記述言語学に基づいた方言音声合成評価枠組みの構築と池間西原方言音声合成の評価

佐藤なな子 [1]　阪井瞭介 [1]　中田亘 [2,3]　高道慎之介 [1,2]
中川奈津子 [4]　林由華 [5]　宮川創 [6]　坂井美日 [7]
[1] 慶應義塾大学　[2] 東京大学　[3] 産総研　[4] 九州大学　[5] 岡山大学　[6] 筑波大学　[7] 鹿児島大学
nanakosato@keio.jp　shinnosuke_takamichi@keio.jp

概要

方言テキスト音声合成（text-to-speech; TTS）は消滅危機方言を保護する目的で近年研究されている．本研究は，方言特徴を踏まえた方言 TTS の評価枠組みを構築し，宮古語池間西原方言 TTS に対し，枠組みを適用する．また，データの少ない方言録音資料を用いた方言 TTS 構築手法についても提案する．方言合成音声を評価した結果，提案手法によって音質の大幅な改善が見られたものの，方言としての自然さについては低下傾向にあることを報告する．

1　はじめに

日本語には様々な地域方言が存在する．しかし，話者減少により伝承が難しくなっている方言も多い [1]．言語や方言の消滅は，文化・アイデンティティを失うことに等しく [2]，その消滅から守る一つの手段として，近年では方言テキスト音声合成（text-to-speech; TTS）の研究が進められている [3]．

TTS モデルの性能はその合成音声により評価され，それに資する評価用データセットを用いることが重要である [4]．方言 TTS モデルにおいては，その合成音声の**方言らしさ**に対して主観評価を実施することが一般的であるため，評価文にも**方言らしさ**が求められる [5]．しかしながら，先行研究における評価文は，データセットからランダムに選択する [5, 6, 7] か，合成音声の特徴を考慮した評価文を選択する [3] か，そもそも選択方法が明記されていない [8, 9, 10] ことがほとんどで，いずれも**方言らしさ**を十分に考慮できているとは言い難い．

これらを踏まえ本研究では，方言特徴を記述言語学の視点から列挙し，**方言らしさ**を備えた評価文セットの構築法を提案し，実際に池間西原方言の評価文セットを構築する．さらに，TTS 構築用ではない音声資源から池間西原方言 TTS モデルを構築する方法を述べ，前述の評価文セットにて評価した結果も報告する．

2　評価文の策定

2.1　記述言語学に基づいた評価枠組み

記述言語学とは，ある特定言語の，ある時代の性質を，ありのままに記述する学問であり [11]，これに則ることで，言語や方言特有の音韻規則を体系的に表現できる．本研究では，“方言の現代での性質を，ありのままに記述する” ことで，方言特徴を以下の 5 種類に分類し，評価枠組みを構築する．

- **音価**: 国際音声記号（international phonetic alphabet; IPA） [12] を用いて表記し，条件異音や有声・無声の違いを評価する．
- **音素**: 東京方言（日本標準語）に存在しない音素をはじめ，その方言に出現する全音素の発音正確性を評価する．
- **モーラ**: 東京方言に存在しないパターンのモーラの発音正確性を評価する．
- **アクセント**: アクセントのみが異なる同音異義語などの言い分けを評価する．
- **韻律**: 文に応じて変わる高低（イントネーション等）の言い分けを評価する．

2.2　池間西原方言への適用

本研究は，宮古島北部西原地域で話される池間西原方言に着目し， 2.1 節で示した評価枠組みを適用する．先行研究 [13, 14, 15, 16, 17, 18, 19, 20] に基づき，池間西原方言に対し，評価文を 100 文用意した．全文はプロジェクトページ[1)]にて入手可能である．

