

## YourTTS:

# ゼロショット多人数TTSとゼロショット音声変換をみんなで実現するために

*Edresson Casanova<sup>1</sup>, Julian Weber<sup>2</sup>, Christopher Shulby<sup>3</sup>, Arnaldo Candido Junior<sup>4</sup>, Eren Çelikkaleli<sup>5</sup> and Moacir Antonelli Ponti<sup>1</sup>*

<sup>1</sup>サンパウロ大学計算機科学研究所（ブラジル）

<sup>2</sup>Sopra Banking Software、フランス

<sup>3</sup>Defined.ai、米国

<sup>4</sup>パラナ連邦工科大学（ブラジル）

<sup>5</sup>ドイツ・Coqui

[edresson@usp.br](mailto:edresson@usp.br)

## 概要

YourTTSは、多言語アプローチの力をゼロショット多言語TTSのタスクにもたらしめます。本手法は、VITSモデルをベースに、ゼロショット多言語話者および多言語学習用にいくつかの新しい改良を加えています。その結果、VCTKデータセットにおいて、ゼロショット多言語音声合成でSOTA（state-of-the-art）、ゼロショット音声変換でSOTAと同等の結果を得ることができました。さらに、本アプローチは単一話者のデータセットでターゲット言語において有望な結果を達成し、低リソース言語におけるゼロショット多言語TTSおよびゼロショット音声変換システムの可能性を開くものである。最後に、1分以内の音声でYourTTSモデルを微調整し、音声の類似性と妥当な品質で最先端の結果を達成することが可能である。これは、学習時の音声や録音特性が大きく異なる話者に対して低合成を行うために重要である。

**索引用語：** 異言語ゼロショット多人数TTS、テキスト音声合成、異言語ゼロショット音声変換、話者適応。

[3]は、様々な参照標本から詳細なスタイルを抽出するための注目機構を備えた細粒度エンコーダと、粗粒度エンコーダを提案しました。その結果、複数の参照サンプルを用いることで、未視聴話者に対するより良い音声類似度を実現しました。

## 1. はじめに

近年、TTS（Text-to-Speech）システムは、深層学習のアプローチにより著しく進歩し、音声ベースの仮想アシスタントなどのアプリケーションを成功させることができるようになりました。多くのTTSシステムは、1人の話者の音声から調整されていましたが、現在、数秒間の音声を使用して、新しい話者の音声を作成することに関心が集まっています（学習中に見たことがない）。このアプローチは、[1, 2, 3, 4]のように、ゼロショット多人数TTS（ZS-TTS）と呼ばれています。

深層学習を用いたZS-TTSは、DeepVoice 3方式[6]を拡張した[5]が最初に提案しました。一方、Tacotron 2 [7]は、一般化エンドツーエンド損失（GE2E）を用いて学習した話者エンコーダから抽出した外部話者埋め込みを用いて適応し[8]、ターゲット話者と類似した音声生成を可能にした[1]。同様に、Tacotron 2は、異なる話者埋め込み手法[2]を用いて、LDE埋め込み[9]により、未見話者に対する音声の類似性と自然さを向上させました[10]。また、性別に依存したモデルを用いることで、未視聴の話者に対する類似度が向上することを示した[2]。このような背景から、Attentron

ZSM-SS [11]は、Wav2vec 2.0 [12]に基づくノーマライゼーションアーキテクチャと外部話者エンコーダを備えたTransformerベースのアーキテクチャである。このアーキテクチャでは、スピーカのエンベッディング、ピッチ、エネルギーを条件として正規化を行っている。本論文では、Wav2vec 2.0 をベースにした外部スピーカエンコーダと正規化アーキテクチャを構成する。SC-GlowTTS [4]は ZS-TTS におけるフローベースのモデルの最初の適用例である。本論文では、SC-GlowTTSをZS-TTSに適用することで、従来の研究と比較して、未知の話者に対する音声類似度を改善し、同等の品質を維持した。

このような進歩にもかかわらず、学習時に観測される話者と観測されない話者の間の類似性のギャップは、まだ未解決の研究課題である。また、ZS-TTSモデルは、学習時に相当数の話者を必要とするため、低リソース言語において高品質なモデルを維持することが困難である。さらに、[13]によれば、現在のZS-TTSモデルの品質は、特に学習時の発話特性と異なるターゲット話者に対しては、十分に良いとは言えないという。また、SC-GlowTTS[4]はVCTKデータセット[14]の11話者のみを用いて有望な結果を得たが、訓練話者の数と種類を制限すると、未知の音声に対するモデルの汎化がさらに妨げられる。

