# 継続時間を利用した話者埋め込みによるボイスクローニングの向上

Qin, Zhe / 秦, 哲

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
19
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030616

# 継続時間を利用した話者埋め込みによるボイスクローニングの向上

# Incorporating Speaker's Speech Rate Features for Improved Voice Cloning

秦 哲 (Qin Zhe)*

法政大学 情報科学研究科
zhe.qin.7w@stu.hosei.ac.jp

## Abstract

We investigate a neural network-based text-to-speech (TTS) synthesis system that aims to simulate the Mandarin voice of different speakers using short voice samples. Our system introduces new attempts in two key parts.: (1)Unlike prior studies, we not only capture the unique voice characteristics of each speaker but also integrate a speaker feature extraction model to improve our model's effectiveness. Importantly, we utilize a duration model for accurately capturing the speech rate and rhythm of each individual, thus allowing our synthesized speech to demonstrate enhanced realism and credibility.(2) Based on the Tacotron model, we combine additional properties such as "speech rate" and "D-vector" to generate mel spectrograms from a given text, thereby enhancing the ability to clone through mechanisms such as location-sensitive attention. This ensures the quality and authenticity of the synthesized speech. Experiments have shown that this innovative approach uses less data to train and has better sound quality than previous models, and has a certain degree of reproduction in speech speed.

## 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。

「ボイスクローニング」として呼ばれるが、この技術が最初に模倣の表現として登場したと考えている。その時、それは人間の天賦や訓練に大きく依存していた。例えば、ものまねタレントは学習と模倣を通じて、類似の音声効果を実現していた。人の声だけでなく、動物の声や物体が発する音まで模倣されていた。

そして、デジタル信号処理の進化と共に、音声の分析、修正、再合成を試みる研究が活発になった。FFT 等の技術により、音声の詳細な分析が可能となり、より高度な音声処理を達成した。その後、ディープラーニングやニューラルネットワークの進展は、音声の深い特性を捉え、真の音声クローニング技術の発展が促進された。膨大なデータと計算能力の向上は、ディープラーニングのモデルトレーニングを支え、音声クローニング技術の発展をさらに推進している。

研究により、役に立ちたいところはいくつかある：

- 個別化された音声アシスタント：音声クローニング技術を利用することで、ユーザーは彼らのデジタルアシスタントとして好きな音声を選択することができ、自身の音声でさえも選択することができる。
- 映画・テレビ制作：映画やテレビ、アニメの制作において、故人となった俳優や声優の音声が不足している場合、音声クローニング技術は効果的な代替手段となる可能性がある。また、制作経費を削減できると考えている。
- 医療分野での応用：病気や傷害で言語能力を失った患者にとって、音声クローニング技術を用いることで再び「話す」ことが可能となる。
- 文化遺産の保護：一部の言語や方言は徐々に失われつつあり、音声クローニング技術はこれらの独特な音声を保存するのに役立つ可能性がある。

しかしながら、この技術には潜在的な乱用のリスクも存在する。例えば、詐欺行為や偽の証拠を作成するために使用されることも考えられる。そういうところに濫用されると、他のに損害を与えるかもしれない、適切な防犯手段が必要である。

## 2 先行研究

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 [5]。約 5 秒ぐらいのサンプル

* 指導教員：伊藤 克亘 教授

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 [6] もある。 しかし、実用上の制約から、ユーザがこのような長時間のサンプルを入力することは非常に困難である。 本研究では、より短いサンプルから話者の特徴を抽出し、音声を再現することを目的とする。

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[11]、d-vectors[12] などが提案された。d-vectors は、ボイスクローニング [5] で使われている手法である。

Fujita[4] 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。著者らは、話者埋め込みベクトル生成と生成された埋め込みベクトルを用いた音声合成の 2 つの実験を通じて、音声リズムのみに依存した場合でも、話者認識においてわずか 10.3% の EER（Equal Error Rate）を達成することを実証している。

さらに、埋め込みベクトルを視覚的に分析することで、類似したリズムを持つ発話間で類似した特徴が明らかになった。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。

