日本語方言の音声認識における学習データの規模と特性

松﨑孝介 [1]　谷口雅弥 [2,1]　坂口慶祐 [1,2]
[1] 東北大学　[2] 理化学研究所
matsuzaki.kosuke.r7@dc.tohoku.ac.jp
masaya.taniguchi@riken.jp　keisuke.sakaguchi@tohoku.ac.jp

概要

本研究は，方言音声認識において低資源のモデルが汎用的なモデルの水準に達するために必要なデータ規模の指針を提示する．そのため，日本語諸方言コーパス（COJADS）のみを用いて学習した音声認識モデルを対象に，学習データ量と性能の関係を文字誤り率（CER）に基づいて定量的に評価し，共通語中心の大規模モデル Whisper と比較した．学習データ量を段階的に変化させた実験の結果，COJADS のような実環境の方言音声において Whisper の CER 中央値（0.578）に達するには，約 35 時間の学習データが必要であることがわかった．また，方言データのみで学習したモデルは，Whisper と比較して都道府県間の性能差を有意に縮小した．

1　はじめに

音声認識（ASR）は Whisper [1] をはじめとする大規模モデルにより性能が向上したが，事前学習に依存できない低資源言語や方言では認識性能が低い．

日本語の諸方言は，語彙・音韻・文法にわたる体系的な差異と地理的多様性を有する．2022 年の『日本語諸方言コーパス』（COJADS）[2] の公開により国内 66 地点の統一的な評価が可能となったが，地点ごとのデータ規模は依然として限定的である．

近年は，大規模な事前学習済みモデルを方言データへ適応させるファインチューニングが主流である．しかし，これらのモデルは事前学習データ由来の共通語のバイアスを内包しており，共通語との言語的な距離によって性能差が生じる課題がある．加えて，低資源環境における学習特性や，性能に寄与するデータ規模についての知見が不十分である．特に，共通語のバイアスを排除した条件下では，データ量のみならず，音声のノイズや発話長分布，発話形式といった質的な側面が認識性能に与える影響についても，十分に検討されていない．

本研究では，事前学習モデルの言語的バイアスを排除した条件下で，方言における音声認識の学習特性を明らかにする．具体的には，CTC に基づく RNN モデルを方言データのみで学習する（スクラッチ学習する）ことで，データ規模と認識性能の関係を定量化する．さらに，Whisper との性能比較や共通語コーパスを用いた対照実験を通して，低資源言語における音声認識システム構築に向けたデータ収集の目安となる知見を提供する．

2　関連研究

2.1　低資源言語の音声認識

低資源言語では，ラベルなしデータを活用する自己教師あり学習フレームワーク wav2vec 2.0 [3] の適用 [4–6] や，その事前学習の強化 [7] が進められてきた．また，リソース補完のための半教師あり学習 [8,9]，音声合成による擬似データの生成 [9]，正書法が未確立な言語に対する国際音声記号（IPA）[10] を介した認識手法 [11] なども提案されている．こうした低資源環境に向けた諸提案は，地理的変異の大きい日本語方言でも重要な基盤となる．

2.2　日本語方言の音声認識と課題

日本語方言では COJADS の登場後，XLSR を用いた方言識別とのマルチタスク学習 [12]，Whisper や XLSR など複数モデルの適応手法の比較 [13]，多段階のファインチューニング [14] など，事前学習モデルを基盤とした認識性能の向上が図られてきた．

これらの研究は事前学習に依拠しているが，低資源環境の特性解明には，対象データのみによるスクラッチ学習を通じて，データの質・量と性能との関係を評価する必要がある．その際，モデルの複雑さや言語的補完の影響を抑え，音響的な対応を直接評価できる標準的な構成による検討が，後続研究の評価基盤として重要となる．

3　実験設定

3.1　タスクの定義と学習アプローチ

本研究は，入力された方言音声に対して，発話内容に忠実なひらがな書き起こしを出力する音声認識タスクを対象とする．そして，大規模事前学習モデルに由来する言語的バイアスを排除するため，方言データのみを用いてモデルをスクラッチ学習した．

3.2　学習アルゴリズムとモデル構成

音声認識の学習アルゴリズムとして，Connectionist Temporal Classification（CTC）[15] を用いた．CTC は，音声と文字の厳密なアライメントを必要とせず，音声全体と文字列の対応関係から直接学習が可能である．一般に，大規模モデルで用いられる Attention ベースの手法では言語モデルによる補完が強く働くのに対し，CTC は音響的な対応関係に基づいて逐次的に文字を出力する．