また、ZS-TTS と並行して、多言語 TTS も発展しており、多言語のモデルを同時に学習することを目指している [15, 16, 17, 18]。これらのモデルのなかには、コードスイッチング、すなわち、同じ音声を維持しながら文の一部でターゲット言語を変更することを可能にするものがあり、特に興味深い[17]。これは、ある言語から別の言語で合成された音声を使用することを可能にするため、ZS-TTSに有用である。

本論文では、ゼロショット多言語音声学習と多言語学習に焦点を当てたYourTTSを提案する。本論文では、VCTKデータセットにおけるゼロショット多言語音声合成の結果と、SOTAに匹敵する結果を報告する。

私たちの新しいゼロショット多言語TTSアプローチは、以下のような貢献をしています。

- 英語版での最新成果。
- ゼロショット多言語TTSのスコープで多言語アプローチを提案した最初の作品。
- モデル学習時にターゲット言語の1話者のみを用いて、ターゲット言語において有望な品質と類似性を持つゼロショット多言語TTSおよびゼロショット音声変換を行うことができます。

- モデル学習時の声質・録音特性と大きく異なる話者に対して、1分以内の発話でモデルの微調整を行い、かつ良好な類似度と品質を実現する。

各実験の音声サンプルを公開しています。

デモサイトにて<sup>1</sup>再現性を高めるため、ソースコードはCoqui TTS<sup>2</sup>にて公開されています。<sup>3</sup>

## 2. YourTTSモデル

YourTTSはVITS

[19]をベースにしているが、ゼロショット多言語学習用にいくつかの新しい改良が加えられている。まず、以前の研究[4,

19]とは異なり、我々のモデルでは音素の代わりに生のテキストを入力として使用しています。これにより、オープンソースの優れた書記素-音素変換器がない言語でも、より現実的な結果を得ることができる。

これまでの研究（例えば[19]）と同様、変換器ベースのテキストエンコーダ[20, 4]を用いる。ただし、多言語学習のために、4次元の学習可能な言語埋め込みを各入力文字の埋め込みに連結している。また、変換器のブロック数を10に、隠れチャンネル数を196に増やした。デコーダは、VITSモデルと同様に、4つのアフィン結合層[21]を積み重ねたものを用い、各層はそれ自体が4つのWaveNet残差ブロック[22]を積み重ねたものとなっている。

ボコーダとしては、HiFi-GAN [23] バージョン 1 に [19]で導入された識別器の修正を加えたものを使用する。さらに、効率的なエンドツーエンド学習のために、VAE（variational autoencoder）[24]を用いて、TTSモデルとボコーダを接続する。このために、[19]で提案されたPosterior Encoderを使用する。Posterior Encoderは16個の非因果的WaveNet残差ブロックから構成される[25, 20]。この潜在変数はボコーダおよびフローベースデコーダの入力として使用されるため、中間表現（メルスペクトログラムなど）は必要ない。これにより、中間表現を学習することが可能となり、ボコーダとTTSモデルを別々に学習する2段階アプローチシステム[19]よりも優れた結果を得ることができる。さらに、入力テキストから多様なリズムの音声を作成するために、[19]で提案された確率的継続時間予測器を用いている。

ここで、(+)は連結を、赤の接続はこの接続によって勾配が伝搬されないことを、破線の接続はオプションであることを示す。HiFi-GAN 識別器ネットワークは簡略化のため省略した。

このモデルにゼロショット多弁生成能力を与えるために、フローベースデコーダ、後置エンコーダ、およびボコーダのすべてのアフィン結合層を外部話者埋め込みに条件付けする。結合層の残差ブロックと後置エンコーダにグローバルコンディショニング[22]を用いている。また、テキストエンコーダとデコーダの出力は、それぞれ継続時間予測器とボコーダへ渡す前に、外部話者エンベディングと合計する。要素ごとの和の前に、線形射影層を使って次元を合わせる（図1参照）。

また、[26]に触発され、最終損失におけるSpeaker Consistency Loss (SCL)を調査した。この場合、事前に学習した話者エンコーダを用いて、生成された音声とグラントゥルース

cosine類似度である。形式的には、 $\varphi(\cdot)$ を話者の埋め込みを出力する関数、 $\cos\text{sim}$ を余弦類似度関数、 $\alpha$ を最終損失におけるSCLの影響を制御する正の実数、 $n$ をバッチサイズとすると、SCLは次のように定義される。

$$LSCL = \frac{-\alpha}{n} \sum_i \cos\text{sim}(\varphi(g_i), \varphi(h_i)) \quad (1)$$