## 3　ボイスクローニング

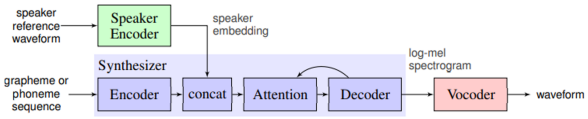


図 1　ボイスクローニング [5] のモデル構造

論文 [5] により、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。

### 3.1　Encoder

論文 [12] 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。

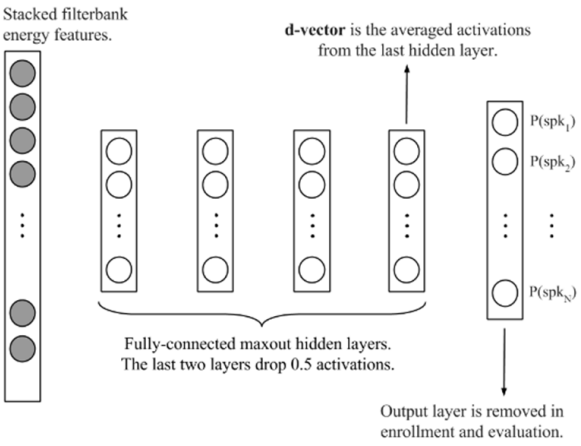


図 2　d-vectors[12] による話者認識の DNN 構造

### 3.2　Synthesizer

論文 [5] により、encoder のところで使ったのは d-vector であり、synthesizer は Tacotron 2[10] 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。

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に変化することができる。

## 4　研究目標

近年の研究では、音声からのアプローチで顕著な成果を上げている。しかしながら、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、非常に類似しているとは言え、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。

したがって、私の研究目的は、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る。先行研究との主な違いは、2 分程度の音声サンプルを使用して、話者の話し方を可能な限りクローニングする。

## 5　提案手法

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、違いがあることがわかった。 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。

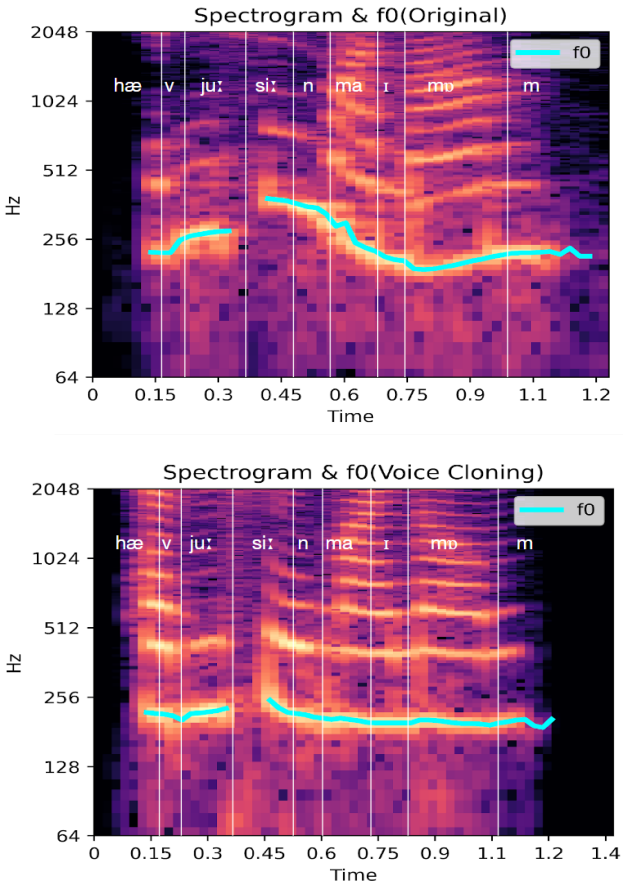


図 3　ボイスクローニング [5] で合成された音声と、同じ内容を読み上げる話者の音声のスペクトルグラムと f0（英語）

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。

以上を踏まえて、まずは Speech rate から話者の特徴を抽出し、シンセサイザーと結合することで、合成音声をより話者本人に近づけることができるようになる。 その後、より詳細な f0 の変化をもとに、話者の特徴を抽出し、類似性を高めていく予定である。

本研究では、wav2vec2 モデルを利用し、アラインメントを取り、漢字ことの継続時間を取る。Peng(2006)[9] と Mok(2009)[7] によると、中国語の等時性はまだ明らかにしていないである。日本語のように明らかなモーラタイミングの特徴がないため、日常で中国語を話す感覚で漢字ことに話速を抽出ことに決めた。