1) https://github.com/takamichi-lab/ikema_tts_evaluation

表 1　池間西原方言の評価枠組み

種類	方言の特徴	文例
音価	母音の無声化，子音の口蓋化， 子音/h/・撥音/N/の条件異音， 子音/r/の二重子音	ふにぬ はー.（舟の刃） ふにぬ ひー.（舟の日和） ふにぬ ほぅー.（舟の帆）
音素	中舌母音 /ɨ/，対立する子音/h/と/f/， 無声鼻音 /n̥/，音素バランス文	ん゜んはっじゃき. （踏み外す）
モーラ	語頭促音，長音後の促音， 東京方言にはない拗音	っさりー にゃーん.（触れてしまった） さりー にゃーん.（枯れてしまった）
アクセント	アクセントの異なる同音異義語	いんまい にゃーん.（海もない）
韻律	疑問文の句末音調	くりゃー むずぐるな？↗（これは麦の茎か？） くりゃー むずぐる？↘（これは麦の茎？）

3　方言音声合成モデル

3.1　要請される条件

本研究では，現存する方言録音資料を用いて方言 TTS モデルを構築する．方言録音資料と理想的な音声コーパスは以下の点で異なる．

- **データ量が学習に十分でない場合がある**: 方言話者の高齢化により，長時間の収録が難しい．
- **テキストと音声が対応しない場合がある**: 公教育によって体系的に表記を習うわけではないため，言語学者の用意したテキストと，母語話者による発話の不一致や，言い間違い・咳き込みが発生しうる．
- **家屋などの一般的な環境で収録されている**: ドアの開閉音，紙をめくる音，車の音といった生活音が背景雑音として含まれる．
- **音素セットとアクセント規則が東京方言と異なる**: 東京方言に含まれない音素の存在や，書記素が異なる場合がある．また，東京方言が多型アクセントであるのに対して異なるアクセント型であることも考えられる．

これらの課題への解決策を 3.2–3.3 節で示す．

3.2　データ前処理

3.2.1　テキスト–音声非対応データの除去

入力テキストと異なる発音となっている音声は，学習データとして不適である．したがって音素アライメントを行い，音素ラベルと音声の対応スコアが低いデータをデータセットから除去する．

3.2.2　音声強調

背景雑音やマイク音響歪みの混在するデータセットで TTS モデルを学習すると，背景雑音やマイク歪みなどの TTS に不必要な要素を学習してしまう．これに対し昨今，任意の雑音や歪みを追加学習なしに除去できる universal speech enhancement の研究が進んでいる [21]．本研究では，学習済みモデルによる音声強調を実施する．

3.2.3　無音区間調整

1 文の音声は，非発話区間に挟まれてファイルに保存されている．しかし非発話区間の時間長にばらつきがあると，TTS モデルは，非発話区間に対し煩雑な時間長予測を強いられる．そこで，発話区間検出を用いて，非発話区間の時間長を一定に揃える．

3.3　モデル学習

3.3.1　G2P 変換の設計と入力の変更

TTS は書記素テキストを入力すると，grapheme-to-phoneme（G2P）で音素列テキストに変換する．しかし一般的な G2P (例えば東京方言の G2P) では未知書記素を<unk>トークンに変換するため，書記素次第では方言テキストを音素に変換できない．したがって，当該方言特化のルールベース G2P を構築する．

3.3.2　調音特徴量の導入

事前学習済みモデルに含まれない音価・音素・モーラに対応するため，調音特徴量を導入する．図 1 のように，当該音素と，その前後の音素から調音特徴量ベクトルを算出し，音素 ID に結合する．音価・音素・モーラ同士の発音がどれだけ似ているかという，音素 ID のみでは表現できない特徴を表現できる．

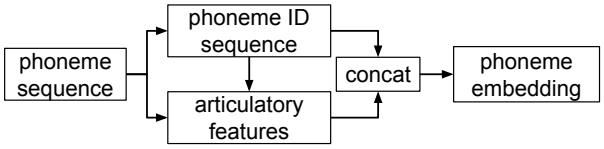

図 1　調音特徴量を導入した音素表現

3.3.3　音響モデルと fine-tuning

方言録音資料のデータ量の乏しさをカバーするため，東京方言で学習された TTS モデルを fine-tuning する．その際，最初段の音素埋め込みを 3.3.2 節の特徴量に置換する．