モデル構造は 3 層の双方向 Gated Recurrent Unit（BiGRU）とし，入力サイズ（特徴量次元数）は 50，隠れ層の次元数は 512 とした．RNN 層の出力に対しては，学習の安定化を目的として層正規化（Layer Normalization）を適用した．最終層には全結合層を配置し，対数 Softmax 関数を介して各時刻における文字トークンの出力確率を算出した．

CTC 損失関数 L_{CTC} は，音声特徴量列 $\boldsymbol{x}$ に対して文字トークン列 $\boldsymbol{l}$ が出力される確率を $P(\boldsymbol{l} \mid \boldsymbol{x})$ とすると，以下のように定義される [16]．

$$L_{\mathrm{CTC}} = -\log P(\boldsymbol{l} \mid \boldsymbol{x}) \tag{1}$$

認識過程の概念図を図 1 に示す．CTC では，通常の文字トークンに加え，どの文字にも対応しないブランクトークンを導入することで，長さの異なる入力と出力の対応関係をモデル化する．

3.3　データセット

本研究では，性質の異なる以下の 2 種類のデータセットの音声と書き起こしテキストを使用した．

COJADS　『日本語諸方言コーパス』有償版 [2] は，2025 年 3 月時点で全国 66 地点・約 105 時間の発話を収録する．方言話者間の自然な対話が中心であり，音声には周囲のノイズや複数人の同時発話を含む．本研究では，前処理（§3.4）により抽出された約 87 時間の有効発話データを対象とした．

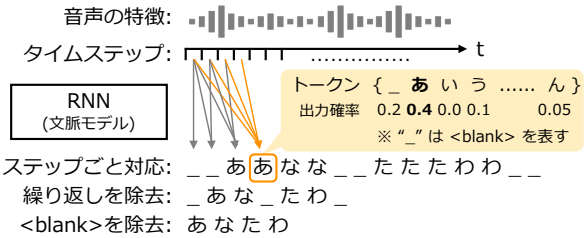

図 1　CTC による音声認識とデコーディングの過程．

Common Voice　Common Voice Corpus [17] は，不特定多数の話者による音声データセットである．文単位で，単独の共通語話者による静音環境での収録が中心である．本研究では，方言差以外の要因（収録条件等）が学習性能に与える影響を分析するための比較用データとして，この日本語版を用いた．

3.4　データセットの前処理

COJADS では，都道府県ごとに発話数比で 8:1:1（Train : Dev : Test）となるよう分割した．なお，ランダム性による影響を軽減するため，3 種類の乱数シードを用いて同様の分割・実験を行った．音声は 16 kHz のモノラル形式に統一した．Whisper (large) による音声認識も行った．表記揺れの影響を軽減するため，テキストは GPT-5 を用いてひらがなに正規化した．前処理および分割の詳細は付録 A に記す．

3.5　評価指標

音声認識の性能評価には，以下で定義される文字誤り率（Character Error Rate; CER）を用いた．

$$\mathrm{CER} = \frac{\text{置換文字数} + \text{挿入文字数} + \text{削除文字数}}{\text{正解テキストの文字数}} \tag{2}$$

なお，分子は正解テキスト・認識結果間の編集距離に等しい．以降では Test セットでの CER を示す．

4　実験結果：学習データ規模と性能

4.1　CTC モデルと Whisper の性能比較

図 2 は，Whisper (large) のゼロショット出力の CER 分布，および COJADS を用いてスクラッチ学習した CTC モデルの出力の CER 分布を，都道府県ごとに示したものである．[1)] まず中央値に着目すると，CTC モデルの CER は 40 都道府県において Whisper よりも低い値を示していることが分かる．[2)]

1)　図 2 では表示できていないが，Whisper の CER 平均値は 2 県において 2.0 を超えた（福井：3.72，徳島：2.04）．
2)　これに対し，Whisper による Common Voice 音声の認識結果は CER 中央値が 0.04 程度と低く，実用レベルであった．