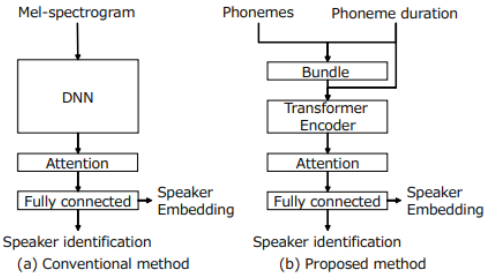


図 4　K. Fujita ら [4] が提出した手法の構造図

K. Fujita ら [4] 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。

K. Fujita らの研究を参考に、話速にこだわりモデルを作りたい。図 5 は提案手法のモデル構造図である。先行研究 [5] のモデルに Duration モデルを入れることで、話速の特徴を抽出する。

## 6　データ処理とモデル構造

### 6.1　データ選択と前処理

継続時間モデルのトレーニングデータセットとして Common Voice13.0 を選択したのは、Common Voice には大規模で多様な音声サンプルセットがあるからである。このデータセットは、複数の言語やアクセントをカバーしているだけでなく、録音環境のバリエーションも含んでいるため、データのバランスが良く、モデルの汎化能力の向上に役立つ。

Common Voice で使用されているデータレビューのメカニズムでは、他の 2 人のユーザーによって「up vote」とマークされたデータをデータセットに含めることができる。しかし、これによってサンプルの質に偏りが生じる可能性がある。例えば、純粋なノイズや早々に終了した音声セグメントが含まれる可能性がある。この問題に対処するため、wav2vecitew2v を使用して各サンプルのトランスクリプ ションと継続時間を抽出する際、抽出されたトランスクリプションと元のラベルの間に長さの大 きな差がある場合、その特定のデータレコードを破棄することにした。

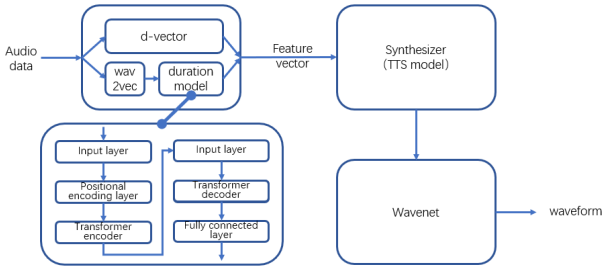


図 5　モデル構造

### 6.2　Duration モデルの学習

#### 6.2.1　モデル・アーキテクチャ

入力シーケンスは、Transformer[13] モデル内でシーケンシャル情報を 捉えるために、最初に位置エンコーダを使ってエンコードされ、リカレント演算や畳み込み 演算を不要にする。エンコーダは 4 つの Transformer エンコーダ層を積み重ねたもので、各層は 8 つのアテンションヘッドを備えている。線形層が入力を固定次元に変換してから、位置エンコードとともにエンコーダに供給される。同様に、デコーダーは 4 つのトランスフォーマーデコーダー層から構成されている。デコーダーに入る前に、ターゲットシーケンスは線形レイヤーを介して変換を受ける。最後に、リニアマッピング層が適用され、デコーダーの出力を、要求された形式に従って、特徴量の予測に関してさらに洗練し、整列させる。

#### 6.2.2　トレーニング

本研究では、音声データから書き起こし情報と継続時間情報を抽出するために、 wav2vec2 学習済みモデルと CTC アルゴリズムを採用した。その後、これらの書き起こしを漢字に基づくワンホットエンコードベクトルに変換した。時間的ダイナミクスを組み込むため、各漢字の対応する継続時間を、それぞれのワンホットエンコーディングベクトルの末尾に付加した。その結果、漢字辞書の長さに 1 を足した長さの拡張ベクトルができた。

最終的に、これらの拡張ベクトルはすべて配列に統合され、モデル学習の入力データとして利用された。データセットのラベルは長さ 1 の話者 ID で構成され、モデルの出力は 512 次元の特徴ベクトルで表現される。モデルの性能を評価するために、我々はクロスエントロピー損失をメトリックとして採用した。このアプローチを通じて、高次元空間内の話者リズム特徴を正確に捉えることを目的としている

### 6.3　Voice Cloning

#### 6.3.1　システムアーキテクチャ

Voice Cloning システムは、Jia Y らの研究に参考した。我々の研究は、この研究の中国語実装 [1] を応用し、元のモデルより良い性能を達成した。図に示すように、我々のモデルの概要を示す。図 6 に示すように、オリジナルのシステムは 3 つの独立した学習コンポーネントから構成されている： (1) 対象話者の数秒間の参照音声を用いて固定次元の埋め込みベクトルを生成する話者エンコーダ、(2) 話者埋め込みベクトルに基づいてテキストの mel スペクトログラムを生成する Tacotron に基づく sequence-to-sequence シンセサイザー、(3)mel スペクトログラムを時間領域の波形サンプルに変換するボコーダ。

音声から継続時間データを抽出するために wav2vec[2] モデルを導入し、Fujita[4] らの研究に基づく Transformer[13] 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[14] 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 らがトレニニーグした pretrain モデルを利用した.

