5R-09

発話速度による話者埋め込みによるボイスクローニングの改善 *

秦 哲 †, 伊藤 克亘 ‡,

1 序論

デジタル技術の発展に伴い、情報伝達の手段が著しく変革してきた中、音声は人間のコミュニケーションの主要媒体としての役割を維持している。その技術応用と研究は、近年、多くの注目を集めている。ボイスクローニング技術は、最近に現れた新しいものではないが、新しい技術の出現やハードウェア性能の向上により、この研究には更なる可能性がもたらされている。

先行研究では、少な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 [7]。約 5 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る。また、話者の声や話し方をリアルに再現できるモデルの学習に 18k 発話を利用した研究 [8] もある。

しかし、短時間（2 分以下）のサンプルを使用して生成された音声は、硬直した発音方法によって合成音声であると簡単に判別されることがある。話し方のクローニングにこだわった研究では、長時間（1 時間以上）の音声サンプルを使用する研究が一般的であり、実際の使用において多くの不便が生じる。

本研究では、中国語のボイスクローニングを目指し、音声の特徴ベクトルを抽出すると同時に、話し方の特徴ベクトルも抽出することである。そして、この新しい特徴ベクトルを利用して、よりリアルな音声クローニングを目指す。

2 発話速度による話者埋め

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[1]、x-vectors[9]、d-vectors[5] などが提案された。

Fujita[6] らによって発表されたアプローチは、音素とその継続時間を利用してリズムベースの埋め込みベクトルを抽出する新しい手法である。客観的および主観的な評価結果から、提案手法が対象話者の音声リズムに近い音声リズムを正確に合成できることが示された。

2.1 Duration モデル

2.1.1 モデル・アーキテクチャ

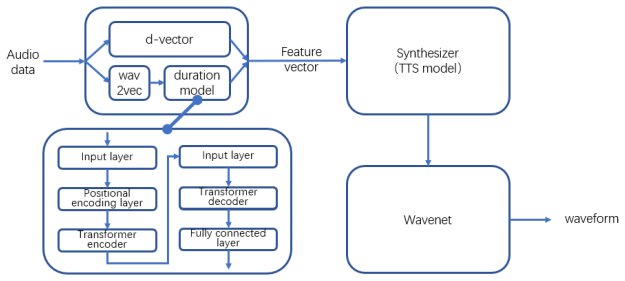

図 1. モデル構造

入力シーケンスは、Transformer[10] モデルで位置エンコーダを用いてエンコードされ、リカレントや畳み込み演算が不要になる。4 つの Transformer エンコーダ層と各層に 8 つのアテンションヘッドが含まれ、線形層で入力が固定次元に変換されてからエンコーダに供給される。デコーダーも 4 つのトランスフォーマーデコーダー層で構成され、ターゲットシーケンスは線形レイヤーで変換される。最終的に、リニアマッピング層が適用され、デコーダーの出力を特定の形式の特徴量予測に整える。

2.1.2 データの準備と学習

Common Voice 13.0 は質の偏りが生じる恐れがあるため、wav2vec を使用し、抽出されたトランスクリプションと元のラベルに大きな差がある場合、そのデータレコードを破棄することにした。

本研究では、音声データから書き起こしと継続時間情報を抽出するために wav2vec2 と CTC を使用し、文字ベースのワンホットエンコードベクトルに変換した。各漢字に対応する継続時間をベクトルに追加し、話者 ID でラベル付けされたデータセットを使用してモデルを訓練した。最終的に、512 次元の特徴ベクトルを出力し、クロスエントロピー損失をメトリックとして使用してモデルの性能を評価した。

2.2 Voice Cloning

音声から継続時間データを抽出するために wav2vec[4] モデルを導入し、Fujita[6] らの研究に基づく Transformer[10] ベースの継続時間モデルを別途学習した。そして、このモデルの出力と GST(Global Style Token) の出力を合成し、発話率特徴ベクトルとして Tacotron[11] モデルに渡す。Synthesizer は、Duration モデルの学習が終了した後、抽出された特徴ベクトルを用いて独自に学習する。エンコーダとボコーダには、Jia Y らがトレーニングした pretrain モデルを利用した.

元の中国語実装 [3] では、話者埋め込みベクトルは、GST の注意メカニズム処理を経た後、テキスト埋め込みベクトルと統合される。この統合により、話者固有の声の特徴を効果的に捉え、具現化したスペクトルが得られる。その後、このスペクトルは音声合成に利用される。

3 実験

継続時間モデルと合成器の訓練に Common Voice 13.0[2] の中国語データセット（1050 時間）と aidatatang_200zh データセット（600 話者、200 時間）を使用した。学習済みモデルの評価には、THCHS30 と AISHELL-1 の 2 つの中国語音声データセットを使用した。

mel-spectrogram を生成するために、MockingBird と再学習した synthesizer を利用する。MockingBird は 3 つのデータセットで 75k ステップ学習された。我々のシンセサイザーは、aidatatang_200zh データセットで 75k ステップ学習した。話者エンコーダには、Jia Y et al. によって提案されたエンコーダを利用し、MockingBird によっ

---

* Incorporating Speaker's Speech Rate Features for Improved Voice Cloning *Qin Zhe (Grad. School of CIS, Hosei Univ.) et al.*

