LLM の生成する方言テキストを音声合成したデータによる音声言語モデルの方言理解能力向上

三森 俊祐 [1]　藤田 雄介 [1]　水本 智也 [1]
[1]SB Intuitions 株式会社
{shunsuke.mitsumori,yusuke.fujita,tomoya.mizumoto}@sbintuitions.co.jp

概要

方言音声に対する音声言語モデル（SLM）の性能低下は実方言音声データ不足に起因する．方言音声を合成することでデータを増やす手法が提案されてきているが，いずれの手法でも結局，方言音声合成モデルの訓練に実方言音声データを必要とする．本研究は実方言音声を用いず，LLM で生成した方言テキストを標準語韻律で音声合成した擬似方言音声を作成し，SLM の学習に追加する．CPJD の 20 方言での BLEU スコアが，標準語翻訳で 48.89 から 52.64，英語翻訳で 24.21 から 26.24 と改善した．

1　はじめに

近年，LLM を基盤とした音声言語モデル（Speech Language Model; SLM）[1] の発展は著しい [2, 3, 4, 5]．LLM を用いたパイプラインの採用により，音声認識，音声翻訳，音声対話などのタスクで従来手法を凌駕している．しかしその一方で，方言音声に対する理解能力は，標準語音声と比較して低下する問題が存在する [6]．この標準語と方言間での技術格差は，モデルの学習に必要な各方言の実音声データが不足していることに主として起因している．

従来このようなデータ不足に対しては，低リソース言語の研究分野において，音声合成によるデータ拡張アプローチが取られてきた．主な手法として，LLM やウェブ収集により得たテキストを，対象言語の実音声で学習した音声合成モデルを用いて音声化する方法 [7, 8] が挙げられる．しかしこれらの手法を方言に適用する場合，音声合成モデルの学習に十分な実音声収集を全方言で行うことは困難であり，このデータ依存性は依然として対応方言拡大の障壁となっている．

そこで我々は，テキストから得る言語的特徴のみで SLM を方言に適応させられるという仮説に基づき，実方言音声を一切用いずに合成する擬似方言音声による SLM の方言適応を提案する．具体的には，LLM を用いて合成した方言テキストを一般的な音声合成モデルで音声合成を行い，言語的特徴は方言であるが韻律的特徴は標準語の擬似方言音声を作成する．最終的に，得られた擬似方言音声を学習データに追加して SLM を学習する．本手法は音声合成モデルの訓練に実方言音声は用いないため，実音声が皆無の方言に対してもデータ拡張が可能となる．

本研究では日本語の方言を対象に，方言音声を入力として標準語テキストへの翻訳及び英語テキストへの翻訳タスクで BLEU スコアを測ることで，提案手法の有効性を検証した．実験の結果，標準語翻訳タスクにおいて平均で BLEU スコアが 48.89 から 52.64（+3.75）に，英語翻訳タスクにおいて平均で BLEU スコアが 24.21 から 26.24（+2.01）に向上した．これにより我々は，方言実音声を一切用いずに合成した標準語韻律の擬似方言音声を用いることで，SLM の方言理解能力を改善できることを我々が知る限り初めて示した．

2　提案手法

本研究では，方言の実音声を一切用いずに合成する擬似方言音声による SLM の方言適応（図 1）を提案する．擬似方言音声の合成パイプラインは，LLM を用いた方言翻訳と標準語韻律の音声合成の 2 段階で構成される．以下に各工程の詳細を述べる．

2.1　LLM を用いた方言翻訳

まず，音声合成の入力となる方言テキストを生成するため，LLM を用いて標準語テキストを方言に翻訳する．具体的には，標準語テキストを入力とし，LLM に対してそれらを対象の方言に変換するよう指示を与える．この際，生成品質を向上させるために，別途 LLM を用いて作成した標準語から対象方

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

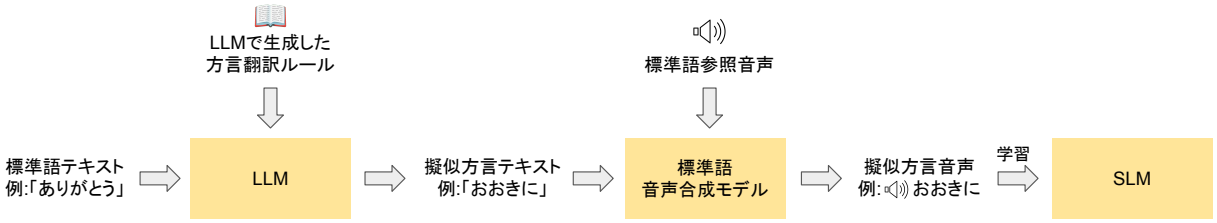

図 1　提案手法の概略図

言への文法的な翻訳規則を，翻訳指示のプロンプトとして同時に与える．[1)]