4　実験・評価

4.1　実験条件

4.1.1　データセット

『池間方言辞典』[13] の著者の一人で母語話者である仲間博之氏が辞書の内容を読み上げた，「南琉球宮古語池間西原方言 音声語彙データベース」[2] を学習データとした．学習は，言い間違いが含まれるデータなどを取り除き，語彙 5279 語，例文 7567 文から成る全 12846 発話を，学習 80%，検証 10%，テスト 10%に分割して行った．

4.1.2　方言音声合成モデル

本研究では，学習データを前処理なしに fine-tuning する手法をベースラインとし，以降に示す各種の前処理を施した手法を提案手法とする．学習済みモデルは，ESPnet [22, 23] を用いて学習された `kan-bayashi-jsut-full-band-vits-prosody` [24] である．

- **音素アライメント**: Julius [25] を用いて，音素と音声の対応スコアを算出した．その後，下位約 22% をデータセットから除去し，学習は語彙 4272 語，例文 5278 文，全 10000 発話で行った．
- **G2P を変更**: 池間方言特有の音素である中舌母音 /ɨ/，無声鼻音/n̥/を他の音素と区別して変換できるよう，G2P の辞書を変更した．また，表 1 にあるような文節区切りの空白文字は，発話中の間と無関係であるため，考慮しない形に変更した．
- **音声強調**: Sidon [26] を用いて，学習データの背景雑音と歪みを除去した．
- **非発話区間の時間長調整**: sileroVAD [27] を用いて発話区間を検出し，発話区間前後の非発話区間が 0.4 – 0.6[s] となるように非発話区間を追加あるいは削除した．
- **調音特徴量の導入**: 先行研究 [28] を参考に，池間西原方言の音素を，以下の 26 次元の調音特徴を 2 値ベクトルで表現した．
  - 母音，半母音，子音，記号
  - 舌の位置が前，中央，後ろ
  - 唇の開き方が広い，普通，狭い
  - 無声音，有声音
  - 鼻音，破裂音，破擦音，摩擦音，接近音，弾き音，側面接近音
  - 両唇音，唇歯音，歯茎音，硬口蓋音，軟口蓋音，口蓋垂音，声門音

表 2　客観評価指標

手法	評価内容
MCD; mel-cepstral distortion [29]	スペクトル歪み
log F_0 RMSE	ピッチ歪み
SpeechBERTScore [30]	音声埋め込みの一致度
SpeechTokenDistance [30]	音声離散トークンの一致度
UTMOS [31]	音質
PER; phone error rate	発音の正確さ
音素アライメントスコア [25]	音素列と音声の対応度

4.2　客観評価

4.2.1　評価手法

各種モデルそれぞれで 2.2 節で述べた評価文 100 文を合成し，表 2 に示す評価を行った．

PER は POWSM [32]，その他はオープンソース[3] を使用して算出した．

4.2.2　実験結果

客観評価結果を表 3 に示す．音声強調の導入前後でスコアが顕著に変化しており，音声強調が，方言 TTS モデルの性能，特に音質の向上に一定の寄与があると言える．また，調音特徴量の導入を行ったモデルが，speechBERTscore, speech token score, PER の 3 指標において最も性能が高く，提案手法の有効性を示している．一方で，方言特有のアクセントや韻律の再現に対するアプローチを行えていないため，MCD や log F_0 RMSE については，性能向上があま

2）https://hdl.handle.net/2324/7238327, 例文の録音も近日公開予定.