から話者埋め込みを抽出し、その上でSCLを最大化する

<sup>1</sup> <https://edresson.github.io/YourTTS/> <sup>2</sup>

<https://github.com/coqui-ai/TTS>

<sup>3</sup><https://github.com/Edresson/YourTTS>

ここで、 $g$ と $h$  はそれぞれ、グラントゥルースと生成された話者音声を表す。

この潜在変数と話者埋め込みは、波形を生成するGANベースのボコーダジェネレータの入力として使用される。効率的なエンドツーエンドのボコーダ学習のために、[23, 27, 28, 19]と同様に、 $z$  から一定長の部分列をランダムにサンプリングする。Flow-based

decoderは、潜在変数 $z$ と話者埋め込みを $PZp$ 事前分布に関して調整することを目的とする。この $PZp$ 分布をテキストエンコーダの出力と整合させるために、MAS (Monotonic Alignment Search) [20, 19]を使用する。確率的継続時間予測器は、話者埋め込み、言語埋め込み、およびMASによって得られた継続時間を入力として再認識する。人間のような音声のリズムを生成するために、確率的継続時間予測器の目的は、音素（ここでは擬似音素）の継続時間の対数尤度の変分下界である。

推論中、MASは使用されない。その代わりに、 $PZp$ 分布はテキストエンコーダーによって予測され、継続時間は確率的継続時間予測器の逆変換によってランダムノイズからサンプリングされ、その後、整数に変換される。このようにして、潜在変数 $zp$ が分布 $PZp$ からサンプリングされる。逆フローベース復号器は、潜在変数 $zp$ と話者埋め込みを入力として受け、潜在変数 $zp$ を潜在変数 $z$ に変換し、ボコーダ生成器に入力として渡すことで、合成波形を得ることができる。

### 3. 実験風景

#### 3.1. スピーカーエンコーダ

話者エンコーダには、VoxCeleb 2 [31]データセットでPrototypical Angular [30]とSoftmax損失関数を用いて学習した、一般に公開されているH/ASPモデル [29]を使用しました。このモデルは、Vox-Celeb 1 [32]のテストサブセットにおいて、最先端の結果を達成するために選択されました。また、Multilingual LibriSpeech (MLS) [33]のテストサブセットでは、全言語を使用してモデルを評価した。このモデルは平均EERが1.967であるのに対し、SC-GlowTTS論文[4]で用いられた話者エンコーダはEERが5.244であった。

#### 3.2. オーディオデータセット

我々は3つの言語を調査し、各言語につき1つのデータセットを使ってモデルを学習した。全てのデータセットにおいて、前処理を行い、類似のラウドネスを持つサンプルを作成し、長い無音時間を削除した。すべての音声を16Khzに変換し、Webrtcvadツールキットを用いて音声アクティビティ検出 (VAD) を適用した。<sup>4</sup>を適用し、後続の無音部分をトリミングした。さらに、Pythonパッケージffmpeg-normalizeのRMSベースの正規化を使用して、すべての音声を-27dBに正規化した。<sup>5</sup>

