少量データを用いた
音声合成におけるボコーダの一考察

奈良女子大学
生活情報通信科学コース
髙田研究室

少量の音声から学習データの拡張

- 少量の音声データの場合でも高性能な音声認識システムを作成
- 合成音声の利用で、認識できなかった単語学習の追加

音声認識システムの問題

通常の音声認識システムは..

- 学習に大量の音声データ+テキストデータが必要
- 少量の音声データだと認識率が低い

高性能な音声認識システムを作ろうとすると
音声データだけで10時間以上のものが
必要になることも...

目標

- 合成音声を人間の音声（自然音声）の代わりに使用
- 少量学習データから合成音声を生成
- 最適な音声合成ボコーダを調査
- 自然音声との精度を比較

音声生成手順

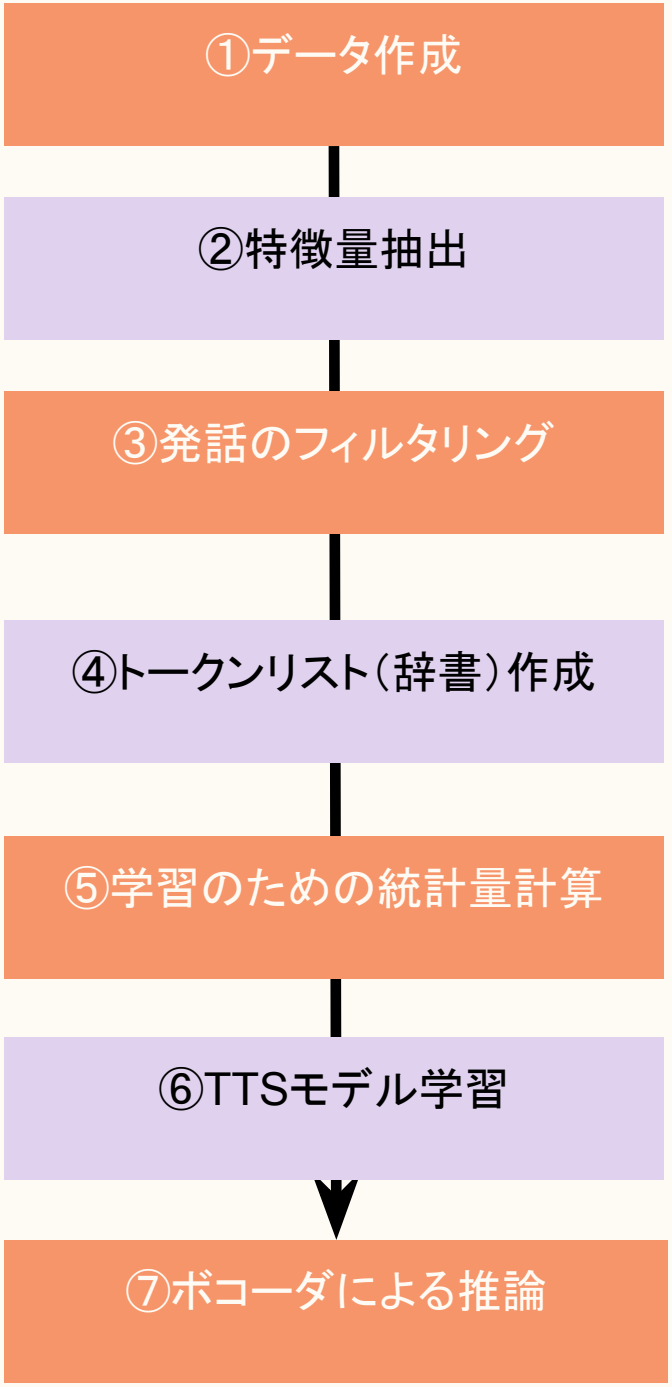
TTS (テキストから音声を作成するツール)をもとに開発

手順1～6：Tacotron2

- 注意機構を含むEncoder-Decoderモデル
- LSTMによる自己回帰
- 実装：ESPnet2を使用

手順7：ボコーダ

ボコーダとは…
音声合成で学習した特徴量から音声波形を生成する部分
音声合成の品質や速度を決める重要な部分



使用ボコーダ

- Griffin-Lim
  - 非DNNモデル
  - 振り幅スペクトログラムを元に波形を生成

- Parallel WaveGan
  - Parallel(並列→高速)+WaveNet(自己回帰型)+GAN
  - WaveNetと同等以上の品質+並列処理による高速処理

使用ボコーダ

- Hifi-Gan
  - Generator に Multi-Receptive Field (MRF) fusion を導入
  - 異なる受容野を用いた多数の residual block の和として音声波形を生成
- VITS
  - E2E-TTS
  - 敵対的学習の中で正規化フローと変分推論を利用
  - 確率的継続長予測器によって、多様なリズムの音声を生成
  - GlowTTSエンコーダ+Hifi-GANボコーダ

