社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

VQ-VAEに基づく解釈可能なアクセント潜在変数を用いた多方言音声合成

山内　一輝†　　齋藤　佑樹†　　猿渡　　洋†

† 東京大学　〒113–8656 東京都文京区本郷 7-3-1

あらまし　本稿では, 目的話者の母方言と同じ方言のテキスト音声合成 (Text-to-Speech: TTS) を目的とする “Intra-dialect TTS” および，話者の声質を保ったまま目的話者の母方言と異なる方言の TTS を目的とする “Cross-dialect TTS” という 2 つのタスクに取り組む．従来法は，東京方言（標準語）を除く方言には入力テキストにアクセントラベルを付与するために必要なアクセント辞書が存在しないという困難を克服するため，アクセント潜在変数 (Accent Latent Variable: ALV) を参照音声から抽出するかテキストから予測して方言 TTS に利用する．しかし，従来法では参照音声は学習データに含まれる話者による音声に限られ，Cross-dialect TTS については検討されていない．本稿では，任意の話者による参照音声入力や方言に応じた ALV 予測が可能な多方言 TTS 手法を提案する．実験的評価により，提案手法が特に Cross-dialect TTS において合成音声の方言らしさを向上させることを示す．

キーワード　DNN 音声合成，方言音声合成，Cross-dialect TTS，VQ-VAE，アクセント潜在変数，Prosody transfer

Kazuki YAMAUCHI†, Yuki SAITO†, and Hiroshi SARUWATARI†

† The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan

1. は じ め に

テキスト音声合成 (Text-to-Speech: TTS) [1] とは，任意のテキストから対応する自然な読み上げ音声を合成する技術である．音声はその内容に関する情報である言語情報の他に，感情や話者性などのパラ/非言語情報を含む [2]．パラ/非言語情報は音声の韻律を多様にし，音声によるコミュニケーションにおいて非常に重要な要素となる．したがって，TTS はテキストと対応する多様な音声との one-to-many mapping 問題であり，自然な韻律の予測は重要かつ困難な課題となっている．本研究の対象となる日本語は，ピッチの高低によってアクセントを表現するピッチアクセント言語であり，同音異義語の弁別や方言音声としての知覚において音声の韻律は重要な要素となる．そのため，典型的な日本語 TTS モデルは，自然な韻律を再現するために入力としてテキストに対応するアクセントラベルを用いる．図 1 に示すように，典型的な日本語の TTS モデルは，アクセント辞書を用いて入力テキストにアクセントラベルを付与し，その埋め込みベクトルを音素埋め込みに加算して TTS モデルに入力する．

日本語には様々な方言が存在する．方言はそれぞれ異なる韻律体系をもち，音声の話者性はその話者の母方言に大きく依存するため，人間と音声合成モデル間の音声コミュニケーション技術において，方言らしい韻律の再現は重要な課題となる．また，話者の人口が減少している方言を保存するための 1 つの手段としても，方言音声合成モデルの構築は非常に重要な役割をもつ．先述の通り方言は標準語とは異なるアクセント体系をもつため，同一テキストに対するピッチパターンが方言によって異なる場合がある．例えば,「雨が」は東京方言では「あ (H) め (L) が (L)」（H は高いピッチを表し，L は低いピッチを表す）と発音されるが，大阪方言では「あ (L) め (H) が (L)」と発音される．そのため，方言音声の合成には目的方言専用のアクセント辞書を用いることが望ましいと考えられる．しかし，東京方言（標準語）を除く方言には入力テキストにアクセントラベルを付与するために必要なアクセント辞書を存在しないという問題がある．したがって，方言で書かれたテキストが入力されたとしても，特に目的話者の母方言がその方言と異なる場合，目的方言らしいアクセントの音声を合成することは困難である．

方言にはアクセント辞書が存在しないという困難を克服するため，音声から抽出したアクセント潜在変数 (Accent Latent Variable: ALV) という音声のアクセントに関する潜在表現を用いる方言音声合成手法が提案されている [3], [4]．ALV は Vector Quantised-Variational AutoEncoder (VQ-VAE) [5] によって音声から抽出される量子化された潜在ベクトル列である．郡らによる日本語は 4 段階アクセントである [6] という主張に従い，VQ-VAE における量子化クラス数は 4 とされており，これにより ALV は High-Low ラベルに近く解釈性・可制御性の高い潜在変数表現となっている．彼らは，参照音声から抽出された ALV を音声の合成時に入力することで，合成音声のアクセントの自