3）https://github.com/Takaaki-Saeki/DiscreteSpeechMetrics

表 3　客観評価結果

Model	MCD [dB] ↓	log F_0 RMSE ↓	SpeechBERTScore ↑	SpeechTokenDistance ↑	UTMOS ↑	PER ↓	Alignment ↑
参照音声	—	—	—	—	2.15 ± 0.52	0.311	143 ± 20.1
Fine-tuning のみ	5.91 ± 1.28	0.418 ± 0.222	0.769 ± 0.038	0.538 ± 0.059	1.97 ± 0.45	0.362	121 ± 4.82
+ アライメント	4.16 ± 0.78	**0.310 ± 0.117**	0.793 ± 0.031	0.553 ± 0.057	2.17 ± 0.40	0.372	129 ± 8.37
+ 新 G2P	4.10 ± 0.87	0.310 ± 0.117	0.796 ± 0.034	0.554 ± 0.060	2.23 ± 0.41	0.369	129 ± 7.53
+ 音声強調	**2.15 ± 0.67**	0.351 ± 0.147	0.879 ± 0.034	0.755 ± 0.093	3.05 ± 0.47	0.359	**136 ± 6.15**
+ 無音区間調整	2.35 ± 0.71	0.383 ± 0.136	0.880 ± 0.032	0.831 ± 0.077	**3.13 ± 0.38**	**0.348**	123 ± 3.61
+ 調音特徴量	2.34 ± 0.58	0.340 ± 0.121	**0.882 ± 0.035**	**0.837 ± 0.072**	3.03 ± 0.41	**0.348**	126 ± 3.63

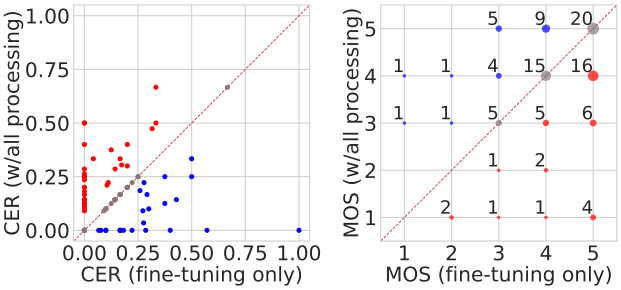

図 2　主観評価結果の散布図．各点は各テキストを表す．青点は，fine-tuning のみのモデルよりも，全処理を加えたモデルが優れた場合である．赤点はその逆である．

り見られなかった．

4.3　主観評価

4.3.1　評価方法

本研究では，表 3 で示した 6 種類のモデルのうち，fine-tuning のみのモデルと，下端に示した全ての処理を加えたモデルのそれぞれに対して，100 文を合成し主観評価を行った．なお，提示順序によるバイアスが生じないよう，200 発話をランダムに提示した．評価者は池間西原方言の専門家 1 名（著者の 1 人：林）で，以下の 2 点について評価を依頼した．

- **発音の正確さ**：まず，評価文テキストを見ずに，合成音声のみを聞いてテキストに書き起こす．次に評価文テキストを提示し，発音の不正確だった箇所を記述する．記述された結果を元に，character error rate (CER) を算出し，モデル間の性能比較を行う．
- **方言音声としての自然さ**：合成音声を聴取し，方言音声としての自然さを 5 段階 MOS（mean opinion score）で評価を行う．

4.3.2　実験結果

主観評価結果を図 2, 表 4 に示す．

表 4　主観評価結果詳細

モデル	置換誤り	挿入誤り	削除誤り	CER	MOS
fine-tuning のみ	73	37	18	0.141	4.16
w/ all processing	134	3	9	0.163	3.86

