G-6

Text-To-Speech における日本語音声の生成に関する検討
生成音声の品質向上および感情表現
A Study on Generating Japanese Speech of Text-To-Speech
Improvement of Quality of Generated Speech and Expression of Emotion

○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 our previous study, we employed AlignTTS and Multi-Band MelGAN as the model used for TTS, and achieved speech generation much faster than existing methods. However, the generated speech contained noise and silence, and the quality of the speech remained a problem. In this report, we examine how to improve the quality of the generated speech by improving the model and the dataset. In addition, we examined the possibility of expressing emotion in the generated speech by fine-tuning AlignTTS in the dataset for emotional speech.

１．はじめに

近年，音声合成技術には深層学習が導入され発展が著しい．音声合成技術のひとつにテキストから音声を生成する Text-To-Speech（TTS）がある．TTS にはテキスト解析器，音響モデル，ボコーダの 3 つのモジュールが存在し，このうち音響モデル，ボコーダには深層学習モデルが採用され品質向上，リアルタイム性の向上を目的とした研究が盛んに行われている．

文献 [1]で著者らは，TTS に用いるモデルに AlignTTS[2]，Multi-Band MelGAN[3]（MB-MelGAN）を採用し既存手法を大きく上回る速度での音声生成を実現した．しかしながら生成される音声にはノイズや無音部分が存在し，品質に課題が残る結果となった．

本報告ではモデルの改良，データセットの改良を行うことで生成音声の品質向上を検討した．

また，感情音声用データセットで AlignTTS を Fine-Tuning させることで生成音声で感情表現が行えるか検討した．

２．生成音声の品質改善

文献 [1]で生成した音声には不要な無音部分やノイズ，発音性が失われているといった課題があった．そこでデータセット，AlignTTS，MB-MelGAN を改良することで課題解決を図った．

２．１．データセットの改良

本研究ではデータセットには東京大学の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[4]を使用した．このデータセットには各データの最初と最後に無音部分が存在し，話者の息継ぎやボタンを押す音といったものも含まれていた．このような部分はモデルの学習に悪影響を与えると考え，発話部分のみを切り取り無音部分の消去を行った．

２．２．AlignTTS の改良

AlignTTS はテキスト情報からメルスペクトログラムを生成するモデルである．ここで入力されるテキスト情報は日本語テキストを音素表記へと変換されたものである．Fig.1 にテキストを音素表記へと変換した例を示す．テキスト解析器から出力される音素表記には英小文字と大文字が混ざるが，AlignTTS の文字識別部には英大文字のみとなっており，入力されたテキストをすべて英大文字へと変換していた．英大文字と英小文字に区別をつけることで発音性が向上できると考え，AlignTTS の文字識別部に英小文字を追加し，音素表記そのままを入力とするように変更した．

Figure 1. Example of phoneme

２．３．MB-MelGAN の改良

AlignTTS および，MB-MelGAN は学習前に音声データをそれぞれの前処理でメルスペクトログラムへと変換している．その際，対数の底に違いがあったため，同一音声の学習データを比較すると Fig.2 に示すように大きな違いが生じていた．本研究の TTS において MB-MelGAN には AlignTTS で生成されたメルスペクトログラムが入力される．この学習データの違いが生成音声のノイズに影響すると考え，MB-MelGAN の学習データを AlignTTS のデータと統一した．

１：日大理工・院（前）・情報，２：日大理工・教員・情報

```
==================================
dataset of AlignTTS
tensor([[ -5.9476,  -5.9455,  -5.0718,  ...,  -3.2708,  -3.3946,  -3.6957],
        [ -6.3682,  -5.4866,  -5.2026,  ...,  -6.3787,  -6.2487,  -5.9850],
        [ -5.7202,  -5.5704,  -5.7798,  ...,  -8.0930,  -6.7239,  -5.8535],
        ...,
        [ -6.9969,  -7.0016,  -7.1710,  ...,  -9.8945, -10.6272,  -8.9010],
        [ -6.6292,  -6.6477,  -6.5915,  ..., -10.1292, -10.3193,  -8.8820],
        [ -6.0836,  -6.0233,  -6.1644,  ..., -10.2204, -10.3494,  -8.8831]])
torch.Size([80, 275])
==================================
```