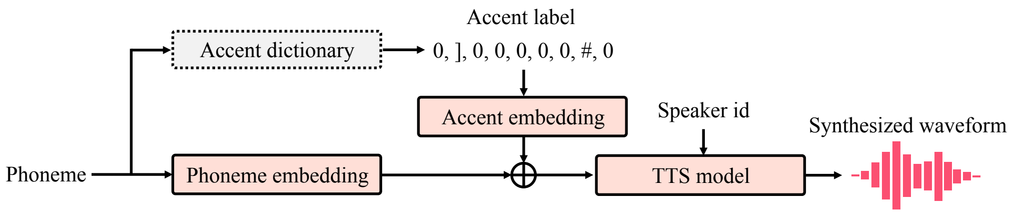

図 1　典型的な日本語 TTS モデル．アクセント辞書を用いて入力テキストにアクセントラベルを付与し，その埋め込みベクトルを音素埋め込みに加算して TTS モデルに入力する．

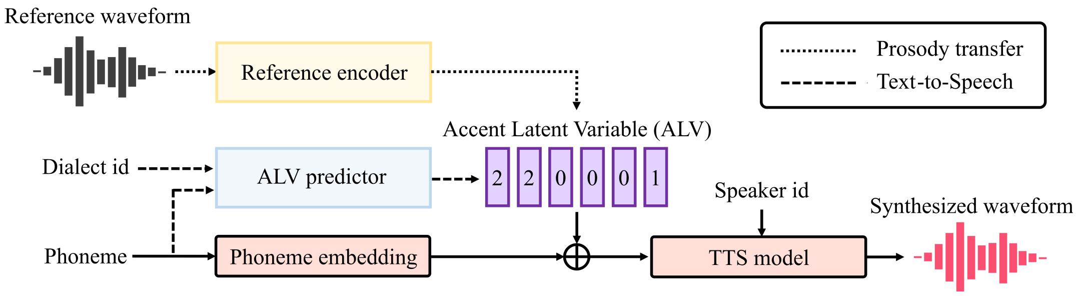

図 2　提案モデルの概略図．参照音声が入力された場合，Reference encoder が参照音声から ALV を抽出し，音素埋め込みに加算して TTS モデルに入力する (prosody transfer)．参照音声を用いない場合，ALV predictor が目的方言に応じてテキストから対応する ALV 列を予測し，音素埋め込みに加算して TTS モデルに入力する (Text-to-Speech)．

然性が向上することを示した．ただし，参照音声は学習データに含まれる話者による音声に限られており，目的話者の母方言と異なる方言の TTS については検討されていない．

学習済みの TTS モデルによる合成音声を，所望の韻律を持つ参照音声を用いて適応させる技術は prosody Transfer と呼ばれる [7]．特に，方言音声を対象とした prosody transfer では，より fine-grained な（すなわち，音素や単語，フレーム単位の）韻律特徴量を用いた適応 [8], [9] が有効であると考えられる．CopyCat2 [10] は，モデルの学習を 2 段階に分けており，第 1 段階では prosody transfer のためのモジュールを学習し，第 2 段階ではテキストから韻律特徴量を予測するためのモジュールを学習する．Accent-VITS [11] は，事前学習済みの Automatic Speech Recognition (ASR) モデルから抽出された bottleneck 特徴量由来の韻律特徴量でテキスト由来の言語特徴量に KL 制約をかけて学習することで，prosody (accent) transfer を可能にしている．ただし，これらのモデルで用いられている韻律特徴量は連続的な特徴量であり，人間が解釈することは困難である．合成音声のアクセントに誤りが含まれていた場合，それを人間が簡単に訂正できることが望ましいが，韻律特徴量の解釈が難しい場合それは困難である．また，日本語のアクセントは 4 段階であるとされている [6] ことから，韻律特徴量を連続のまま扱うのではなく，量子化して扱う方が効率的である可能性が考えられる．実際，湯舟らはアクセント情報を抽出するモデルとして，連続潜在ベクトルを扱う VAE [12] よりも潜在ベクトルを量子化して扱う VQ-VAE の方が韻律の再現において優れていることを示している [3]．

