DNN による方言テキストからの方言音声合成の検討

Examination of dialect speech synthesis from dialect text using deep neural network

妹尾 和真

Kazuma Senoo

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では，Deep Neural Network （DNN） を用いて方言テキストから方言音声を合成する Text to Speech (TTS) モデルの作成を目指す．本報告では，大阪方言をアクセントラベルなしの音素と音声データとを用いて学習し，アクセントラベルの必要性の検討を行った．

1 はじめに

日本には多様な方言が存在するが，現在マスメディアやインターネットが広く普及したことに伴い，主に標準語（東京方言）が使用され各地域で古くから使用されてきた方言は，近年衰退傾向にある．方言という文化の保存という点で方言の音声合成モデルの作成は非常に有益な手段であると考える．また，方言の音声合成モデルの作成は，ドラマや映画などでの方言指導にも利用でき，方言の音声合成モデルの作成は方言の保存以外にも意義があると考える．

しかし，方言は標準語と違いアクセントラベルを付与するためのアクセント辞書がないという問題がある．方言の音声合成の先行研究は，ほとんどがアクセントラベルを付けて行うものであった．ただ，方言特有のアクセントラベルを手動でつけるとすると 1 つの方言に対して非常に労力がかかり，多様な方言の保存という目的には適していないと考えた．また，アクセントラベルとしてアクセント潜在変数 （Accent Latent Variable:ALV） を与えるという方式 [1] が提案されており，音素へ手動でのアクセントラベルの付与をしない学習方式も提案されている．その方式で学習した TTS モデルが推論したものは，アクセントの方言らしさに関する MOS が有意に高かった．そこで，アクセントラベルなしで TTS 合成を行った場合でも方言らしさの表現は可能だと考え，図 1 のような方言のテキストから変換したアクセントラベルなしの音素と音声から取り出した音声特徴量を学習データとして与える方式で方言音声合成モデルの作成の検討を行うことにした．

本報告では，VITS[2] を用いて大阪方言テキストの音素と大阪方言音声を学習データとして与え，大阪方言音声の TTS モデルの作成の検討を行った．

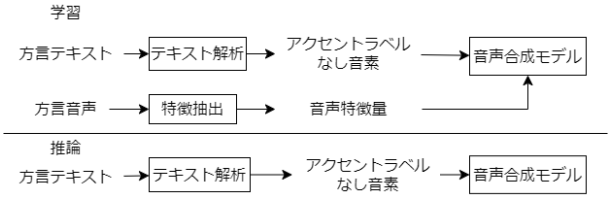

図 1: アクセントなしの音素と音声での方言音声の合成の概要

2 学習データ

本研究では，方言音声合成モデルの学習のためのデータセットとして JMD コーパス [3] を用いる．JMD コーパスは大阪方言と熊本方言のネイティブ方言話者による各 1300 発話の各 2 時間以上の読み上げ音声からなる方言音声コーパスである．本報告では，大阪方言の発話データを学習用 1100 発話，検証用 100 発話，評価用 100 発話と分けて用いた．また，発話内容のファイルはかな漢字交じり文で書かれており，地名や店名など固有名詞も含まれていたため，音素へ変換する際に誤った読み方での音素に変換されてしまったため，すべての漢字のひらがなへの書き換えをひらがな変換ツールと手動で行った．

3 大阪方言音声合成モデルの作成

VITS は，HiFi-GAN V1 の Decoder と GAN、Glow-TTS から Text Encoder や音素継続時間推定、アライメントを組み合わせ改良したもので，音素と音声の潜在変数を導入し，VAE により損失関数を介して接続して学習する．今回利用した VITS では，grapheme-to-phoneme (g2p) を行うプログラムが英語テキストに対するものしか実装されていなかったため，日本語テキストの音素変換を実装するために pyopenjtalk[4] の g2p を用いて行い，音素記号に関しても日本語の音素記号をシンボルに追加した．それらの処理をしたのち，学習データを用いて大阪方言音声の TTS モデルの作成を行った．

4 メルスペクトログラムの比較

作成した大阪方言音声の TTS モデルを用いて評価用データの文章での音声を生成した．生成した音声と評価用音声のメルスペクトログラムの比較を行い，大阪方言音声の TTS モデルの評価を行った．図 2 と図 3