2.2　標準語韻律の音声合成

2.1 節で生成した方言テキストを音声化する．本研究では方言の実音声データを一切使用せず，標準語のリソースのみを用いた音声合成を行う．具体的には，生成された方言テキストを音声合成モデルに入力し，その際の条件付けとして標準語の音声のみを参照音声として与える．生成される音声は，テキスト内容は方言でありながら，韻律は標準語の参照音声のスタイルが反映される．つまり，方言の語彙・文法構造を持ちながらも，韻律のみが標準語である擬似方言音声が生成される．

3　実験

3.1　タスク

本研究では SLM の方言理解能力を多角的に評価するため，入力の日本語音声に対して，標準語テキストへの翻訳と英語テキストへの翻訳の 2 つのタスクを行う．

3.2　SLM のモデル構造

本研究では，音声エンコーダと LLM をプロジェクターで接続したエンドツーエンドの SLM を構築した．音声エンコーダには Whisper large-v3[9] を，LLM には日本語の指示追従能力に優れた Llama-3.1-Swallow-8B-Instruct-v0.3[10] を採用した．そして，音声エンコーダの最終層から得られる 1280 次元の音声特徴量を，LLM の 4096 次元の入力埋め込み空間へ射影するため，軽量なプロジェクターを構築した．構造としては，カーネルサイズ 4，ストライド 2 の 1 次元畳み込み層（Conv1d）を 2 層（間に GELU 活性化関数を配置）重ね，最後に線形層を通す構成とした．なお学習に際しては，音声エンコーダおよび LLM のパラメータは凍結し，プロジェクターのパラメータのみを更新した．

3.3　データセット

3.1 節で述べた 2 つのタスクを遂行するため，学習用および評価用のデータセットを以下の通り構築した．

3.3.1　学習用データセット

ReazonSpeech large v2[11]　大規模日本語音声コーパスであり，フィルタリング処理後の約 260 万発話を使用した．本コーパスに含まれる，音声に対応する書き起こしを標準語翻訳の正解ラベルとした．一方で，英語対訳は付与されていないため，Qwen2.5-32B-Instruct[12] を用いて日本語テキストから英語テキストへの翻訳を行い，これを英語翻訳タスクの正解ラベルとして使用した．なお，提案手法およびベースライン手法では，このデータセットのテキスト部分をデータ合成元としても利用する．

Speech-BSD[13]　ビジネスシーンを想定した日本語音声データセット (2 万発話) である．本データセットには，人間によって作成された高品質な書き起こしおよび英語対訳が存在するため，これらを両タスクにおいて正解ラベルとした．

CoVoST2[14]　多言語音声翻訳データの日本語部分（1,119 発話）を使用した．人手による書き起こしと英語対訳が付与されているため，これらを両タスクの正解ラベルとした．

3.3.2　評価用データセット

提案手法で作成した擬似方言音声で訓練した SLM の，実方言音声に対する方言理解能力の向上を評価するため，学習には一切使用していない実方言音声コーパスとして CPJD (Crowdsourced Parallel Speech Corpus of Japanese Dialects)[15] を使用した．

1)　翻訳規則を LLM で生成するの際のプロンプトは付録 A で述べる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1　方言音声から標準語テキストへの翻訳タスクにおける BLEU スコア

学習に追加する合成データ	Avg.	秋田	阿波	福井	福岡	広島	北海道	いわき	伊予	出雲	金沢
なし	48.89	35.17	41.70	62.61	63.67	56.32	60.47	55.15	57.74	43.07	44.71
標準語音声（ベースライン）	48.91	34.64	42.54	62.41	64.07	56.30	**60.52**	53.79	57.04	41.87	44.64
擬似方言音声（提案手法）	**52.64**	**38.80**	**46.80**	**64.22**	**67.52**	**61.33**	59.93	**58.97**	**61.55**	**47.80**	**46.34**