本稿では，任意の話者による参照音声を用いた prosody transfer および方言に応じた ALV 予測が可能な多方言 TTS 手法を提案する．図 2 に示すように，提案モデルには “Reference encoder” と “ALV predictor” という 2 つのモジュールが組み込まれている．Reference encoder は任意の話者による参照音声から ALV を抽出するためのモジュールであり，ALV predictor は目的方言に応じてテキストから対応する ALV 列を予測するためのモジュールである．これらのモジュールによって得られた ALV は音素埋め込みに加算され，後段の TTS モデルに入力される．提案モデルは複数の方言の音声を含むデータセットを用いて学習され，推論時は目的話者と目的方言を指定して音声を合成することができる．実験では，目的話者の母方言と同じ方言の TTS を目的とする “Intra-dialect TTS” および，話者の声質を保ったまま目的話者の母方言と異なる方言の TTS を目的とする “Cross-dialect TTS” という 2 つのタスクによって提案モデルを評価する．評価は音声の自然性およびアクセントの目的方言らしさに関する Mean Opinion Score (MOS) を測る主観評価実験と，音声の明瞭性と目的話者との話者類似度を評価するための客観評価実験により行う．実験結果から，我々の提案手法が (1) Cross-dialect TTS において合成音声のアクセントの目的方言らしさを向上させること，(2) 未知の話者による参照音声を用いた prosody transfer により合成音声のアクセントの目的方言らしさを向上させることが示された．なお，サンプル音声はデモページにて公開されている [1]．

2. 提案手法

図 3 に提案モデルのアーキテクチャを示す．提案モデルは主に山内らの先行研究 [4] に基づくが，各モジュールのアーキテクチャの詳細や学習方法に変更がある．以降では提案モデルにおける各モジュールの詳細や学習方法について説明する．

（注 1）：https://kyamauchi1023.github.io/yamauchi24sp03

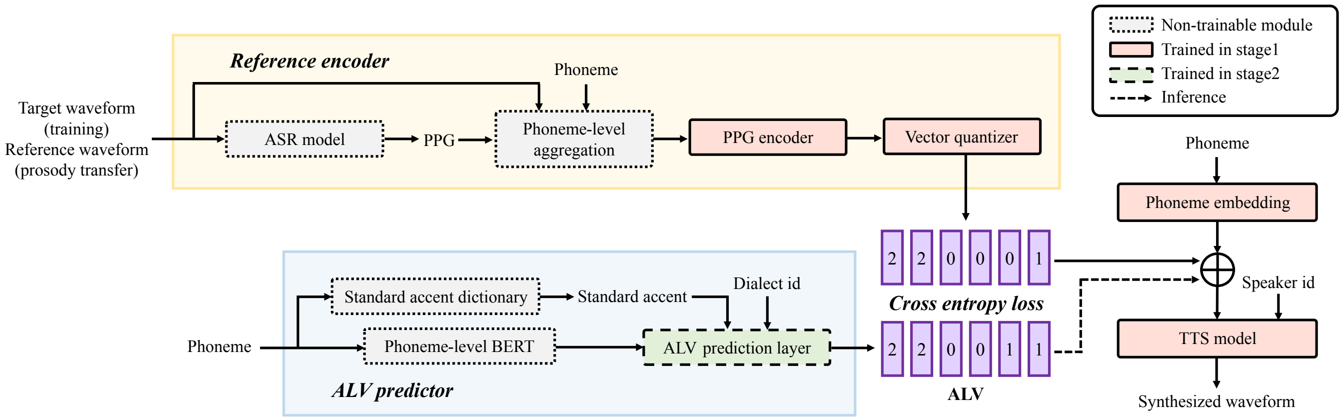

図 3 Reference Encoder と ALV Predictor からなる提案モデルのアーキテクチャ．学習の前半 (Stage1) では Reference encoder と TTS モデルが学習され，学習の後半 (Stage2) では ALV predictor が学習される．

2.1 Reference Encoder

Reference encoder は参照音声から ALV を抽出するためのモジュールである．まず，入力された参照音声から韻律に関する特徴量を抽出する．先行研究 [4] では韻律特徴量として基本周波数 (F0) が用いられていたが，提案モデルは Phonetic Posteriorgrams (PPG) [13] を用いる．PPG は事前学習済み ASR モデル から得られる特徴量であるが，日本語において単語の弁別にはアクセント情報が必要であるため，PPG には韻律に関する情報が十分に含まれていると考えられる．さらに，さまざまな音質や話者を含む大規模な音声認識用のデータで学習された ASR モデルを活用することで，未知話者に対しても頑健なアクセント情報抽出が可能になることが期待される．

