs/2410.04029v1, October 2024.

[10] Cheol Jun Cho, Nicholas Lee, Akshat Gupta, Dhruv Agarwal, Ethan Chen, Alan W. Black, and Gopala K. Anumanchipalli. Sylber: Syllabic Embedding Representation of Speech from Raw Audio, October 2024.

[11] 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, October 2021.

[12] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In **2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7669–7673, 2020.

[13] Felix Kreuk, Joseph Keshet, and Yossi Adi. Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. In **Interspeech 2020**, pp. 3700–3704. ISCA, October 2020.

[14] Tzeviya Sylvia Fuchs and Yedid Hoshen. Unsupervised Word Segmentation Using Temporal Gradient Pseudo-Labels. In **2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1–5, June 2023.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In **2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 5206–5210, 2015.

[16] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[17] Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling, December 2020.

[18] 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. **TACL**, Vol. 8, pp. 377–392, 2020.

[19] Maureen De Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. ProsAudit, a prosodic benchmark for self-supervised speech models. In **INTERSPEECH 2023**, pp. 2963–2967. ISCA, August 2023.

[20] Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually Pretrained Speech Language Models. In **NeurIPS**, 2023.

[21] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **NAACL-HLT**, pp. 839–849, June 2016.

A　sBLIMP, sSC の結果

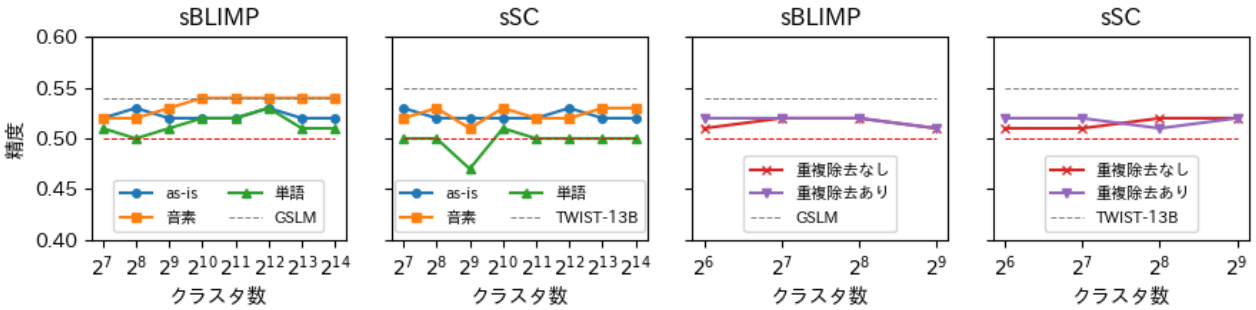

図 4　左 2 つのグラフの各折れ線は表 1 の設定 (A)(B)(C) の結果を表す．右 2 つのグラフの各折れ線は表 1 の設定 (D)(E) の結果を表す．ベースラインとして，sBLIMP は GSLM [1] の数値を，sSC は TWIST-13B [20] の数値を灰色の点線で示している．赤い点線はチャンスレート (0.5) を表す．

表 4　固定分節 (fixed)・ランダム分節 (sampled) 設定の結果．(B)(C) 設定の結果はクラスタ数 2^{14} のものを抜粋．

設定	sBLIMP ↑	sSC ↑
音素-表 1 (B)	0.54	0.53
音素-fixed	0.53	0.53
音素-sampled	0.51	0.52
単語-表 1 (C)	0.51	0.50
単語-fixed	0.51	0.51
単語-sampled	0.50	0.53

図 4 および表 4 に，sBLIMP と sSC の結果を示す．いずれの設定においてもチャンスレートに近い精度となっている．特に TWIST-13B は 13B パラメータを持つ大規模なモデルでありながら sSC の精度は 0.55 に留まっており，両タスクの難しさが伺える．