実験①概要

▸ 単一話者での転移学習を実施

データセット

- つくよみちゃんコーパス（100文，10分58秒）

⇒固有名詞や難しい熟語が多い（例：セーヌ川、橋脚等）

- ITAコーパス（100文，10分32秒）

⇒日常会話が多い

事前学習済みモデル

JSUTコーパス（7696文、10時間）で学習したモデル

実験①概要

- 単一話者での転移学習を実施

モデル	Epooh数	Batch bins	Vocoderのfine-tuning有無
Tacotron2＋Griffin-Lim	1000	3750000	なし(事前学習済みを使用)
Tacotron2＋ParallelWave Gan	1000	3750000	なし(事前学習済みを使用)
Tacotron2＋Hifi-Gan	1000	3750000	なし(事前学習済みを使用)
VITS	50	1000000	あり

実験 ② 評価手法

▸モデルの客観評価手法

MCD（メルケプストラム歪み）：MCEPのユークリッド距離から計算

$$\mathrm{MCD} = \frac{10\sqrt{2}}{\ln 10} \cdot \left(\sum_{i=1}^{n} \sqrt{(\mathrm{MCEP_i} - \widehat{\mathrm{MCEP_i}})^2} \right)$$

Log F0 RMSE（二乗平均平方根誤差）：対数をとったF0配列のRMSE

$$\mathrm{LogF0RMSE} = \sqrt{\frac{1}{n} \cdot \sum_{i}^{n} (\log \mathrm{F0_i} - \log \hat{\mathrm{F0}}_i)^2}$$

WER（単語誤り率）：音声認識器Whisperで認識した単語の誤り率を計算

$$\mathrm{WER} = \frac{(\textit{挿入単語数} + \textit{置換単語数} + \textit{削除単語数})}{\textit{正解語数}}$$

音声認識器Whisper

68万時間の多言語・マルチタスク教師付きデータで学習させた音声認識モデル

エンドツーエンドをベースとしたアーキテクチャ

結果

- MCD, LogF0RMSE, WER を算出
- 各40個の文章を推論し、音声の平均 ±95%信頼区間を求めた

つくよみちゃんコーパス

モデル	MCD	LogF0RMSE	WER*100
Griffin-Lim	11.64±0.99	0.22±0.05	30.42
ParallelWaveGan	8.07±0.58	0.19±0.03	27.15
Hifi-Gan	8.39±0.45	0.21±0.04	33.39
VITS	3.78±0.05	2.17±0.06	10.14
オリジナル音声	—	—	5.29

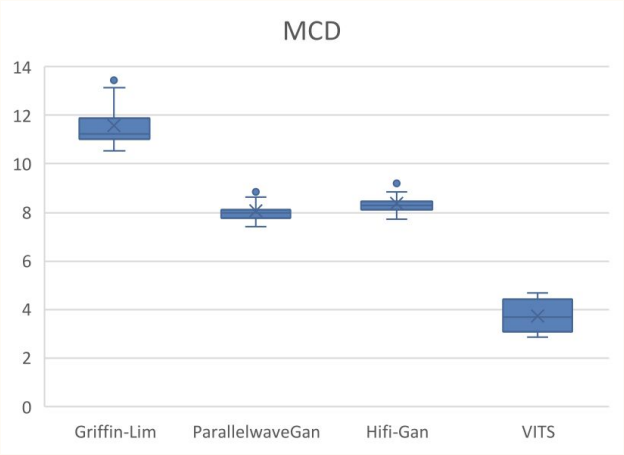
ITAコーパス

モデル	MCD	LogF0RMSE	WER*100
Griffin-Lim	11.41±0.87	0.34±0.08	26.56
ParallelWaveGan	7.96±0.52	0.21±0.05	24.54
Hifi-Gan	8.15±0.42	0.25±0.06	26.41
VITS	3.62±0.05	2.18±0.10	7.21
オリジナル音声	—	—	4.66