提案モデルでは ALV は音素単位の韻律特徴量とするため，PPG を音素単位に aggregation する．具体的には，音素アライメント情報を用いて，同一音素に対応する区間の PPG を平均化したものをその音素に対応する PPG とする．Aggregation された PPG は 1 次元 Convolutional Neural Network (CNN) ベースの PPG エンコーダに入力され，その出力はベクトル量子化モジュールによって特定のクラス数に量子化される．

2.2 ALV Predictor

ALV predictor は目的方言に応じてテキストから対応する ALV 列を予測するためのモジュールである．音素エンコーダには，先行研究 [4] と同様に音素単位の言語モデルである Phoneme-level BERT [14] を用いる．テキストからアクセントを予測ためには音素のみでなく書記素 (grapheme) の情報が有効であると考えられる．例えば，「雨」と「飴」は音素はともに “a m e” だが，異なるアクセントをもつ．Phoneme-level BERT は事前学習時に音素から書記素（単語）を予測するタスクを解くように学習されるため，Phoneme-level BERT から得られる特徴量は ALV 予測に有効であることが考えられる．Phoneme-level BERT により得られた特徴量は，標準語のアクセントラベルと方言 ID の埋め込みベクトルと連結され，Bi-directional Long Short-Term Memory (BiLSTM) ベースの学習可能レイヤーに入力される．

2.3 学 習

提案モデルの学習は 2 段階に分けられる．第 1 段階では Reference encoder と TTS モデルが共同で学習される．ただし，Reference encoder に用いられる ASR モデルは学習済みのものを使用し，パラメータの更新はしない．損失関数は TTS モデルの損失関数にベクトル量子化のための損失関数 [5] を加えたものである．なお，学習時は教師データである目的音声を参照音声として使用する．

第 2 段階では ALV predictor が学習される．ALV predictor は Reference encoder から得られた ALV のクラス ID を予測するように学習される．すなわち，損失関数は ALV predictor の出力と Reference encoder から得られた ALV のクラス ID の Cross entropy loss である．ただし，ALV predictor に用いられる Phoneme-level BERT は学習済みのものを使用し，パラメータの更新はしない．

3. 実験的評価

3.1 実 験 条 件

本実験では，モデルの学習のためのデータセットとして JSUT コーパス [15] および JMD コーパス [16] を用いた．JSUT コーパスは単一の標準語話者（女性話者）による約 7700 発話の読み上げ音声からなるコーパスであり，JMD コーパスは大阪方言話者（女性話者）および熊本方言話者（男性話者）による各 1300 発話の読み上げ方言音声からなるコーパスである．本稿では方言の中でも特に大阪方言の TTS を目的とし，学習の際は JSUT コーパスと JMD コーパスの大阪方言サブセットを混合し，学習用（8484 発話），検証用（256 発話），評価用（256 発話）サブセットに分割して用いた．また，学習データに存在しない未知話者による prosody transfer の有効性を評価するために，prosody transfer のための参照音声として CPJD コーパス [17] [2] 含まれる大阪方言話者（男性話者）による音声を用いた．CPJD コーパスはクラウドソーシングによって集められた多方言音声コーパスであり，各方言毎に 250 発話の音声が収録されている．

（注2）：https://sites.google.com/site/shinnosuketakamichi/research-topics/cpjd_corpus

表 1　実験により得られた 95% 信頼区間付き MOS，x-vector の コサイン類似度 (COSSIM) および CER．入力テキストは大阪方言で書かれており，方言性 MOS は音声のアクセントが大阪方言としてどの程度自然かを示す．REF は参照音声として用いた生の音声サンプルを表す．太字は FS と FS2-AP の評価値に $p < 0.05$ の有意差があったことを示す．

(a) Task 1: Intra-dialect TTS の評価結果

Method	目的話者	Subjective Evaluation		Objective Evaluation	
		自然性 MOS (↑)	方言性 MOS(↑)	COSSIM (↑)	CER (↓)
FS2 (baseline)	JMD (大阪方言話者)	2.91 ± 0.120	3.15 ± 0.145	0.990	11.0
FS2-AP (proposed)	JMD (大阪方言話者)	2.91 ± 0.129	3.15 ± 0.151	0.991	10.7
FS2-REF (proposed)	JMD (大阪方言話者)	2.88 ± 0.131	3.26 ± 0.153	0.991	10.3
REF	CPJD (大阪方言話者)	4.39 ± 0.105	4.18 ± 0.132	-	7.8