- **発音の正確さ**: fine-tuning のみのモデルで最頻出であった，文末に不要な長音が挿入される例は，全ての処理を行ったモデルの音声からは発見されなかった．また不自然な区間 pause の存在も，全処理済みモデルでは指摘されなかった．これは無音区間の調整によって，非発話区間の合成が安定したと考えられる．一方で，全処理済みモデルでは置換誤りの増加が見られた．これに伴い CER の値も悪化した．特に多かったのは文頭の k 子音の欠落 (例えば “か” が “あ” となる) で，音声強調時に文頭の無声子音も除去してしまった可能性や，導入した調音特徴量と実際の発話のズレが原因としてあげられる．
- **方言音声としての自然さ**: こちらも置換誤りの増加により，スコアの低下が見られた．また，発音の言い分けを評価する目的で作成した，文意の通らない評価文に対するスコアが低い傾向にあり，意味の通じる評価文を選択する必要性が示唆された．

5　おわりに

本研究では，方言録音資料を用いた方言 TTS の構築と，それに対する評価枠組みの構築を行った．記述言語学に基づいた方言特徴が反映された評価文を用い，構築したモデルを評価した結果，音質の著しい向上が見られたものの，**方言らしさ**に対してはやや低下が見られた．今後は音素の置換誤りを防ぐ仕組みの導入や，本研究で未検討のアクセントや韻律に対するアプローチを行い，さらなるモデルの改善を目指す．

謝辞： 本研究は，JSPS 科研費 24K00074 の助成を受けたものです．また，本研究にご協力いただいた，仲間博之先生に感謝申し上げます．

参考文献

[1] 木部暢子 et al., **危機的な状況にある言語・方言の実態に関する調査研究事業報告書**. 人間文化研究機構国立国語研究所, 2011. [Online]. Available: https://ndlsearch.ndl.go.jp/books/R100000002-I023425033

[2] 国立国語研究所研究情報誌編集委員会, “国語研ことばの波止場： 国立国語研究所研究情報誌： Ninjal research digest 13 巻 2024 年 3 月,” 2024. [Online]. Available: https://ndlsearch.ndl.go.jp/books/R100000002-I028061246-i31892327

[3] L.-W. Chen et al., “Voxhakka: A dialectally diverse multi-speaker text-to-speech system for taiwanese hakka,” in **2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)**, 2024, pp. 1–6.

[4] P. Varadhan et al., “The state of TTS: A case study with human fooling rates,” in **arXiv**, 08 2025, pp. 2285–2289.

[5] K. Yamauchi et al., “Cross-dialect text-to-speech in pitch-accent language incorporating multi-dialect phoneme-level bert,” in **2024 IEEE Spoken Language Technology Workshop (SLT)**, 2024, pp. 750–757.

[6] X. Di et al., “Bailing-TTS: Chinese dialectal speech synthesis towards human-like spontaneous representation,” **arXiv**, vol. abs/2408.00284, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271600856

[7] T. Bollinger et al., “Text-to-speech pipeline for swiss german - a comparison,” **arXiv**, vol. abs/2305.19750, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258987769

[8] A. Guevara-Rukoz et al., “Crowdsourcing Latin American Spanish for low-resource text-to-speech,” in **Proceedings of the Twelfth Language Resources and Evaluation Conference**, N. Calzolari et al., Eds. Marseille, France: European Language Resources Association, May 2020, pp. 6504–6513. [Online]. Available: https://aclanthology.org/2020.lrec-1.801/

[9] K. Doan et al., “Towards zero-shot text-to-speech for Arabic dialects,” in **Proceedings of the Second Arabic Natural Language Processing Conference**, N. Habash et al., Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 123–129. [Online]. Available: https://aclanthology.org/2024.arabicnlp-1.11/

[10] L. Gutscher et al., in **Neural Speech Synthesis for Austrian Dialects with Standard German Grapheme-to-Phoneme Conversion and Dialect Embeddings**, 08 2023.

