G-6

GAN による日本語音声の生成に関する研究

A Study on Generating Japanese Speech Using Generative Adversarial Network

○齋藤祐哉[1]，　関弘翔[2]，　細野裕行[2]
*Yuya Saito[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: In recent years, deep learning has been introduced and developed in speech synthesis technology. However, the generated voice has problems such as quality and real-time performance. In this research, we examine a method for generating Japanese speech with high quality and high speed in Text-To-Speech.

1．はじめに

近年，音声を利用したアプリケーションが増加している．そこに使用される音声合成技術には深層学習が導入されはじめ発展してきた．しかしながら，生成される音声には品質やリアルタイム性といった課題が挙げられる．

本研究では，音声合成技術のひとつである Text-To-Speech（TTS）において，学習効率及び速度面で優位な GAN（Generative Adversarial Network：GAN）を用いた手法による高品質な日本語音声を生成することを目的とする．

2．Text-To-Speech の概要及び関連研究

TTS とは，テキストから音声を作り出す技術である．Fig.1 に TTS の流れを示す．

テキストから音声を生成するまでに 3 つのモジュールが存在する．テキスト解析器（Text Parser）では日本語の文章を音素などの言語特徴量（Linguistic Features）へと変換する．音響モデル（Acoustic Model）では言語特徴量をメルスペクトログラムなどの音響特徴量（Acoustic Features）へと変換する．ボコーダー（Vocoder）では音響特徴量を音声波形へと変換する．

TTS において，主流の実装とされているのが，音響モデルに Tacotron2[1]，ボコーダーに WaveGlow[2]を採用したものである．この手法では人間の肉声に匹敵する音声を生成することができるが，生成には時間がかかるため，リアルタイム性には課題が残る．特に WaveGlow は大容量モデルであり，GPU を用いることで比較的高速に生成できるが GPU メモリに制約のある環境での動作は困難である．

また多くの音響モデルは英語を対象に構築されているため，日本語生成においては品質低下の恐れがある．

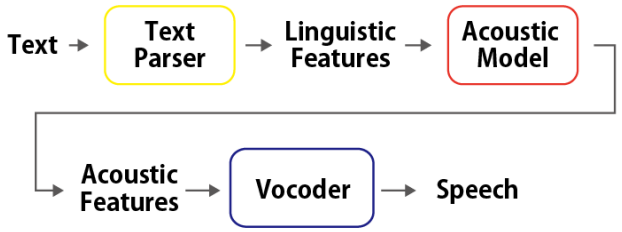

Figure1. Flow of TTS

3．日本語音声の生成手法の検討

本研究では，音響モデルには AlignTTS[3]，ボコーダーには Multi-Band MelGAN（MB-MelGAN）[4]を採用した．どちらのモデルも品質を維持しつつ，前述の Tacotron2，WaveGlow よりも高速に処理を行うことができるが，これらを組み合わせた TTS の報告例は見受けられない．

各モデルにおいて日本語音声で学習させ，提案手法による生成音声の品質，リアルタイム性を検討する．音響モデルの学習には高道らにより公開されている一人の女性話者による 10 時間分の音声データである JSUT コーパス[5]を使用する．音響モデルにおいては生成した音響特徴量と実音声での音響特徴量の比較を行う．ボコーダーにおいては，その性質上，英語を対象としたモデルでも日本語音声の生成には問題ないが，JSUT コーパスでの学習前後における生成音声の品質を比較する．また，提案手法と既存手法による音声の生成速度を比較する．

参考文献

[1] J.Shen, et al:”Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,”IEEE ICASSP 2018 , pp.4779-4783(2018)
[2] R.Prenger, et al:” WaveGlow: A Flow-based Generative Network for Speech Synthesis, ”IEEE ICASSP 2019 , pp.3617-3621(2019)
[3] Z.Zhen, et al:”Aligntts: Efficient feed-forward text-to-speech system without explicit alignment,” IEEE ICASSP 2020, pp.6714-6718(2020)
[4] G.Yang ,et al:”Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech,”IEEE SLT 2021, pp.492-498(2021)
[5] R.Sonobe, et al:”JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354(2017)

１：日大理工・院（前）・情報，２：日大理工・教員・情報