結果（つくよみver）

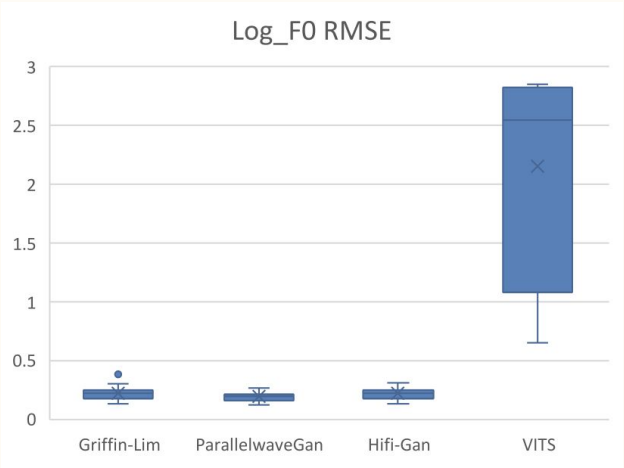
MCD

- VITSが一番値が小さい
- VITS平均値＊2＜他モデル



Log_F0 RMSE

- VITSの平均値，分散が一番大きい
- PWGが最も小さい
- Griffin-Lim, PWG, Hifi-Ganの差はほとんど無い



<u>精度</u>

<u>E2Eモデル>非E2Eモデル</u>

WERの結果の考察

MCD：Griffin-Lim ＞ Hifi-GAN ＞ PWG ＞ VITS

WER：Hifi-GAN ＞ Griffin-Lim ＞ PWG ＞ VITS

つくよみちゃんコーパス

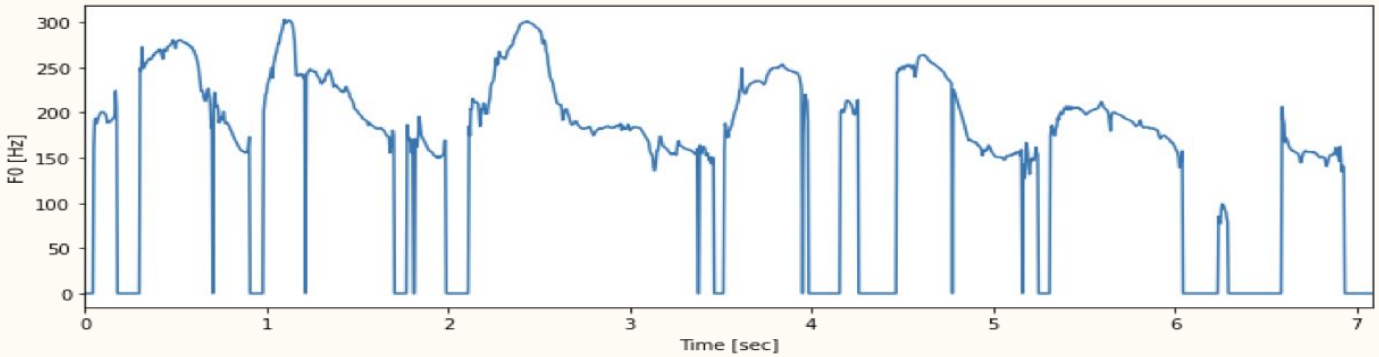
モデル	Griffin-Lim	PWG	Hifi-GAN	VITS	自然音声
MCD	11.64±0.99	8.07±0.58	8.39±0.45	3.78±0.05	-
WER＊100	30.42	27.15	33.39	10.14	5.29

ITAコーパス

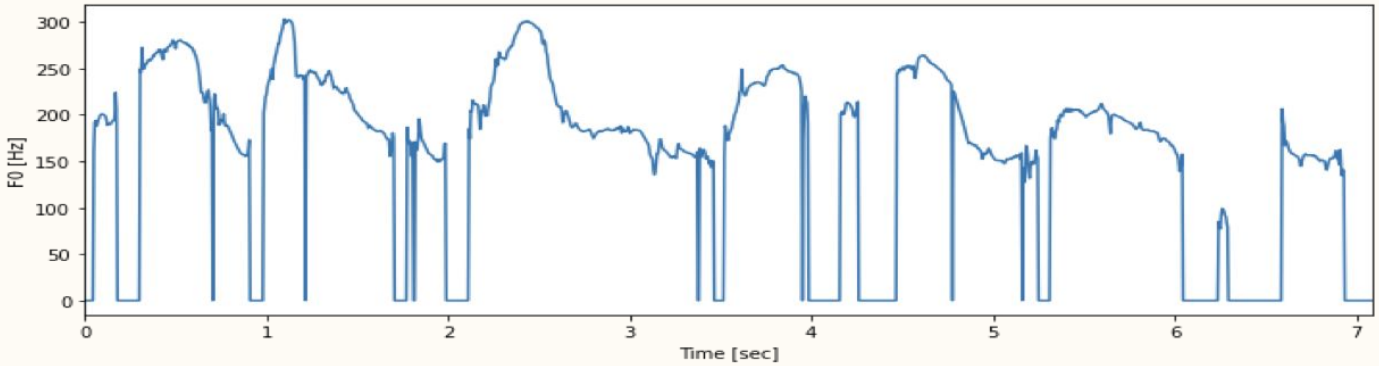
モデル	Griffin-Lim	PWG	Hifi-GAN	VITS	自然音声
MCD	11.41±0.87	7.96±0.52	8.15±0.42	3.62±0.05	-
WER＊100	26.56	24.54	27.41	7.21	4.66