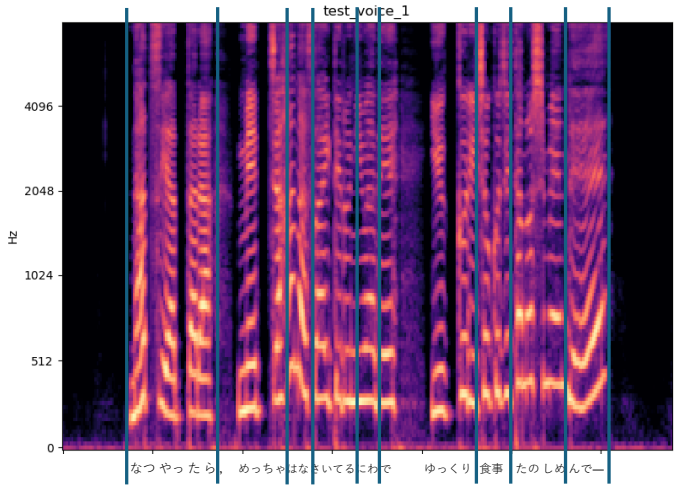

図 2: 評価用音声のメルスペクトログラム

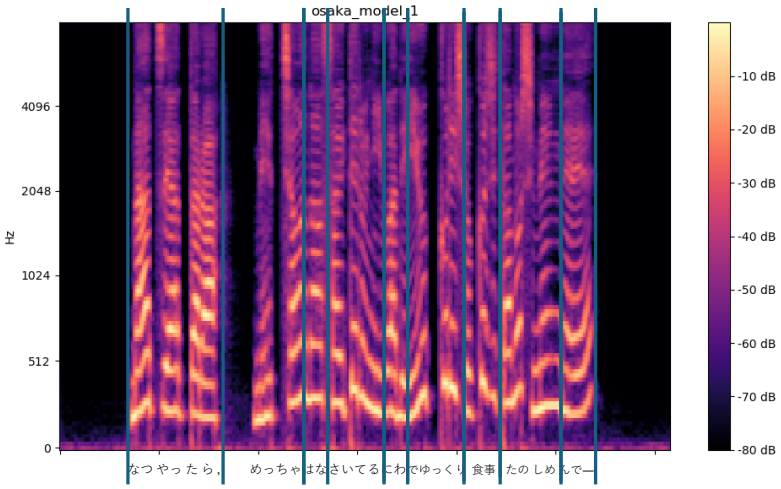

図 3: 大阪方言 TTS モデルの生成音声のメルスペクトログラム

のメルスペクトログラムは，「夏やったら，めっちゃ花咲いてる庭でゆっくり食事たのしめんでー」という文章のものである．図 2 と図 3 の前半と後半部分は類似性がみられるが，中盤はあまり類似性がないように見える．このことより大阪方言の「〜やったら」や「〜やで」といった語尾については，今回の学習方式でも表現可能だと思われる．しかし，図 2 と図 3 の中盤の青線で囲った箇所は単語のアクセントであり，大阪方言の単語のアクセントはあまり再現できていなかった．「花」のアクセントが作成したモデルでは，標準語の花のアクセントとなっており，大阪弁での「花」のはが高くなるアクセントは再現できていない，また反対に「庭」と「食事」と言っている箇所は評価用音声では，標準語のアクセントとなっていたが，作成した TTS モデルでは，どちらも先頭の文字のアクセントが高くなる表現がされていた．「庭」と「食事」の箇所に関しては大阪方言のネイティブ話者ではない筆者は，作成した TTS モデルのアクセントがより大阪方言のように聞こえた．今回の作成した TTS モデルでは，ネイティブ話者にとってはエセ関西弁のような違和感のあるものであるが，非ネイティブ話者からは，方言らしさというものを表現できているものではないかと考える．「花」の部分は，大阪方言特有の発音であり，「花」以外にも多数存在し改善するためには方言データを集める必要があるが，「庭」や「食事」といった標準語のアクセントは，標準語の音声から学習できるため標準語の学習と方言の学習を組み合わせることで改善できるのではないかと考える．

5 まとめ

本報告では，VITS を用いて大阪方言の音素と音声データを学習データとした方言音声 TTS モデルの作成について述べた．本実験で．VITS での大阪方言のアクセントラベルなしでの音声生成では．学習データが少量であったが，大阪方言特有の語尾に方言性が少しみられた．しかし，花などの大阪特有のアクセントを持つ単語のアクセントの表現は再現することができなかった．アクセントの表現が再現できなかったのは学習データの量の要因が大きいと考え，今後は，VITS を用いて，学習データが多数存在する標準語音声でのアクセントラベルありとアクセントラベルなしの音声合成をし，標準語でのアクセントラベルありとなしの比較実験を行い，VITS でのアクセントラベルなしでの学習の評価を検討する．

参考文献

[1] 山内 一輝, 斎藤 佑樹, 猿渡 洋, “VQ-VAE に基づく解釈可能なアクセント潜在変数を用いたた方言音声合成, 電子情報通信学会技術研究報告, vol. 123, no. 403, pp. 220-225, 2024

[2] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech “, Proceedings of the 38th International Conference on Machine Learnibi 2021

[3] 高道 慎之介, 猿渡 洋, ”JMD: Japanese multi-dialect corpus”, https://sites.google.com/site/shinnosuketakamichi/publication/research-topics/jmd_corpus.

[4] ” pyopenjtalk”, https://r9y9.github.io/pyopenjtalk/.