手軽に製作可能かつ、調整可能で品質の高い テキスト音声合成システムの開発

大阪電気通信大学高等学校 工学科3年 芦田 裕飛

1. モチベーション

- 近年、音声合成分野の研究は深層学習の影響により急激に進んでおり、様々な技術が登場している
- これらの技術を活用し、個人でもテキスト音声合成ソフトウェアを制作出来るようになっている
- 例として2021年に**VOICEVOX**という無料で使えるテキスト音声合成ソフトウェアが登場し、声の調整が可能な仕様となっているが、構造が複雑である(複数のパーツがあり深層学習モデルの学習が大変)
- OSSとして公開されているVOICEVOXのUIを用いて調整可能な音声合成モデルを作成し、作成手法を広め、誰でも作れるようにし、**誰でも簡単に声の保存や共有が出来るようにしたい**

図1. 調整可能な音声合成ソフトVOICEVOXの全体UI

図2. VOICEVOXの イントネーション調整UI

図3. VOICEVOXの 音の長さ(音素長)調整UI

2. 音声合成の技術概要

- 近年のテキスト音声合成は、総じて以下のようになっており、ある程度一貫して深層学習モデルを学習できるようになっている

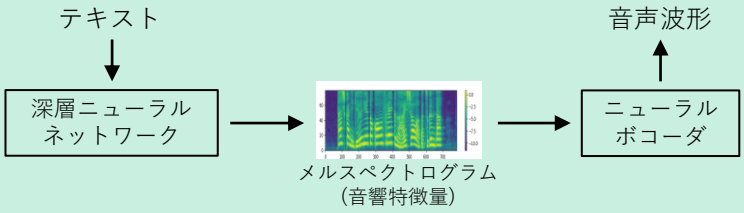

図4. 近年の音声合成の概要的な構造

- 今回、メルスペクトログラムを推論する深層ニューラルネットワークについては、調整可能な構造を持つFastSpeech2という技術を用いる[文献3]

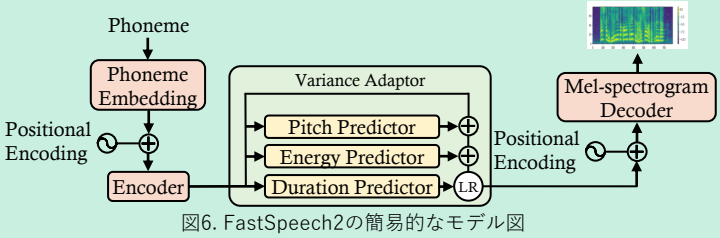

図6. FastSpeech2の簡易的なモデル図

- ニューラルボコーダについては、今回はFre-GANという敵対的生成ネットワークを用いる[文献4]
- 文献5,6を代表とする研究等向けに公開されているデータセットを集め計11005文の音声を学習させた

3. 実験

- 実験過程等は長いため、行った内容を列挙する
- 音質の向上
  - 学習音声のサンプリングレート変更(22.05kHz->48kHz)
  - JSUTコーパスを用いた転移学習(ファインチューニング)
  - 先行研究[文献7]に基づくTransformerのConformerでの置き換え
    - 詳細な実験の末、Decoder部分のみ置き換えた
- 推論精度の向上
  - 先行研究報告[文献8]に基づく韻律(アクセント)の導入
  - 教師データの改善
  - 不要なエナジー推論・埋め込みを削除

4. 結果

- 最終的に以下のモデル図の構成で、そこそこ品質が高く、かつVOICEVOXとも連携が取れる形になった

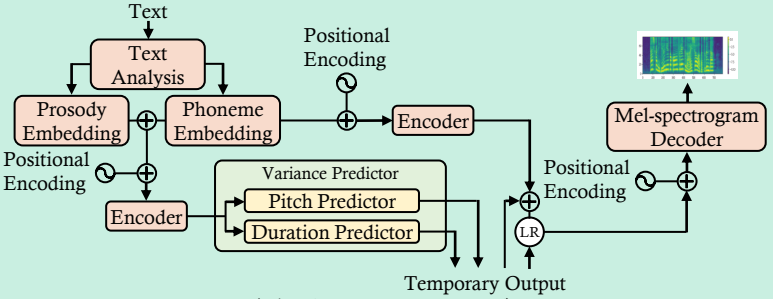

図7. 改造したFastSpeech2のモデル図

図8. VOICEVOXのUI上で動作する、今回制作した音声合成モデル

5. 今後の課題

- 合成音声にノイズが混じっているなど、品質向上の余地が残っている
- パラメータの調整やモデルの更なる改善を試したい

6. 参考文献・謝辞

[1] 山本龍一, 高道慎之介, Pythonで学ぶ音声合成, インプレス社, 2021.
[2] VOICEVOX - 無料で使える中品質なテキスト読み上げソフトウェア, Kazuyuki Hiroshiba, https://voicevox.hiroshiba.jp.
[3] Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, et al., “FastSpeech 2: Fast and high-quality end-to-end text-to-speech,” arXiv preprint arXiv:2006.04558, 2020.
[4] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, Seong-Whan Lee, “Fre-GAN: Adversarial Frequency-consistent Audio Synthesis,” arXiv preprint arXiv:2106.02297, 2021.
[5] Ryosuke Sonobe, Shinnosuke Takamichi and Hiroshi Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint, 1711.00354, 2017.
[6] 小口純矢, 金井郁也, 小田恭央, 齊藤剛史, 森勢将雅, ITAコーパス：パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価,情報処理学会研究報告, vol. 2021-MUS-131, no. 31, pp. 1-6, 2021.
[7] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, Yuekai Zhang, “Recent Developments on ESPnet Toolkit Boosted by Conformer,” arXiv preprint arXiv:2010.13956, 2020.
[8] 藤井 一貴, 齋藤 佑樹, 猿渡 洋, 韻律情報で条件付けされた非自己回帰型End-to-End日本語音声合成の検討, 情報処理学会研究報告, 2021-SLP-138, No. 16, pp. 1-6, 2021.
最後に、利用させていただいたデータセットの公開者・関係者の皆様(あみたろの声素材工房様、つくよみちゃん(夢前黎)様、梅本らく様、RiRi様、ちゅき様、小口純矢様、カノン様、松風様、SSS合同会社様、日本声優統計学会様)にこの場をお借りしてお礼申し上げます。

https://github.com/y-chan/FastSpeech2