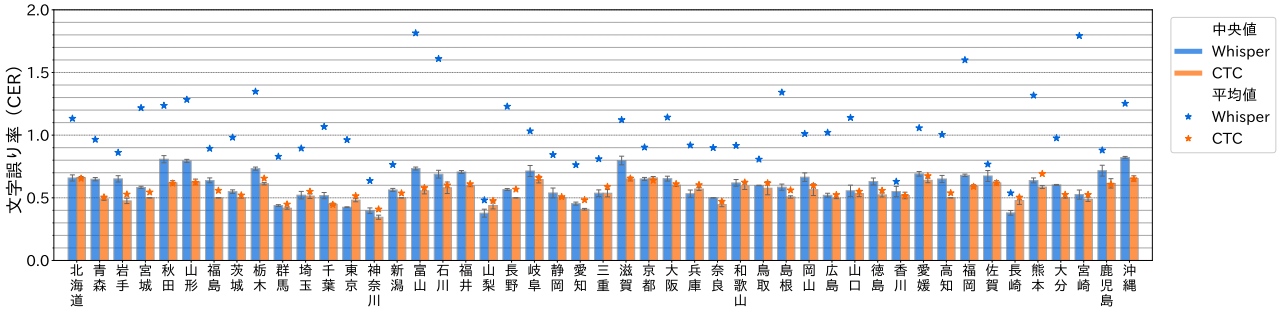

図 2　Whisper (large) のゼロショットと CTC 学習モデルの音声認識結果．エラーバーは 3 seed の標準偏差である．

表 1　Whisper および CTC 学習モデルの出力例．

特徴（収録県）	上段：正解テキスト 中段：Whisper 出力 下段：CTC 出力	CER
(a) フィラー（宮城）	んー	–
	ごしちょうありがとうございました	8.00
	んー	0.00
(b) 非共通語的（青森）	もすわだすか゜すんだら	–
	もしわたしがしんだら	0.55
	むすわだすか゜しんだら	0.18
(c) 共通語的（千葉）	すべるとあぶないがら	–
	すべるとあぶないから	0.10
	すぐのたないがら	0.50

Whisper では，多くの地域で CER の平均値が中央値の 2 倍程度となった．実際，極端に誤りの多い発話が存在し（§4.2），約 3%を占めている．一方，CTC モデルでは CER の外れ値は少ない．この差異は，§4.2 にて述べる Whisper の特性と，音響的対応に基づいて逐次的に出力を行う CTC の性質の違いを反映していると考えられる．参考として，音声認識結果の CER 分布（ヒストグラム）を付録 B に示す．

4.2　認識結果の定性的分析

表 1 に，Whisper および CTC モデルの出力例を示す．意味的な補完を行わず音声に忠実に書き起こす CTC の性質は，(c) のように不利に働く場合もあるが，方言特有の表現を含む (b) では有利に働いた．

対して Whisper には，短時間や無音の入力に対して入力と無関係な定型文を出力する誤りや，方言的な表現を共通語的な表現に置き換える補完が確認された．意味的な補完は共通語としての可読性の観点では有用な場合もあるが，本研究が意図する発話内容に忠実な記録とは乖離する．また，(a) のような誤出力は，Whisper が非発話入力に対して “Thanks for watching” などの定型的な文を頻出させるという既存知見 [18] とも整合する．Whisper にて観察された

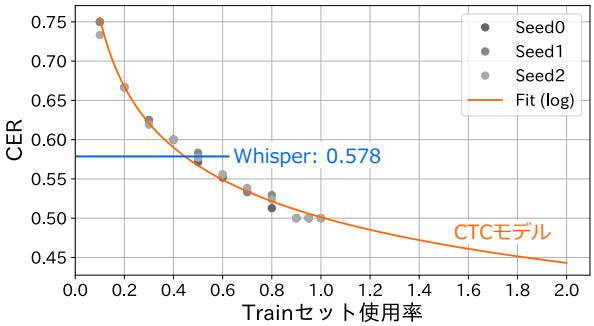

図 3　Train セット使用率と CER 中央値（COJADS）．

CER の外れ値の主因は，こうしたハルシネーションによる挿入誤りであった．編集距離の内訳（挿入・削除・置換）に基づく詳細な考察は付録 C に譲るが，外れ値は平均値を大きく押し上げるため，以降では性能比較の代表値として中央値を用いる．

4.3　都道府県ごとの性能差

図 2 より，CTC モデルは Whisper と比較して中央値が低いうえに，都道府県間のばらつきも小さい．その統計的検証として，モデル間の CER 中央値の差に対して Mann-Whitney U 検定を，都道府県ごとの中央値の分散に対して Brown-Forsythe 検定を行った．その結果，すべてで p 値は 0.05 未満となり，CTC モデルと Whisper の間には，中央値および地域差の双方において有意な差が存在することが確認された．すなわち，方言データのみでスクラッチ学習した CTC モデルは，事前学習モデル Whisper よりも地域差の小さい認識性能を実現できるといえる．