学習に追加する合成データ	Avg.	京言葉	宮崎	諸県	奈良	岡山	大阪	埼玉	土佐	遠州	津軽
なし	48.89	42.19	37.86	19.29	50.47	46.67	59.49	73.97	55.56	44.25	25.99
標準語音声（ベースライン）	48.91	42.88	38.21	18.25	51.26	45.47	58.94	**75.95**	56.68	44.00	25.75
擬似方言音声（提案手法）	**52.64**	**46.66**	**42.63**	**27.34**	**54.25**	**56.13**	**62.54**	73.83	**57.12**	**46.55**	**31.41**

表 2　方言音声から英語テキストへの翻訳タスクにおける BLEU スコア

学習に追加する合成データ	Avg.	秋田	阿波	福井	福岡	広島	北海道	いわき	伊予	出雲	金沢
なし	24.21	19.10	25.34	27.69	26.72	25.97	30.19	27.75	27.49	22.55	**21.68**
標準語音声（ベースライン）	25.38	19.90	26.29	29.57	28.46	26.67	30.17	**30.12**	27.58	**24.92**	20.04
擬似方言音声（提案手法）	**26.24**	**20.12**	**27.23**	**30.13**	**30.39**	**27.39**	**30.78**	30.05	**29.16**	23.28	21.45

学習に追加する合成データ	Avg.	京言葉	宮崎	諸県	奈良	岡山	大阪	埼玉	土佐	遠州	津軽
なし	24.21	24.35	21.73	10.60	26.67	24.83	28.19	28.66	25.36	23.83	14.30
標準語音声（ベースライン）	25.38	25.97	23.23	10.12	**28.81**	25.96	**29.58**	31.03	26.78	24.70	14.65
擬似方言音声（提案手法）	**26.24**	**27.14**	**24.98**	**12.41**	28.40	**28.97**	29.45	**31.98**	**28.14**	**25.22**	**16.70**

CPJD は全国 20 方言・21 名のネイティブ話者によるパラレルコーパス（各話者約 250 発話）である．全音声に対し，対応する標準語の書き起こし文が存在するため，これを標準語翻訳の正解ラベルとした．しかし CPJD には英語対訳が含まれていないため，評価用として高品質なラベルを確保する目的で，GPT-4o を用いて標準語テキストから英語への翻訳を生成し，これを正解ラベルとした．

3.4　比較手法

提案手法の有効性を検証するため，学習データの構成が異なる以下の 3 つのモデルを構築した．

データ拡張なし　3.3 節の実音声データ（Reazonspeech large v2, Speech-BSD, CoVoST2）のみを用いてモデルを学習した．合成データによる拡張を行わないモデルである．

提案手法　データ拡張なしの学習データに加え，Reazonspeech large v2 のテキストに対し，CPJD の 20 方言から各発話につきランダムに 1 方言を選択して作成した約 260 万発話の擬似方言音声を追加したデータセットを用いてモデルを学習した．方言変換ルールの作成および標準語からの方言翻訳には，gpt-oss (120B)[16] を，音声合成には，Tsukasa-Speech[17] を使用した．音声合成時に用いる標準語参照音声には，各合成毎に jvs コーパス [18] からランダムな音声を用いた．

ベースライン　データ拡張なしの学習データに加え，Reazonspeech large v2 のテキストを方言翻訳せずに合成した約 260 万発話の合成標準語音声データを追加したデータセットを用いてモデルを学習した．提案手法と同様に音声合成には Tsukasa-Speech を，音声合成時に用いる標準語参照音声には，各合成毎に jvs コーパスからランダムな音声を用いた．提案手法との差分はテキストが方言に翻訳されていないことのみである．これにより，SLM の方言理解能力向上が単なるデータ量の増加ではなく擬似方言音声が持つ方言特有の言語的特徴に起因していることを検証する．

3.5　実験結果

表 1 および 表 2 に，CPJD テストセットを用いた標準語翻訳および英語翻訳タスクにおける評価結果を示す．なお，両タスクの評価には，sacrebleu[19] を用いて BLEU スコアを算出した．

標準語翻訳（表 1）では，ベースライン手法の平均スコアは 48.91 であり，データ拡張なし (48.89) と比較してわずか+0.02 の変化にとどまった．対して提案手法は，データ拡張なしと比較して平均 BLEU スコアで 48.89 から 52.64（+3.75）の向上を達成した．英語翻訳（表 2）では，提案手法はデータ拡張なしと比較して平均 BLEU スコアで 24.21 から 26.24（+2.01）の向上が見られた．ベースライン手法も一定の改善を示したが，提案手法はそれを上回る結果となった．

