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、ボイス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。

ボイスクローニング技術は、特定の人物の声の特徴を分析し、コピーするための技術である。この技術は、目標とす

* 指導教員：伊藤 克亘 教授

る人物の声のサンプルを収集し、深層学習および機械学習アルゴリズムを使用して声の特徴、例えば音高、リズム、声の色や強度などを分析する。その分析に基づき、入力したテキスト内容を目標の声での音声出力に変換することができる。プロセスとしては、まず多くの目標声音のサンプルを集めることから始まる。これらのサンプルには、話者の異なる声調、感情表現、話速、発音など、声の様々な特性が含まれる。これは、システムが目標声のユニークな特徴を全面的に理解し、学習するために必要である。十分な声のサンプルが収集されると、これらのデータは深層ニューラルネットワークモデルのトレーニングに使用される。モデルは、訓練過程で目標声の特徴を正確に捉え、再現するために、継続的に調整され、最適化される。このプロセスは通常、大量の計算リソースと時間を要し、モデルがデータから声を生成する方法を学ぶためである。これは、声の表面的な特徴を模倣するだけでなく、話者固有の言語習慣や表現方法もとらえる。

トレーニングが完了すると、モデルはテキスト入力を受け取り、それを目標声と似た音声出力に変換することができる。つまり、システムは、ニュース記事、本、またはあらゆる形式のパーソナライズされたメッセージを、目標人物の声で朗読することができる。ボイスクローニング技術はここでその強力な能力を示しており、声の物理的属性だけでなく、ある程度、話者の感情や声調の変化を模倣することで、生成された声をより自然でリアルにする。

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、ボイスクローニング技術の進歩は、新たな挑戦と倫理的な問題をもたらしている。特に個人のプライバシー、身元確認、著作権に関連する問題など、この技術には潜在的な乱用のリスクがある。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。したがって、この技術の発展と応用に伴い、技術の正当で責任ある使用を保証するための法的規制や倫理的指針の策定が非常に重要になる。

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。

Ye Jia たち [5] が紹介した深層学習に基づく多話者テキストから音声への合成システムは、訓練に含まれていない話者の声も生成できる。このシステムは、スピーカーエンコーダー、Tacotron 2 に基づくシンセサイザ、そして WaveNet に基づく自己回帰型ボコーダーネットワークの 3 つの独立した訓練コンポーネントから成り立っている。この研究は、多様な話者データセットを使用することの重要性を強調し、約 5 秒の声のサンプルで話者をクローンする能力を持っている。

一方、Fujita たち [6] が提案した新型テキストから音声へのシステムは、自発的な対話から話し方のスタイルの潜在的な表現を捉えることを目指している。このシステムは、VAE または GMVAE と VITS モデルを組み合わせ、二段階の訓練方法を採用しています。このアプローチにより、特定の話者の特徴だけでなく、自然な対話中のスタイルの変化

も反映した、より自然で表現力のある音声を生成できる。この研究では、約 18000 の発話を含むラジオアナウンサーたちの声のサンプルを使用しており、実際の応用では、これほど大量の声のサンプルを得ることは困難である。

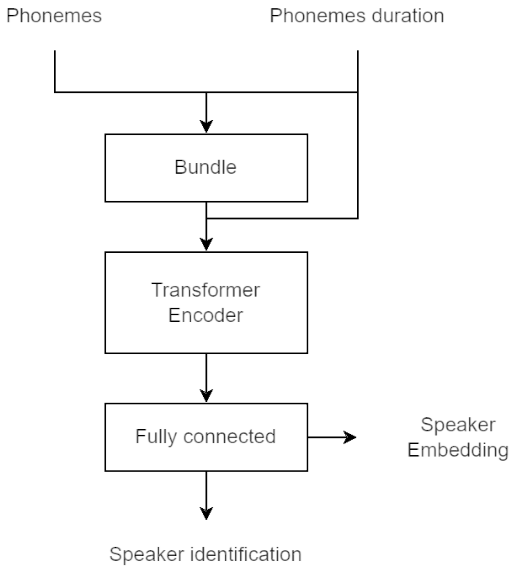

図 1　K. Fujita ら [4] が提出した手法の構造図

3　研究目標

近年、ボイスクローンの分野での進展が著しく、特に限られたオーディオサンプルを使った高品質な音声生成の分野で顕著になっている。しかし、リアルな話し方をシミュレートすることには、現在の技術では限界がある。少量のサンプルに基づいて生成された音声は、音声的にはターゲットの話し手に非常に近いにもかかわらず、口調や話し方が自然ではない。例えば、Jia Y たち [5] が提案したモデルは、わ 5 秒のオーディオサンプルで音声の複製が可能で、印象的な技術の進歩を示している。しかしながら、このモデルでは、5 秒以上のオーディオサンプルを処理する際に顕著な性能向上は見られず、より豊かな音声特徴を抽出する上での潜在的な限界を明らかにした。

一方、具体的な話し方をシミュレートするための現在の研究では、Mitsui らのように、18,000 以上の音声データを使用することもある。実用化から考えると、これほど大量の声のサンプルを得ることは非現実的である。大量のデータへの依存は、資源が制約された環境でのボイスクローニング技術の応用可能性を制限している。