4.4　学習データ量と認識性能

図 3 は，COJADS における Train セット使用率と CER 中央値の関係を示したものである．対数の近似曲線は，データを増やした際の見通しを立てるために 2.0 まで外挿している．図 3 より，学習データ量の増加に伴い CER が単調減少する傾向がみら

表 2　さまざまな条件での認識結果．条件 A，B は Train セット使用率 1.0，0.11 にそれぞれ相当する．* 音声に付加したノイズの信号対雑音比（SNR）．† Train セット全体における 平均 ± 標準偏差．‡ 3 seed における中央値の 平均 ± 標準偏差．

条件	Train 種類	Train 件数	Train 音声長 †	SNR [dB] *	モデル	Test 種類	Test CER 中央値 ‡
A	COJADS	140k	1.79 ± 1.22	–	CTC	COJADS	0.500 ± 0.000
B	COJADS	15k	1.80 ± 1.22	–	CTC	COJADS	0.732 ± 0.005
C	COJADS	15k	4.14 ± 1.52	–	CTC	COJADS	0.617 ± 0.004
D	Common Voice	15k	4.67 ± 1.79	–	CTC	Common Voice	0.373 ± 0.003
E	Common Voice	15k	4.67 ± 1.79	+20	CTC	Common Voice	0.387 ± 0.002
F	Common Voice	15k	4.67 ± 1.79	+10	CTC	Common Voice	0.435 ± 0.006
G	Common Voice	15k	4.67 ± 1.79	0	CTC	Common Voice	0.556 ± 0.006
H	Common Voice	15k	4.67 ± 1.79	−10	CTC	Common Voice	0.762 ± 0.011
I	–	–	–	–	Whisper	COJADS	0.578 ± 0.006
J	–	–	–	–	XLSR	COJADS	0.793 ± 0.006

れた．CER 中央値を基準とすると，CTC モデルは Train セット（有効データの 80%）の使用率 0.5 付近で Whisper の性能に達した．これは本実験での COJADS の有効データ全体の約 40%（約 35 時間），発話数にして約 7 万件に相当する．

4.5　COJADS と Common Voice の性能差

Train セットの件数を同数にした対照実験において，Common Voice の音声で学習したモデルは，COJADS で学習するよりも CER が低くなった（後述の表 2 の条件 B, D）．この要因として，COJADS が会話形式で Common Voice が読み上げ形式であることによる発話の明瞭さの違いや，雑音の含み方といった収録環境の違い，1 発話あたりの音声長の違いが考えられる．これらのどの特徴が COJADS で学習と認識を難しくしているのか，§5 にて分析する．

5　分析：データ特性の性能への影響

表 2 に，さまざまな条件で学習した結果やベースラインの CER をまとめる．以下では便宜上，左列の条件記号を用いて「条件 A」のように言及する．

事前学習モデル　Whisper のほか，CTC モデルどうしでの比較として，CTC で事前学習された wav2vec2-large-xlsr-53-japanese（XLSR）による音声認識も行った（条件 I, J）．その結果，どちらも CTC モデルより CER が高かった．これは，事前学習データの性質や言語的バイアスが，COJADS 音声の認識性能に影響することを示唆する．

収録環境（ノイズ）　ノイズの影響を検証するために，Common Voice の音声にピンクノイズ[3]を付加して学習を行った（条件 E, F, G, H）．その結果，SNR が −10 dB から 0 dB のときに CER が COJADS と同程度となった．実際の COJADS の音声にはこれほど強い定常ノイズは含まれていないが，ノイズが学習難易度を上げる一因であることが示唆された．

音声長分布　COJADS の音声長の平均は Common Voice の半分以下であり，両者の音声長の分布は大きく異なる（条件 B, D）．この違いの学習への影響を検証するため，分布を可能な限り揃えたデータセットを構成[4]して学習・評価を行ったところ，CER が改善した（条件 B, C）．これは，音声長分布が学習難易度を決定する一因であることを示唆する．

評価指標の解釈　以上の分析から，COJADS における音声認識の学習難易度は，発話形式や音声長分布の特性など複数の要因が重なって規定されていることが示唆された．また，複数のモデルの比較結果から，同一の CER であっても，モデルによって誤りの生じ方や出力の性質が異なることが確認された．これは，低資源条件におけるモデル選択や性能比較では，単一の指標によらず，誤りの傾向や分析の目的を踏まえた解釈が重要であることを示している．今後は，収録条件や発話形式の異なるデータセットを用いたさらなる検証を通じて，これらの影響をより体系的に整理する必要がある．