(a) Train data of AlignTTS

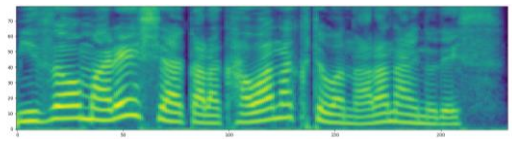
```
==================================
dataset of MB-MelGAN
tensor([[-2.5830, -2.7657, -2.4843,  ..., -3.0387, -2.8790, -2.6421],
        [-2.5821, -2.3828, -2.4192,  ..., -3.0408, -2.8871, -2.6159],
        [-2.2026, -2.2595, -2.5101,  ..., -3.1143, -2.8627, -2.6772],
        ...,
        [-1.4205, -2.7702, -3.5147,  ..., -4.2971, -4.3991, -4.4387],
        [-1.4742, -2.7138, -2.9202,  ..., -4.6154, -4.4816, -4.4947],
        [-1.6050, -2.5993, -2.5421,  ..., -3.8657, -3.8574, -3.8579]])
torch.Size([275, 80])
==================================
```

(b) Train data of MB-MelGAN

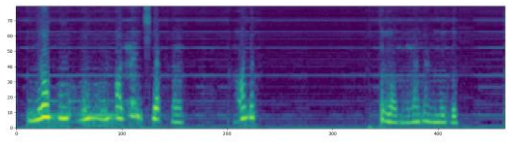
Figure 2. Difference of train data

２．４．生成結果

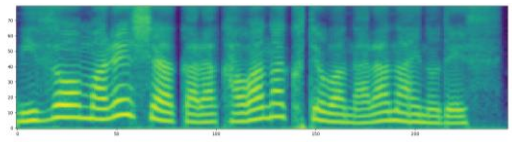
Fig.3 に生成結果を示す．改良後の生成音声 Fig. 3 (c)は Fig.3(b)にあったノイズや無音部分等の問題点もなくなり，学習に使用した音声 Fig.3(a)と比較しても遜色ないレベルで生成することができた．



(a) Target



(b) Our previous generation rsult

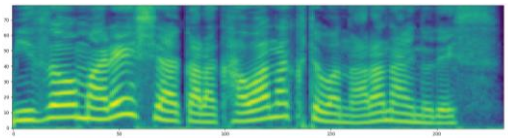


(c) Generated speech

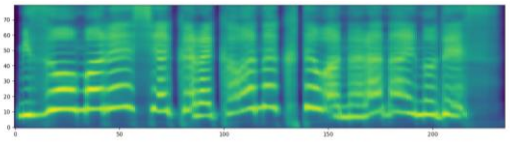
Figure 3. Result of generation speech

３．感情音声の生成

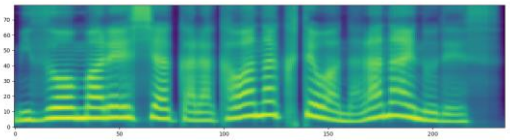
TTS において生成音声の中に感情表現するには，その感情の音声でモデルを学習させることが一般的である．感情表現用のデータセットには日本声優統計学会から配布されている声優統計コーパス[5]を使用した．声優統計コーパスは，独自に構築された 100 種類の音素バランス文を 3 人の女性声優が angry・happy・normal3 パターンの感情で読み上げた音声データである．このうち，1 名の話者による angry，happy 音声それぞれで AlignTTS を Fine-Tuning させて感情表現ができるか検討した．Fine-Tuning とは学習済みモデルの重みを初期値としてモデルを再度学習させる手法である．



(a) before Fine-tuning



(b) angry



(c) happy

Figure 4. Generated Mel-spectrogram of emotional speech

Fig.4 にメルスペクトログラムの生成結果を示す．生成した音声には学習した感情によって口調や声の高さに変化が生まれ，感情表現を変化させることができた．しかしながら，happy においては通常音声と声の高さが大きく異なるためか，ノイズが発生した．また，JSUT コーパスと声優統計コーパスとでは話者が異なるため，Fine-Tuning 前後で話者性が変化した．

４．まとめ

本報告では生成音声の品質向上および感情音声の生成を検討した．生成音声の品質についてはアクセント等に課題が残るが人間の音声と遜色ないレベルまで向上した．感情音声については生成音声で感情表現は実現したが学習データの都合上、話者性まで変換される結果となった．

今後は，生成音声のさらなる品質向上や話者性を変えずに音声の感情を変化させる手法の検討を行う．

参考文献

[1] 齋藤他，「GAN による日本語音声の生成に関する研究」，令和 3 年度日本大学理工学部学術講演会予稿集，G-6，p.352，(2021)
[2] Z.Zhen, et al:”Aligntts: Efficient feed-forward text-to-speech system without explicit alignment,” IEEE ICASSP 2020, pp.6714-6718(2020)
[3] G.Yang ,et al:”Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech,”IEEE SLT 2021, pp.492-498(2021)
[4] R.Sonobe, et al:”JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354(2017)
[5] y_benjo and MagnesiumRibbon, “Voice-actress corpus,” http://voice-statistics.github.io/