4S-07

話速のモデル化に基づく、自然なボイスクローニングの実現 *

秦哲 (法政大学情報科学部), 伊藤克亘 (法政大学情報科学部)

1 まえがき

現在、一人暮らしのひとが多くなり、一人暮らしの孤独感を耐えられない人もいる。令和二年の国勢調査によると「65 歳以上人口のうち，単独世帯の人口は 671 万 7 千人となっている。また，65 歳 以上人口に占める割合は 19.0 %となっており，65 歳以上人口の約 5 人に 1 人が一人 暮らしとなっている」という。一人暮らしの高齢者にとっては、SNS などの使い方に慣れていないかもしれない。音声アシスタントの硬い声をよりリアルな声に切り替えることで、より家らしい感じがするだろう。

TTS(Text To Speech) 技術で人間がよく使う交流手段で情報を伝達することできる。音声アシスタントやナビゲーションアプリなどでよく使っている。しかし、既存の TTS 技術では、設定されたいくつかの音声が使われることが多い。 ボイスクローニング (Voice Cloning) はこの問題を解決する方法の一つである。ボイスクローニングとは、話す声から特徴を抽出し、話者の声で話す TTS を作る技術である。

2 先行研究

先行研究では、短い音声サンプルから類似の音声を生成できるボイスクローニングがある [1]。約 15 秒ぐらいのサンプルから話者の特徴を抽出することができます。 また、より長いサンプルでより良い結果を合成する研究 [2] もある。 使いやすさを考慮した上で、より短い音声サンプルで、より良い発話結果を得ることを目標としています。

ボイスクローニングの重要な部分である特徴抽出について、いくつかの既存の特徴抽出技術について理解を深めてきた。話者の特徴を抽出する方法と話者認識の方法は強い関係がある。話者認識の先行研究として、機械学習技術を用いた特徴抽出技術 i-vectors[3]、x-vectors[4]、d-vectors[5] などが提案された。d-vectors は、ボイスクローニング [1] で使われている手法である。

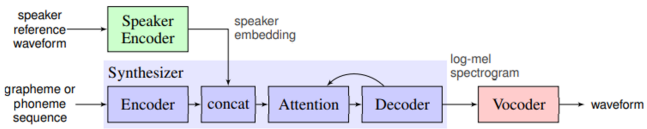

図 1 ボイスクローニング [1] のモデル構造

github にあるリポジトリ [6] と作者の論文 [1] について理解し、トレニンーグを試した。

論文によって、このボイスクローニングの構造は 3 つのモデルに分けていた。システムの構造は図 1のように示した。話者の特徴を抽出する encoder, 抽出した特徴を利用し、話者の声で話すスペクトルグラムを生成する synthesizer と音声を合成する vocoder である。

2.1 Encoder

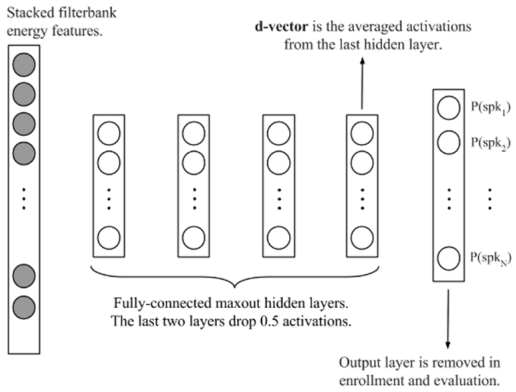

図 2 d-vectors[5] による話者認識の DNN 構造

論文 [5] によって、このモデルは 4 つ hidden layer を持つ maxout DNN である。DNN の構造は図 3のように示した。各 hidden layer で 256 ノードがある。出力の数はトレーニングサンプルの中にある話す人の数と同じ。三番目と四番目の hidden layer が dropout をし、50 %のアクティベーションを無視することでより良い効果が得られる。最後の hidden layer の各ノードのアクティベーション平均値を計算して、これは d-vector である。なぜ最後の hidden layer 出力を使うと言えば。一つ目はモデルのサイズを減ることができる。二つ目は作者らが最後の hidden layer の結果はより良い汎化効果がある。

2.2 Synthesizer

論文 [1] によって、シンセサイザーは Tacotron 2[7] で実装されていると述べている。エンコーダーを通じて、音声から話者の特徴を抽出する。抽出した特徴ベクトルとシンセサイザーのエンコーダーの出力は、各タイムステップでつながっている。元の Tacotron 2 モデルのエンコーダーの入力はテキスト列である。ボイスクローニングのところは前処理でテキスト列を音素列に変換する。これにより、コンバージェンスがより速くなる。

話者の特徴ベクトルは、Tacotron 2 のエンコーダーのが生成した特徴とつながって、新しい特徴になって、次のデコーダーに入力する。デコーダーでメルスペクトルグラムを生成する。

2.3 Vocoder

ボコーダとして WaveNet[8] を用い、Synthesizer が生成したメルスペクトラムを音声波形を変化することができる。

3 提案手法について

ボイスクローニングで合成された音声と、同じ内容を読み上げる話者の音声の違いを比較することで、両者の基本周波数 F0(fundamental frequency) には、大きな違いがあることがわ

* **Realization of natural voice cloning based on speech rate modeling by Qin Zhe. (Hosei Univ) et. al.**