GANの比較の考察（音声波形）

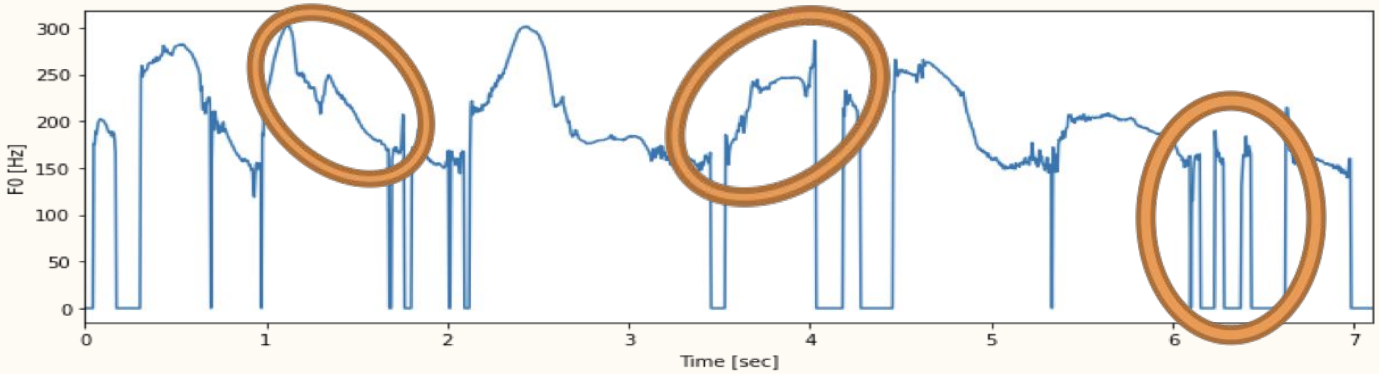
▶自然音声



▶ParallelWaveGan



▶Hifi-GAN



合成音声で認識率が悪い単語の特徴

▸全ての合成音声モデルで認識率が悪い

- 母音が続く（例：王（おう）、降雨（こうう）等）→**音が消えやすい**
- 「ら」行（例：拾い→ひどい、狙う→ねがう等）→**音が濁りやすい**

▸Tacotron2を使用したモデル（3個）で認識率が悪い

- 半音を含む（例：ウェイトレス→レイトレス、山脈→さんらく等）→**音が単純化する**

▸Griffin-Limボコーダ

- 「ち」から始まる（例：長母音（ちょうぼいん）、鎮痛（ちんつう）等）→**音が濁りやすい**
- 「む」→**「ん」に変化**しやすい

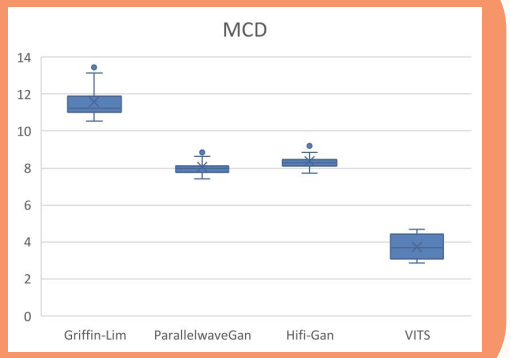
▸Hifi-GANボコーダ

- 濁音（例：軍→うん、擬古→いこ）→**音が単純化する**
- 中間音が入る（岩壁→がんぺいき、化粧→けいしょう等）→**音が間延びする**

▸ParallelWaveGanボコーダ

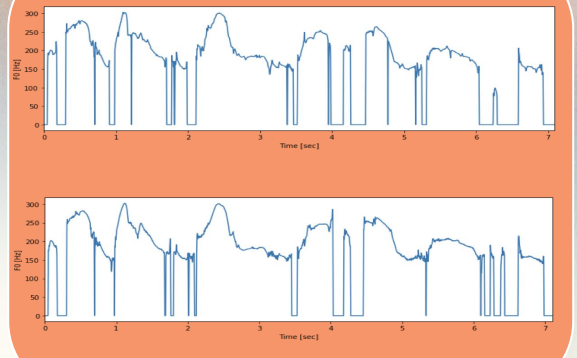
- 濁音が続く（例：お辞儀（おじぎ）、語尾（ごび）など）→**発音が不明瞭**になりやすい

まとめ



MCD：Griffin-Lim ＞ Hifi-GAN ＞ PWG ＞ VITS

WER：Hifi-GAN ＞ Griffin-Lim ＞ PWG ＞ VITS



E2Eモデルの精度
＞＞
非E2Eモデルの精度

MCDの評価
≠
音声認識の精度

発音が明瞭
＝
認識率が高い