Fujita らの研究は、音声持続時間をモデル化することで、話し手の特定の話し方を反映する特徴ベクトルを抽出することに成功し、短いオーディオサンプルからより豊富な話し方情報を抽出する新たなアイデアを提供した。この研究を参考し、Jia Y らのボイスクローニングン技術と Fujita らの話し方の特徴抽出方法を組み合わせて、限られた音声データから、より自然で、より話者の個性的な特徴を生かした音声クローンの実現を目指す。

そこで、本研究の目的は、中国語のボイスクローニングモデルを向上することである。このモデルは、音声のスペクトル特徴だけでなく、話し方の特徴ベクトルも統合的に抽出することを目指している。このように、比較的短いオーディオサンプル (秒単位のオーディオサンプル) を用いて、音声だけ類似するでなく話し方のスタイルにおいても、目標話者と似た音声を生成することが期待されている。そして、長さの不確定の音声入力に対応できると期待されている。より長い音声を処理する際には、効率的な特徴抽出モデルにより話し方の詳細をキャプチすることができ、合成音声の

自然さとリアリティがさらに向上したい。

4　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。

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)[8] 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。

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 2 は提案システムのコンポネントを示す図である。先行研究 [5] のモデルにデュレーションモデルを入れることで、話速の特徴を抽出する。右にある青い部分は Jia Y たち [5] の研究と中国語実装 MockingBird[2] にあるコンポネントである。左にある紫色の部分は私たちが提案したコンポネントである。

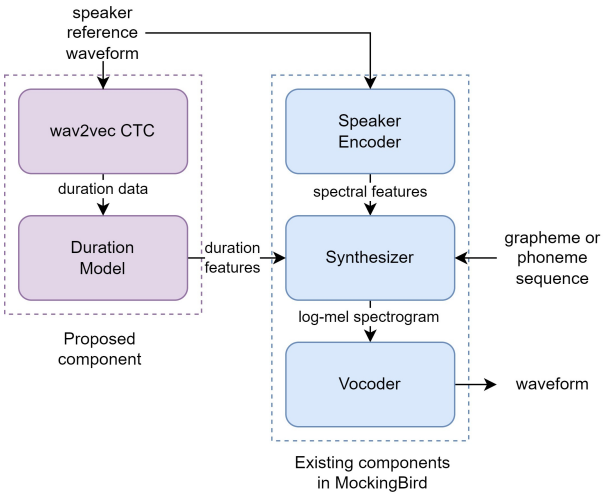

図 2　システムコンポネント

5　データ処理とモデル構造

5.1　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。

5.2 Duration モデルの学習

5.2.1 モデル・アーキテクチャ

入力シーケンスは、Transformer[9] 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。

5.2.2 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。

5.3 Voice Cloning

5.3.1 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 [2] 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 3 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[3] モデルを導入し、Fujita[4] らの研究に基づく Transformer[9] 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[10] 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 らがトレニーングした pretrain モデルを利用した.

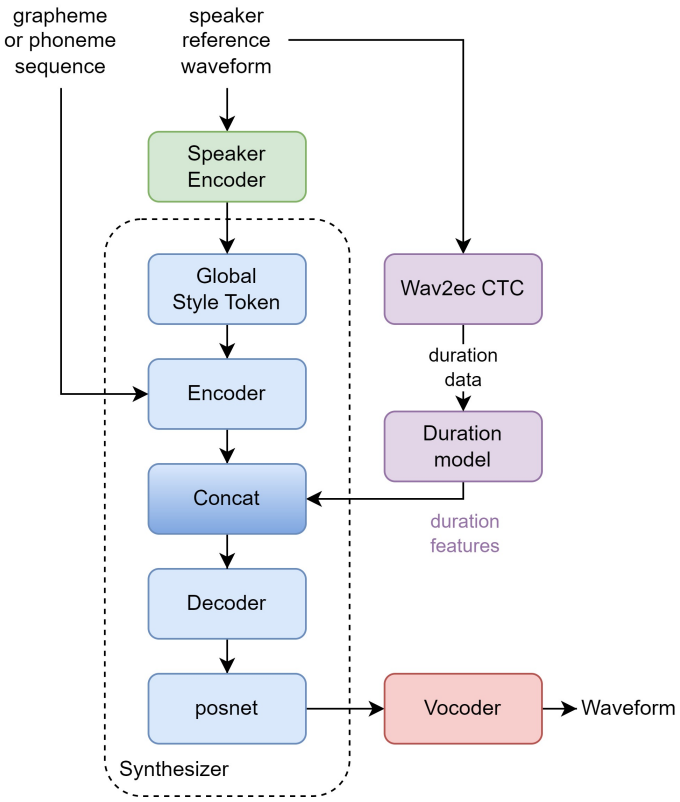

図 3　システム概要：シンセサイザーでは、デュレーションモデルから得られた埋め込みベクトルを組み込み、その後、音声合成の出力品質を向上させるために再トレーニングした。

5.3.2　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 [2] 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 3 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。

6　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[1] 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。

6.1 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。

6.2 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 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。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。

7　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。

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] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

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

[3] 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.

[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] 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.

[9] 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.

[10] 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.