6　おわりに

日本語諸方言コーパス（COJADS）を用いた音声認識において，学習データ規模と性能の関係を定量的に評価し，スクラッチ学習にて Whisper の性能に達するための目安となるデータ量を示した．また，データ特性の認識性能への影響を分析した．

3）　ピンクノイズは，$S(f) \propto 1/f$ と低周波成分が強い特性をもち，自然界の雑音に近いとされる．SNR（Signal-to-Noise Ratio）は $\mathrm{SNR}_{\mathrm{dB}} = 20\log(S/N)$ で定義する．20 dB は信号がノイズの 10 倍強く，0 dB は信号とノイズの強度が等しい．

4）　Common Voice と似た音声長分布となるよう元データから音声を抜粋した．調整により，分布の形状の違いを表す JSD は，Train セットにおいて 0.732 から 0.174 に縮小した．

謝辞

本研究は JSPS 科研費 JPMJCR20D2，JP25K03175，JP24K16077 の助成を受けたものです．本研究を進めるにあたり，東北 NLP グループの皆様から有益なコメントをいただきました．ここに深く感謝申し上げます．

参考文献

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 28492–28518. PMLR, 23–29 Jul 2023.

[2] 日本語諸方言コーパス（有償版）Ver.2025.03. https://www2.ninjal.ac.jp/cojads/index.html.

[3] 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.

[4] 堀元優花, 張逸群, 斎藤博昭. 深層学習に基づいた富山弁音声認識とその標準日本語への変換. 人工知能学会全国大会論文集, Vol. JSAI2024, No. 2C1-GS-7-01, 2024.

[5] Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. Development of automatic speech recognition for the documentation of Cook Islands Māori. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 3872–3882, Marseille, France, June 2022. European Language Resources Association.

[6] Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In Sarah Moeller, Antonios Anastasopoulos, Antti Arppe, Aditi Chaudhary, Atticus Harrigan, Josh Holden, Jordan Lachler, Alexis Palmer, Shruti Rijhwani, and Lane Schwartz, editors, **Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages**, pp. 41–51, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Anuroop Sriram, Michael Auli, and Alexei Baevski. Wav2Vec-Aug: Improved self-supervised training with limited data, 2022. arXiv:2206.13654.

[8] 今泉遼, 増村亮, 塩田さやか, 貴家仁志. End-to-end 方言音声認識のための方言ラベルを考慮した半教師あり学習. 情報処理学会 音声言語情報処理（SLP）, Vol. 2022-SLP-140, No. 15, 2022.

[9] Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 715–729, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] International Phonetic Association. https://www.internationalphoneticassociation.org/content/ipa-chart.

[11] Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. Universal automatic phonetic transcription into the international phonetic alphabet, 2023. arXiv:2308.03917.

[12] Shogo Miwa and Atsuhiko Kai. Dialect speech recognition modeling using corpus of japanese dialects and self-supervised learning-based model xlsr. In **Interspeech 2023**, pp. 4928–4932, 2023.

[13] Naoki Takahashi, Shogo Miwa, Yuta Kamiya, Takumi Toyama, Raufun Nahar, and Atsuhiko Kai. Comparison of large pre-trained models and adaptation methods for japanese dialects asr. In **2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)**, pp. 811–814, 2024.

[14] Yuta Kamiya, Shogo Miwa, and Atsuhiko Kai. A parameter-efficient multi-step fine-tuning of multilingual and multi-task learning model for japanese dialect speech recognition. In **2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)**, pp. 1–6, 2024.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In **Proceedings of the 23rd International Conference on Machine Learning**, ICML '06, p. 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[16] 高島遼一. Python で学ぶ音声認識 機械学習実践シリーズ. インプレス, 2021.

[17] Common Voice Corpus 日本語版 v21.0. https://commonvoice.mozilla.org/ja/datasets.

[18] Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. Investigation of whisper asr hallucinations induced by non-speech audio. In **ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1–5. IEEE, 2025.