(b) Task 2: Cross-dialect TTS の評価結果

Method	目的話者	Subjective Evaluation		Objective Evaluation	
		自然性 MOS (↑)	方言性 MOS(↑)	COSSIM (↑)	CER (↓)
FS2 (baseline)	JSUT (標準語話者)	3.48 ± 0.114	2.46 ± 0.141	0.990	7.1
FS2-AP (proposed)	JSUT (標準語話者)	3.44 ± 0.100	**3.04 ± 0.156**	0.989	7.9
FS2-REF (proposed)	JSUT (標準語話者)	3.49 ± 0.104	3.11 ± 0.154	0.989	7.6
REF	CPJD (大阪方言話者)	4.08 ± 0.114	4.10 ± 0.130	-	7.8

Reference encoder で用いる PPG を抽出するための ASR モデルとしては，学習済みの Whisper [18] large-v2 モデル [3]を用いた．また，PPG を 音素レベルに aggregation するための音素アライメント情報は Julius [19] で取得した．ベクトル量子化モジュールの量子化クラス数（ALV クラス数）は 4 とした．ALV predictor で用いる PL-BERT には日本語 Wikipedia コーパス [4] によって事前学習された事前学習済みモデル [5]を用いた．また，テキストに標準語のアクセントラベルを自動付与するために OpenJTalk [6] を用いた．TTS モデルには FastSpeech 2 [20] を用い，FastSpeech 2 から出力されたメルスペクトログラムを波形に変換するためのボコーダには，学習済みの HiFi-GAN [21] UNIVERSAL_V1 モデル [7]を使用した．

実験では，目的話者の母方言と同じ方言の TTS のを目的とする “Intra-dialect TTS” および，話者の声質を保ったまま目的話者の母方言と異なる方言の TTS のを目的とする “Cross-dialect TTS” という 2 つのタスクによって提案モデルを評価した．本稿では，Intra-dialect TTS と Cross-dialect TTS における目的話者をそれぞれ JMD コーパスの大阪方言話者，JSUT コーパスの標準語話者として定義した．本実験では以下の 3 つの手法を評価した．

- **FS2**: 通常の FastSpeech2
- **FS2-AP**: ALV を ALV Predictor で予測する提案手法
- **FS2-REF**: ALV を参照音声から抽出する提案手法

FS2 は図 1 のように標準語のアクセントラベルを用いて音声を合成する．FS2-AP は音声の合成時に ALV predictor に入力する方言 ID として大阪方言を指定する．FS2-REF は音声の合成時に CPJD コーパスの音声を参照音声として用いる．

3.2 主 観 評 価

クラウドソーシングを用いて，音声の自然性およびアクセントの大阪方言らしさに関する MOS テストを実施した．各手法の合成音声をランダムに提示し，音声の自然性（人間らしく自然な音声に聴こえるか）およびアクセントの大阪方言らしさ（標準語ではなく大阪方言として自然なアクセントか）を 5 段階で評価させた．Intra-dialect TTS および Cross-dialect TTS ともに，MOS 評価の受聴者数は 35 人，1 人の評価回数は 24 とした．また，先述の 3 つの手法による合成音声の他に参考として CPJD コーパスの自然音声も評価させた．

3.3 客 観 評 価

合成音声の明瞭性と目的話者との話者類似度を評価するための客観評価を行った．話者の類似度は x-vector [22]（話者表現ベクトル）のコサイン類似度によって評価した．具体的には，目的話者による自然音声サンプルのうち学習に使われていない評価用サブセットに含まれている全ての音声から得られる x-vector を平均し，平均化された x-vector と各合成音声から得られた x-vector のコサイン類似度を計算し，それらの平均を計算した．x-vector は JTubeSpeech コーパス [8]という多話者日本語音声コーパスによって学習された学習済みモデル [9]を用いて取得した．音声の明瞭性は文字誤り率 (Character Error Rate: CER) によって評価した．CER の計算は仮名漢字交じり文に対して行った．CER を測るための ASR モデルは Reference encoder で PPG を抽出するために用いた学習済みの Whisper large-v2 モデルを用いた．