これらの結果は，単に合成によりデータ量が増えたこととは別に，提案手法によって実方言データを

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3　方言音声から英語テキストへの翻訳タスクにおける出力比較

分類	入力音声	標準語訳	英語正解文	アブレーション手法の出力	提案手法の出力
改善例 1 (京都弁)	ほんで食材も **たりひんかった**んどす.	ほんで食材も 足りませんでした.	And we also ran out of ingredients.	And I also **bought** some ingredients.	And we also **ran out of** ingredients.
改善例 2 (いわき弁)	きっと素敵な休日に **なっぺよ**.	きっと素敵な休日に なりますよ.	I'm sure it will be a wonderful holiday.	**I hope you have** a wonderful holiday.	**I'm sure it will be** a wonderful holiday.
失敗例 1 (金沢弁) (方言語彙の不足)	**がめる**ものがめるもの, 古いものが多いがやて.	とるものとるもの, 古いものが多いんだよね.	The things I **pick up** tend to be old.	There are a lot of old things here.	There are a lot of old things here.
失敗例 2 (青森弁) (方言音韻への未適応)	**赤**のスマホでも いいがもな.	赤色のスマホでも よいかもね.	A **red** smartphone might be nice, too.	I might as well use my smartphone.	Maybe it's okay if it's an iPhone.

用いずに合成された擬似方言音声が SLM の方言理解能力向上に有効であることを示している．

3.6　分析

本節では，提案手法による SLM の方言能力の向上が具体的にどのような方言能力の獲得によるものか，また課題は何かを明らかにするため，実際の翻訳結果を用いた定性分析を行う．表 3 に，提案手法とベースライン手法の出力例を示す．

3.6.1　提案手法による改善要因

改善例からは，提案手法が方言特有の文法や語彙を学習できていることが読み取れる．改善例 1（京都弁）では，ベースライン手法が「たりひんかった（足りなかった）」という否定を含む方言表現を認識できず，買った（bought）と誤認識しているのに対し，提案手法は方言表現を正確に理解して「ran out of」と翻訳できている．これは，擬似方言データに含まれる方言特有の否定形（〜ひん）の学習効果が現れていると考えられる．改善例 2（いわき弁）では，ベースライン手法が「なっぺよ」をおそらく標準語の「なってよ（願望）」と混同し「I hope」と訳出しているのに対し，提案手法は「なっぺよ」が持つ推量や確信のニュアンスを理解し，「I'm sure it will be」と正しく訳出している．

これらの結果は，標準語韻律であっても方言の言語的特徴を持つ擬似方言音声での学習を通じて，SLM が方言の文法・語彙構造に対する知識を獲得できることを裏付けている．

3.6.2　提案手法の課題

一方で，失敗例からは本手法の二つの課題が示唆された．第一の課題は網羅されていない方言語彙である．失敗例 1 では，両モデルとも「がめる（＝取る）」という語彙を認識できず翻訳に反映できていない．学習データを確認したところ，元となる標準語テキストには「がめる」と翻訳すべき「取る」の表現が複数含まれていたが，実際に正しく翻訳された例は無かった。これは，LLM を用いた方言翻訳の性能が提案手法の課題であることを示している．第二の課題は方言の韻律変化に対する脆弱性である．失敗例 2 では，テキスト上は標準語と同じ「赤」という単語が含まれているが，青森弁訛りの発音により，両手法ともに「赤」を認識できていない．提案手法は，合成方言テキストを標準語韻律で音声化して学習しているため，語彙や文法は学習できるものの，実音声特有の音韻的な変化に対する適応には限界があることが示唆される．

4　おわりに

本研究では，実方言音声を一切用いずに合成する擬似方言音声による SLM の方言適応を提案した．本手法は，LLM により生成した方言テキストを標準語参照音声で条件付けして音声化することで，標準語韻律を持つ擬似方言音声データを生成するものである．CPJD の 20 方言を用いた評価実験の結果，本手法により生成したデータを学習に追加することで，方言音声から標準語テキストへの翻訳は BLEU スコアで平均 48.89 から 52.64 へ，英語テキストへの翻訳は平均 24.21 から 26.24 へと向上した．実験結果と分析から，本手法が生成する「言語的特徴は方言であるが韻律的特徴は標準語である擬似方言音声」が，SLM の方言理解能力向上に有効であることを示した．これは，実方言音声収集が困難な方言に対しても，SLM の方言学習に有効な合成データ生成が本手法によって可能になったことを意味する．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, et al. Recent advances in speech language models: A survey. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13943–13970, Vienna, Austria, July 2025. Association for Computational Linguistics.