[11] 小学館国語辞典編集部, **日本国語大辞典. 第 1 巻 (あ-こ)**, 精選版 ed. 小学館, 2006, no. 第 1 巻 (あ-こ). [Online]. Available: https://ndlsearch.ndl.go.jp/books/R100000002-I000008032265

[12] International Phonetic Association, **Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet**. Cambridge University Press, 1999. [Online]. Available: https://books.google.co.jp/books?id=33BSkFV_8PEC

[13] 仲間博之 et al., **南琉球・宮古語 池間方言辞典： 西原地区版**. 国立国語研究所 言語変異研究領域, 2024, no. 2. [Online]. Available: [https://ndlsearch.ndl.go.jp/books/R100000002-I000008032265](https://repository.ninjal.ac.jp/records/2000172)

[14] 田窪行則, “池間方言形態音韻論–名詞提題形，対格形を中心に,” **日本音声学会公開講演**, 2018.

[15] 林由華, “琉球語宮古池間方言の談話資料,” **大西正幸・稲垣和也 (編) 『地球研言語記述論集 1』**, pp. 153–199, 3 2009.

[16] 五十嵐陽介, “名詞の意味が関わるアクセントの合流－南琉球宮古語池間方言の事例－,” **音声研究**, vol. 20, no. 3, pp. 46–65, 12 2016.

[17] 藤本雅子 et al., “南琉球宮古島池間方言の鼻子音の調音,” **音声研究**, vol. 27, no. 1, pp. 13–26, 2023.

[18] 五十嵐陽介 et al., “琉球宮古語池間方言のアクセント体系は三型であって二型ではない (<特集>n 型アクセント研究の現在),” **音声研究**, vol. 16, no. 1, pp. 134–148, 2012. [Online]. Available: https://cir.nii.ac.jp/crid/1390282679763725056

[19] トマ・ペラール, 林由華, “宮古諸方言の音韻：体系と比較,” **消滅危機方言の調査・保存のための総合的研究 南琉球宮古方言調査報告書**, pp. 13–52, 8 2012.

[20] 五十嵐陽介, “南琉球宮古語池間方言・多良間方言の韻律構造,” **言語研究 = Journal of the Linguistic Society of Japan / 日本言語学会 編**, no. 150, pp. 33–57, 2016. [Online]. Available: https://ndlsearch.ndl.go.jp/books/R000000004-I027745307

[21] K. Saijo et al., “Interspeech 2025 URGENT speech enhancement challenge,” in **26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025**, O. Scharenborg et al., Eds. ISCA, 2025. [Online]. Available: https://doi.org/10.21437/Interspeech.2025-1363

[22] S. Watanabe et al., “ESPnet: End-to-end speech processing toolkit,” in **Proceedings of Interspeech**, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[23] T. Hayashi et al., “Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2020, pp. 7654–7658.

[24] kan bayashi, “Espnet2 pretrained model, kan-bayashi/jsut_tts_train_full_band_vits_raw_phn_jaconv_pyopenjtalk_prosody_train.total_count.ave, fs=44100, lang=jp,” Sep. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5521340

[25] A. Lee, T. Kawahara, “julius-speech/julius: Release 4.5,” Jan. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2530396

[26] W. Nakata et al., “Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing,” in **arXiv**, 09 2025. [Online]. Available: https://arxiv.org/abs/2509.17052

[27] S. Team, “Silero VAD: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier,” https://github.com/snakers4/silero-vad, 2024.

[28] M. Staib et al., “Phonological features for 0-shot multilingual speech synthesis,” in **Interspeech**, 10 2020, pp. 2942–2946.

[29] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in **Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing**, vol. 1, 1993, pp. 125–128 vol.1.

[30] T. Saeki et al., “SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics,” in **Interspeech 2024**, 2024, pp. 4943–4947.

[31] ——, “UTMOS: Utokyo-sarulab system for voicemos challenge 2022,” in **arXiv**, 09 2022, pp. 4521–4525.

[32] C.-J. Li et al., “POWSM: A phonetic open whisper-style speech foundation model,” 2025. [Online]. Available: https://arxiv.org/abs/2510.24992