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

かった。 同じ声であっても、f0 の変化により、両者を区別するのに十分な差異をもたらすことができる。

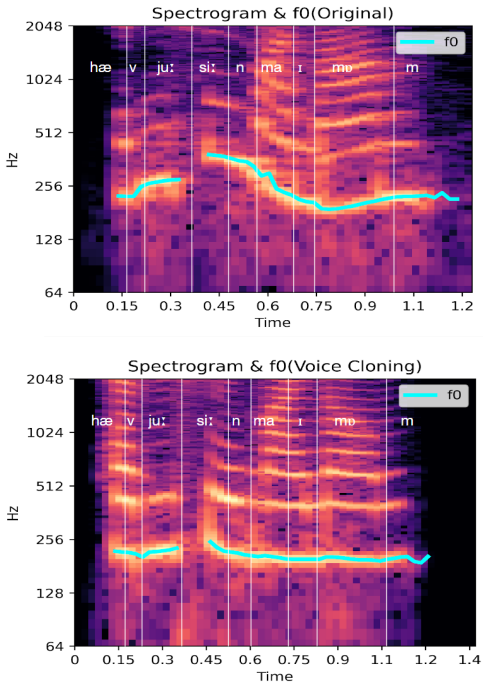

図 3　ボイスクローニング [1] で合成された音声と、同じ内容を読み上げる話者の音声のスペクトルグラムと f0

特に、音声の Speech Rate は、ボイスクローニングでうまく処理できなかった。 1 音素にかかる時間はほぼ同じ、長い単語は継続時間が長い。 話者の習慣による違うが、 多くの人長い単語の各音素は、より速く発音する。図 3はボイスクローニングで合成された音声と、同じ内容を読み上げる話者の音声のスペクトルグラムと f0 である。上は話者の話す声で、下は合成音声である。図 3の 0.3 秒付近から継続時間と音高はうまく処理出来なかったことがわかる。

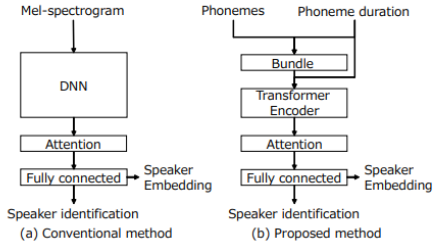

図 4　K. Fujita ら [9] が提出した手法の構造図

K. Fujita ら [9] は、音声リズムに基づく新しい話者エンベディング抽出手法を提案した。その上で、話し方は話者とより近い音声を生成された。

具体的には、音素のワンホットベクトルと継続時間を特徴とし、それらを連結して適切な特徴量ベクトルに変換する。Bundle block は、複数の先行および後続の入力特徴ベクトルを連結し、局所特徴量を抽出する。Transformer encoder block で、系列の特徴を抽出する。従来のモデルと比べて、TDNN を Transformer encoder block に変更することで、全体を見ることで特徴を抽出するようにした。

今考えているのは音声から抽出した特徴ベクトルと話し方を抽出した特徴ベクトルを結合し、話者の声も話し方もより近い音声を生成すること。

4　あとがき

先行研究であるボイスクローニングの中、短い音声から特徴を抽出し、音声を生成するモデルと、長い、特定な音声からよりいい効果をもらえるモデルがある。実用性から考えると、短いデータでより良い結果が得られれば、価値があると思う。私が研究するのは、短い、任意の音声から話し方の特徴を抽出することで、より話者に近い音声を生成することである。

Y. Jia ら [1] は主観的なリスニングテストに基づく crowd-sourced MOS（Mean Opinion Score）で評価した。すべての MOS 評価は、Absolute Category Rating scale を準拠しており、評価点数は 0.5 点刻みで 1〜5 点である。このフレームワークを用いて、合成音声を「自然さ」と「実際の話者による発話との類似性」の 2 つの側面から評価する。

K. Fujita ら [9] は、Speech rate と　RMSE of　Phoneme duration の箱ひげ図とを用いて、生成された音声を客観的に評価する。主観評価は、XAB リスニングテストを利用した。刺激によるバイアスを排除するため、全ての順列の合成音声ペアを 2 つの順序（XAB と XBA）で提示した。

モデルができた後は、これらの手法を比較した上で適切な評価方法を使う予定である。

参考文献

[1] Jia *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *Advances in neural information processing systems*, vol. 31, 2018.

[2] Mitsui *et al.*, “End-to-end text-to-speech based on latent representation of speaking styles using spontaneous dialogue,” *arXiv preprint arXiv:2206.12040*, 2022.

[3] “Speaker verification using i-vectors,” [Accessed June 07, 2022]. [Online]. Available: https:// ww2. mathworks. cn/ help/ audio/ ug/ speaker- verification- using- ivectors. html? s_ tid=srchtitle_ i- vector_ 4

[4] Snyder *et al.*, “Deep neural network embeddings for text-independent speaker verification.” in *Interspeech*, vol. 2017, 2017, pp. 999–1003.

[5] E. Variani *et al.*, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. ICASSP*, 2014.

[6] CorentinJ, “Corentinj/real-time-voice-cloning: Clone a voice in 5 seconds to generate arbitrary speech in real-time,” [Accessed Nov 11, 2022]. [Online]. Available: https:// github. com/ CorentinJ/ Real- Time- Voice- Cloning

[7] Shen *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[8] Oord *et al.*, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[9] K. Fujita *et al.*, “Phoneme duration modeling using speech rhythm-based speaker embeddings for multi-speaker speech synthesis,” in *Proc. Interspeech 2021*, 2021, pp. 3141–3145.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.