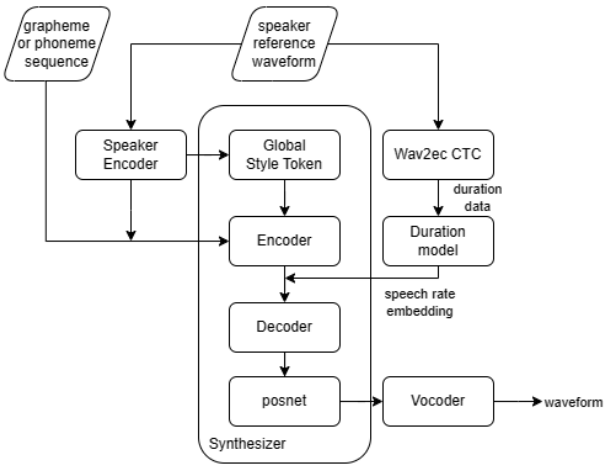


図 6　システム概要：シンセサイザーでは、デュレーションモデルから得られた埋め込みベクトルを組み込み、その後、音声合成の出力品質を向上させるために再トレーニングした。

#### 6.3.2　システムに Duration モデルを統合する

Jia Y et al. は、話者の音声を複製するために、d-vector をモデルに組み込んだ。

元の中国語実装 [1] 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、図 6 に示すように、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。特に、継続時間モデルから抽出された速度特徴ベクトルを組み込むことで、この合成プロセスをさらに強化し、生成される音声により複雑な詳細を吹き込む。

## 7　実験

デュレーションモデルとシンセサイザーを別々に訓練するために、2 つのデータベースを使用した。継続時間モデルの学習には、1050 時間の音声データからなる common voice 13.0[**?**] の中国語データセットを利用し、227 時間を検証した。合成器の学習には、600 話者 200 時間の音声データを含む

aidatatang_200zh データセットを用いた。各文の書き取り精度は 98% 以上である。

モデル学習終了後、2 つの中国語音声データセットをサンプルとして用いた： THCHS30 データセットと AISHELL-1 データセットである。入力テキストとして音声の書き起こしを入力テキストとし、同じ内容を発話した話者の録音と比較した。THCHS30 データセットは主に女性の音声データから構成されており、他の話者の録音による背景雑音が聞こえる可能性があるため、若干の干渉が生じる。一方、AISHELL-1 データセットは、干渉の少ない静かな環境で収集されている。モデルに入力する前に、各スピーチの最初と最後にある長い無音部分をトリミングし、モデルがよりスピーチの内容に集中できるようにした。

mel-spectrogram を生成するために、MockingBird シンセサイザーと再トレーニングしたシンセサイザーを使用した。MockingBird シンセサイザーは 3 つのオープンソースデータセットで 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。

### 7.1 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 | $2.53 \pm 0.22$ | $2.69 \pm 0.24$ |
| Proposed model | $3.09 \pm 0.22$ | $3.49 \pm 0.24$ |

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。

### 7.2 声の自然さと話し方の類似性

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 MockingBirdmkb 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。

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 | $2.49 \pm 0.22$ | $2.50 \pm 0.22$ |
| Proposed model | $3.35 \pm 0.22$ | $3.24 \pm 0.24$ |

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。

## 8 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 7 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。

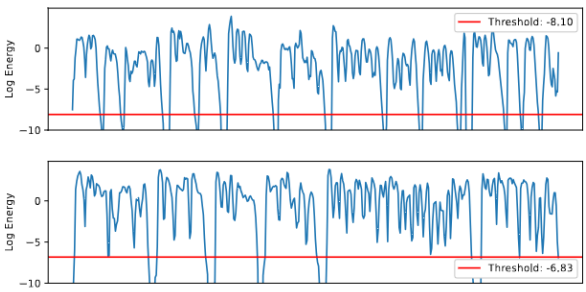


図 7 従来モデル (上) と提案モデル (下) で生成した音声の対数短時間エネルギー

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

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[4] K. 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.

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

[6] 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.

[7] P. Mok. On the syllable-timing of cantonese and beijing mandarin. *Chinese Journal of Phonetics*, 2:148–154, 2009.

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

[9] G. Peng et al. Temporal and tonal aspects of chinese syllables: A corpus-based comparative study of mandarin and cantonese. *Journal of Chinese Linguistics*, 34(1):134, 2006.

[10] 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)*, pages 4779–4783. IEEE, 2018.

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

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

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.