（注3）：https://huggingface.co/openai/whisper-large-v2
（注4）：https://dumps.wikimedia.org/
（注5）：https://github.com/kyamauchi1023/PL-BERT-ja
（注6）：https://open-jtalk.sp.nitech.ac.jp
（注7）：https://github.com/jik876/hifi-gan
（注8）：https://github.com/sarulab-speech/jtubespeech
（注9）：https://github.com/sarulab-speech/xvector_jtubespeech

表 2　韻律特徴量として PPG を用いた場合と F0 を用いた場合の prosody transfer の性能比較の結果．太字は手法間に $p < 0.05$ の有意差があったことを示す．

	Subjective Evaluation		Objective Evaluation	
韻律特徴量	自然性 MOS (↑)	方言性 MOS(↑)	COSSIM (↑)	CER (↓)
F0	3.33 ± 0.109	2.87 ± 0.143	0.989	7.4
PPG	$\mathbf{3.49 \pm 0.104}$	$\mathbf{3.11 \pm 0.154}$	0.989	7.6

3.4 実験結果

評価結果を表 1 に示す．まず，表 1(a) に示されている，目的話者を大阪方言話者とした Intra-dialect TTS の評価結果から，Intra-dialect TTS においては FS2 と FS2-AP の間に評価値の有意差が見られなかった．一方で，参照音声入力による prosody transfer によって，アクセントの大阪方言らしさに関する MOS がやや向上する傾向が見られた．また，表 1(b) に示されている，目的話者を標準語話者とした Cross-dialect TTS の評価結果から，Cross-dialect TTS において FS2-AP は FS2 に対してアクセントの大阪方言らしさに関する MOS が有意に高かった．これにより，ALV predictor が大阪方言らしいアクセント表現を学習していることを示し，提案手法が Cross-dialect TTS において合成音声の目的方言らしさを向上させるのに有効であるということが示された．さらに，参照音声入力による prosody transfer によって，FS2 に対してアクセントの大阪方言らしさに関する MOS が有意に向上し，さらに FS2-AP をも上回った．これにより，我々の提案手法により未知の話者による参照音声入力による prosody transfer が可能であることが示された．

3.5 韻律特徴量に関する Ablation study

提案モデルの Reference encoder では，音声から ALV を抽出するための韻律特徴量として PPG を用いた．しかし，音声のアクセント情報を含む特徴量として，従来手法 [3], [4] のように音声の F0 を用いるという手法も考えられる．ここでは，韻律特徴量として PPG を用いることの有効性を評価するため，韻律特徴量として PPG を用いた場合と F0 を用いた場合の比較評価を実施する．実験条件は 3.4 節と同様とした．ただし，Reference encoder で用いる韻律特徴量は話者に依存しない特徴量である必要があるため，F0 は発話単位で正規化した．また，F0 を音素レベルに aggregation する際，無声区間は線形補間によって補間した．なお，F0 の抽出には WORLD [23] を用いた．

評価結果を表 2 に示す．PPG を用いた場合は F0 を用いた場合に対して，音声の自然性およびアクセントの大阪方言らしさに関する MOS が有意に高かった．要因の 1 つとして，F0 を話者非依存な特徴量にするため発話単位で正規化しているが，それだけでは未知話者に対して頑健な韻律特徴量となっていないということが考えられる．

4. おわりに

本稿では，任意の話者による参照音声を用いた Prosody transfer および方言に応じた ALV 予測が可能な多方言 TTS の手法を提案し，Intra-dialect TTS と Cross-dialect TTS という 2 つのタスクによって提案手法を評価した．評価結果から，我々の提案手法が (1) Cross-dialect TTS において合成音声の目的方言らしさを向上させること，(2) 未知の話者による参照音声を用いた Prosody transfer により合成音声の目的方言らしさを向上させることを示した．今後は，ユーザから提供される ALV を教師データとする適応的音声合成 [24] への拡張や，ALV predictor の継続学習 [25] を検討する．

謝辞: 本研究は，JST，ACT-X，JPMJAX23CB の支援を受けたものである．

文　献

[1] Y. Sagisaka: “Speech synthesis by rule using an optimal selection of non-uniform synthesis units”, Proc. ICASSP, New York, U.S.A., pp. 679–682 (1988).