† 法政大学大学院 情報科学研究科
‡ 法政大学情報科学部

て再学習させた。ボコーダも Wavernn と Hifi-gan に基づいて MockingBird で訓練した。

我々は主に主観的なリスニングテストに基づく MOS（Mean Opinion Score）評価に依存している。スコアは 1 から 5 の範囲で、最低から最高までのパフォーマンス品質を表する。合成音声の評価では、音声の質と類似性（主に話者エンコーダに影響される）、音声スタイルの類似性、音声の自然さ（主に合成器に影響される）の 4 つの次元を考慮する。これらの基準は、評価プロセスの重要なパラメータとなる。

3.1 音声品質と声の類似度

表 1. 音声品質と類似度 MOS(95%信頼区間)

	Voice Quality	Voice Similarity
MockingBird	2.53 ± 0.22	2.69 ± 0.24
Proposed model	3.09 ± 0.22	3.49 ± 0.24

音の類似性はエンコーダが抽出した特徴ベクトルの品質に影響され、音質は主にボコーダに影響される。ただし、シンセサイザーは特徴ベクトルの処理とスペクトログラムの生成に重要な役割を果たし、音声の類似性と音質の両方に一定の影響を与えることに留意する必要がある。表 1 に示すように、本シンセサイザは Mockingbird と比較して、類似度・音質ともに優れている。

3.2 声の自然さと話し方の類似性

表 2. 声の自然さと話し方の類似度 MOS(95%信頼区間)

	Voice Naturalness	Speaking Style Similarity
MockingBird	2.49 ± 0.22	2.50 ± 0.22
Proposed model	3.35 ± 0.22	3.24 ± 0.24

表 2 の結果は、我々のモデルが、短いサンプル入力を持つ中国語ボイスクローニングシステ ムである Mocking-Birdmkb を、音声の自然さと発話スタイルの類似性の両方において上回っ ていることを示している。この優位性は、抽出された発話スタイル特徴を我々のモデルに組み込んだことに起因する。さらに、我々の実験により、入力サンプル音声の時間を 8 秒から 20 秒に延長することで、生成された音声の自然さが大幅に向上することも明らかになった。

4 結論

従来モデルは、アクセント句を考慮していなかった。「アクセント句」とは、発話のリズムや強調を決定する言語単位である。このため、生成される音声は自然さや流暢さに欠ける場合があった。提案モデルは、アクセント句を考慮して発音を生成する。これにより、人間の話し言葉のように、上下文に基づき発音を自然に調整することが可能である。

音声合成において適切なポーズは非常に重要である。これは、必要な休息のスペースを提供するだけでなく、言語の理解性と自然さを高めるためにも役立つ。図 2 は音声波形の対数短時間エネルギーを示し、フレーム長は 512 である。図内の閾値以下の部分は停止を意味する。閾値の計算は平均値に分散を加える方法で指定されている。従来のモデルでは、自然でない、または過度な停止を生成することが問題であった。提案モデルは、より自然で少

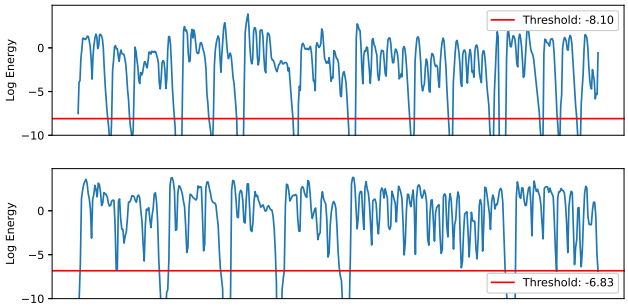

図 2. 従来モデル (上) と提案モデル (下) で生成した音声の対数短時間エネルギー

ないポーズを生成できる。これにより、音声はより自然で聞きやすくなることが期待される。

我々は、Transformer ベースの持続時間モデルを使って、話者の持続時間特徴を抽出し、 ニューラルネットワークベースの多話話者 TTS 合成システムの中国語実装に統合した。 このシステムの合成器に継続時間特徴を組み込むことで、少ない学習データ要件で強化された合成性能を達成した。

提案モデルは、スピーチの間や長いセンテンスを処理する際に、比較的自然に動作する。これは、シーケンスモデリングの最適化により、言語リズムと韻律を捉えることができるためである。Mockingbird モデルと比較すると、提案モデルは長文の生成に優れており、より長いシーケンスを合成する際に行った改良が実証された。このことは、このモデルが音声生成に豊富な文脈情報をよりうまく利用できることを示している。

提案モデルはモッキンバードモデルに対して大きな利点を示すが、解決すべき課題も残っている。特に、スピードアップと、句読点への過度の依存である。 これらの観察結果は、今後の研究を示唆している。

参考文献

[1] Speaker verification using i-vectors. [Accessed June 07, 2022].

[2] Ardila et al. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[3] babysor. Mockingbird. `https://github.com/babysor/MockingBird`, 2023.

[4] Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[5] Ehsan et al. Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ICASSP*, 2014.

[6] Fujita et al. Phoneme duration modeling using speech rhythm-based speaker embeddings for multi-speaker speech synthesis. In *Proc. Interspeech 2021*, pages 3141–3145, 2021.

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

[8] 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.

[9] Snyder et al. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, volume 2017, pages 999–1003, 2017.

[10] Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Wang et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.