A　データの前処理と分割の詳細

音声・テキストの整形　COJADS において，書き起こしテキストから「{笑}」などのメタ情報を除去し，音声は付随する時間情報に基づき発話単位で切り出した．特徴量抽出時のサンプル数を確保するため，0.5 秒未満の発話は除外している．音声ファイルは 16 kHz へのリサンプリングを行い，モノラルの wav 形式に統一した．Common Voice については，音声は同様のサンプリング処理を行い，書き起こしテキストは漢字かな交じりのため後述の方法でひらがなに正規化した．

データ分割　COJADS については，都道府県ごとに発話数比で 8:1:1（学習：検証：評価）となるよう分割した．分割のランダム性による影響を軽減するため，3 種類の乱数シードを用いて同様の分割および学習・評価を実施した．Common Voice については，提供元の分割（Train / Dev / Test）に従っている．

ひらがな正規化　漢字かな交じりのテキストは GPT-5（`gpt-5-2025-08-07`）を用いてひらがなに変換した．[5] 推論負荷（reasoning effort）は `low` に設定した．変換の際のシステムプロンプトを以下に示す．

```
あなたは日本語テキストをすべてひらがなに変換する専門家です。ルールは次の通りです：
1. すべての漢字とカタカナはひらがなに変換する。
2. 数字は日本語読みで書く（例：5 つ→いつつ、10.24 →じってんによん、10,000 →いちまん）。
3. 英語は日本語読みで書く（例：Hello world. →はろーわーるど。、CER →しーいーあーる）。
4. 改行を含め、記号はそのまま残す（例：「」()、。）
```

B　CER 分布

図 4，図 5 に，COJADS の Test セットにおける，CTC と Whisper の音声認識結果の CER 分布を示す．§4.2 にて言及した Whisper のハルシネーションは，図 5 に示す Whisper の CER 分布において外れ値として確認できる（図中では 3.0 以上をまとめて表示しているが，CER 最高値は 3797，次点は 308 であった）．

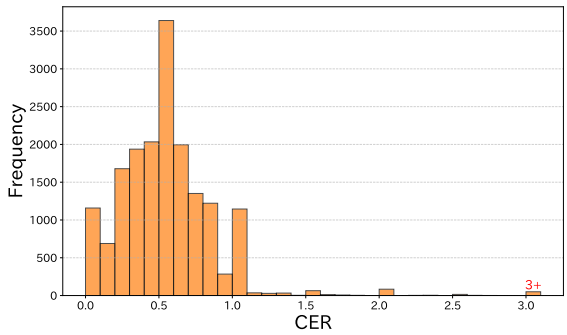

図 4　COJADS の Test セットにおける CTC の音声認識結果（表 2 の条件 A）．3 seed の平均．

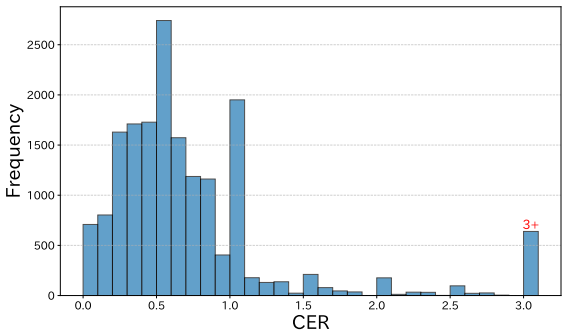

図 5　COJADS の Test セットにおける Whisper の音声認識結果（表 2 の条件 I）．3 seed の平均．

C　編集距離の内訳

表 3 に，CER 算出時に得られた編集距離の内訳（挿入・削除・置換）を示す．一般に，挿入誤りが多い場合は予測テキストが参照テキストよりも長くなり，削除誤りが多い場合は予測テキストが短くなる傾向がある．表 3 より，Whisper は挿入および置換の回数が多く，出力文字列の長さが入力発話の内容から逸脱する傾向が確認された．一方，CTC モデルでは削除および置換が中心であり，文字列長は比較的入力に忠実，あるいは短くなる傾向にある．CER は分母に参照テキストの文字数を用いるため，挿入誤りが多い場合には CER が 1 を超えるが，削除誤りが多い場合でも 1 を超えることはない．この点からも，CER 分布において観察された高 CER 側の外れ値の主因が，Whisper における挿入誤りを伴う誤出力にあることが裏付けられる．

表 3　出力テキストの編集距離の内訳 [%]（CTC は 3 回の平均）．

モデル	挿入	削除	置換
CTC	4.66	46.43	49.36
Whisper	29.81	25.90	44.29

5）　GPT-5 のほか，pykakasi（https://github.com/miurahr/pykakasi）や SudachiPy（https://github.com/WorksApplications/SudachiPy）による変換も行い，Common Voice の 100 例で目視確認したところ，GPT-5 が文脈を考慮できて最も正確であったためこれを採用した．