[2] A. S. Cowen, H. A. Elfenbein, P. Laukka and D. Keltner: “Mapping 24 emotions conveyed by brief human vocalization”, American Psychologist, **74**, 6, pp. 698–712 (2019).

[3] K. Yufune, T. Koriyama, S. Takamichi and H. Saruwatari: “Accent modeling of low-resourced dialect in pitch accent language using variational autoencoder”, Proc. SSW, Budapest, Hungary, pp. 189–194 (2021).

[4] 山内一輝, 齋藤佑樹, 猿渡洋：“アクセント潜在変数の予測と制御が可能な tts モデルによる方言音声合成の検討”, 日本音響学会 2023 年秋季研究発表会講演論文集, pp. 1255–1256 (2023).

[5] A. van den Oord, O. Vinyals and K. Kavukcuoglu: “Neural discrete representation learning”, Proc. NIPS, Vol. 31, Long Beach, California, USA, pp. 6309–6318 (2017).

[6] 郡史郎：“日本語のイントネーション–しくみと音読・朗読への応用”, 大修館書店 (2020).

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark and R. A. Saurous: “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron”, Proc. ICML, Proceedings of Machine Learning Research, pp. 4693–4702 (2018).

[8] Y. Lee and T. Kim: “Robust and fine-grained prosody control of end-to-end speech synthesis”, Proc. ICASSP, pp. 5911–5915 (2019).

[9] V. Klimkov, S. Ronanki, J. Rohnke and T. Drugman: “Fine-grained robust prosody transfer for single-speaker neural text-to-speech”, Proc. INTERSPEECH, pp. 4440–4444 (2019).

[10] S. Karlapati, P. Karanasou, M. Lajszczak, S. Ammar Abbas, A. Moinet, P. Makarov, R. Li, A. van Korlaar, S. Slangen and T. Drugman: “CopyCat2: A Single Model for Multi-Speaker TTS and Many-to-Many Fine-Grained Prosody Transfer”, Proc. Interspeech 2022, pp. 3363–3367 (2022).

[11] L. Ma, Y. Zhang, X. Zhu, Y. Lei, Z. Ning, P. Zhu and L. Xie: “Accent-VITS: accent transfer for end-to-end TTS”, **abs/2312.16850**, (2023).

[12] D. P. Kingma and M. Welling: “Auto-encoding variational bayes”, Proc. ICLR (2014).

[13] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng: “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training”, Proc. ICME, pp. 1–6 (2016).

[14] Y. A. Li, C. Han, X. Jiang and N. Mesgarani: “Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions”, Proc. ICASSP, pp. 1–5 (2023).

[15] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari: “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research”, Acoustical Science and Technology, **41**, 5, pp. 761–768 (2020).

[16] S. Takamichi and H. Saruwatari: “JMD: Japanese multi-dialect corpus” (2021).

[17] S. Takamichi and H. Saruwatari: “CPJD corpus: Crowdsourced parallel speech corpus of japanese dialects”, Proc. LREC, Miyazaki, Japan, pp. 434–437 (2018).

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever: “Robust speech recognition via large-scale weak supervision” (2022).

[19] A. Lee, T. Kawahara and K. Shikano: “Julius — an open source real-

time large vocabulary recognition engine", Proc. EUROSPEECH, Aalborg, Denmark, pp. 1691–1694 (2001).

[20] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu: "FastSpeech 2: Fast and high-quality end-to-end text to speech", Proc. ICLR, Vienna, Austria (2021).

[21] J. Kong, J. Kim and J. Bae: "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis", Proc. NeurIPS, Vol. 33, Virtual Conference, pp. 17022–17033 (2020).

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur: "X-Vectors: Robust dnn embeddings for speaker recognition", Proc. ICASSP, pp. 5329–5333 (2018).

[23] M. Morise, F. Yokomori and K. Ozawa: "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications", IEICE Trans. Inf. Syst., **E99-D**, 7, pp. 1877–1884 (2016).

[24] K. Fujii, Y. Saito and H. Saruwatari: "Adaptive end-to-end text-to-speech synthesis based on error correction feedback from humans", Proc. APSIPA ASC, Chiang Mai, Thailand, pp. 1702–1707 (2022).

[25] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning and C. Finn: "Direct preference optimization: Your language model is secretly a reward model", Proc. NeurIPS (2023).