[2] Dong Zhang, Shimin Li, Xin Zhang, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 15757–15773, Singapore, December 2023. Association for Computational Linguistics.

[3] Alexandre Défossez, Laurent Mazaré, Manu Orsini, et al. Moshi: a speech-text foundation model for real-time dialogue, 2024.

[4] Qingkai Fang, Shoutao Guo, Yan Zhou, et al. LLaMA-omni: Seamless speech interaction with large language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[5] Wenxi Chen, Ziyang Ma, Ruiqi Yan, et al. SLAM-omni: Timbre-controllable voice interaction system with single-stage training. In **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 2262–2282, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] Tomoya Mizumoto, Yusuke Fujita, Hao Shi, Lianbo Liu, Atsushi Kojima, and Yui Sudo. Evaluating Japanese dialect robustness across speech and text-based large language models. In **2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**. IEEE, 2025.

[7] Tianduo Wang, Lu Xu, Wei Lu, and Shanbo Cheng. From tens of hours to tens of thousands: Scaling back-translation for speech recognition. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 12461–12475, Suzhou, China, November 2025. Association for Computational Linguistics.

[8] Alan Dao, Dinh Bach Vu, Huy Hoang Ha, et al. Speechless: Speech Instruction Training Without Speech for Low Resource Languages. In **Interspeech 2025**, pp. 3239–3243, 2025.

[9] Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision, 2022.

[10] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, et al. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, University of Pennsylvania, USA, October 2024.

[11] Y. Yin, D. Mori, and S. Fujimoto. ReazonSpeech: A Free and Massive Corpus for Japanese ASR. In **Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing**, 2023.

[12] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025.

[13] Shuichiro Shimizu, Chenhui Chu, Sheng Li, and Sadao Kurohashi. Towards speech dialogue translation mediating speakers of different languages. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1122–1134, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.

[15] Shinnosuke Takamichi and Hiroshi Saruwatari. CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[16] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

[17] Respair Soshyant, Auto Meta, Cryptowooser, and Buttercream. Tsukasa speech: Engineering the naturalness and rich expressiveness. https://huggingface.co/Respair/Tsukasa_Speech, 2024. Accessed: 2026-01-06.

[18] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, et al. Jvs corpus: free japanese multi-speaker voice corpus, 2019.

[19] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　方言翻訳規則生成用プロンプト

本研究で LLM を用いた方言翻訳の際に，標準語から対象方言への翻訳規則を生成するために LLM に与えたプロンプトを以下に示す．プロンプト内の「○○弁」には，各実験対象の方言名（例：津軽弁，大阪弁）を挿入して実行した．

方言翻訳規則生成プロンプト

あなたはプロの○○弁 → 標準語翻訳者です．○○弁について，指定された以下の項目の標準語 → ○○弁の変換ルールを作成し，JSON 形式で出力してください．○○弁以外の方言との混同は避けてください．指定の方言では変化の起こらない項目は飛ばしてください．

 # 制約

- 固有名詞，数字，専門用語は絶対に変更しないこと
- 階層構造は作らず，フラットなリストにすること
- 値がない場合は出力しないこと

 # 項目キーワード

断定, 否定, 説明, 接続, 質問, 情報提示, 同意・共感, 依頼, 丁寧依頼, 終助詞・語尾・語頭, 推量・確認, 念押し・通知, 相手への確認・軽い反駁, 説明・理由, 念押し・感嘆, 文法・機能表現, 五段活用, 一段活用, サ変, 可能否定, 過去否定, 丁寧否定, 説明・理由の終止, 過去説明, 断定・推量・質問, 接続（順接・逆接・添加）, 依頼・勧誘, 否定依頼, 禁止・許可・義務, 存在・尊敬（おる／〜はる）, 進行・状態, 語彙（代表的な置換・意味拡張）, 人称表現, 格助詞の変換, 形容詞の活用, 音韻変化, その他特別な方言表現

 # 出力形式 (JSON)

解説は不要．以下のキーを持つオブジェクトの配列のみを出力せよ．

key:　項目キーワード
rule:　具体的な変換ルールや特徴
sample:　標準語 → 方言の変換例（短い文で）

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).