<sup>4</sup> <https://github.com/wiseman/py-webrtcvad>

<sup>5</sup> <https://github.com/slhck/ffmpeg-normalize>

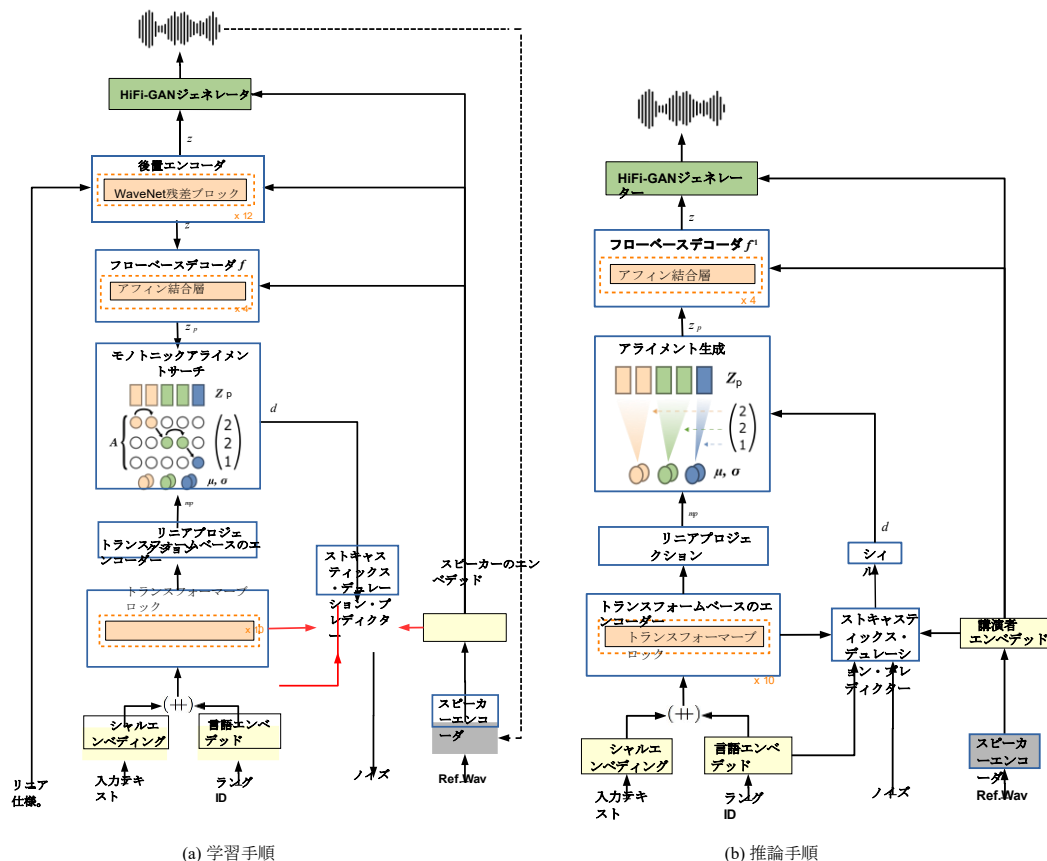


図1: YourTTSの図 (a) 学習手順と (b) 推論手順を示す。

**英語。** VCTK[14]データセットは、44時間の音声と109人の話者を含み、48KHzでサンプリングされている。VCTKデータセットは、訓練、開発（訓練セットと同じ話者を含む）、テストに分けられる。テストセットでは、[1], [4]の提案に従い、各アクセントから女性7名、男性4名の計11名の話者（話者225, 234, 238, 245, 248, 261, 294, 302, 326, 335, 347）を選び、開発及び学習セットには含まれていない話者（話者234, 238, 245, 261, 292, 326, 347）をテストセットとして選択しました。

さらに、いくつかの実験では、モデルの学習における話者の数を増やすために、LibriTTS データセット [34]のサブセット *train-clean-100* と *train-clean-360* を使用しました。

#### ポルトガル語。TTS-

Portugueseコーパス[35]は、ブラジル・ポルトガル語の単一話者によるデータセットで、約10時間の音声を48KHzでサンプリングしたものである。著者らはスタジオを使用していないため、このデータセットには周囲の雑音が含まれている。我々はFullSubNetモデル[36]をノイズ除去に使用し、データを16KHzに再サンプリングした。開発には500サンプルをランダムに選択し、残りのデータセットをトレーニングに使用した。

**フランス語：** LibriVoxを利用したM-AILABSデータセット[37]のfr FRセット<sup>6</sup>。女性2人（104h）、男性3人（71h）の音声を16KHzでサンプリングしている。

英語における本モデルのゼロショット多言語能力を評価するために、テスト用に確保された11人のVCTK話者を用いる。さらに、VCTK以外の領域での性能を検証するために、LibriTTSデータセット[34]のサブセット *test-clean* から10話者（5F/5M）を選択しました。ポルトガル

語については、以下のサンプルを選択した。

<sup>6</sup> <https://librivox.org/>

は、Multilingual LibriSpeech (MLS) [33] データセットの10話者 (5F/5M)から取得した。フランス語については、セクション4で説明した理由により、評価用データセットを使用しませんでした。最後に、話者適応の実験では、より現実的な設定を模倣するために、Common Voiceデータセット[38]から4人の話者を使用しました。

### 3.3. 実験セットアップ

YourTTSを使った学習実験を4回行った。

- **実験1** : VCTKデータセット (モノリンガル) を使用。
- **実験2** : VCTKとTTS-Portugueseの両データセット (バイリンガル) を使用。
- **実験3** : VCTK、TTS-Portuguese、M-AILABS frenchデータセット (3ヶ国語) を使用。
- **実験4** : 実験3で得られたモデルから、LibriTTSパーティションtrain-clean-100とtrain-clean-360の両方から1151人の英語話者を追加して学習を継続する。

学習速度を上げるため、全ての実験において、トランスファー学習を用いた。実験1では、LJSpeech[39]で1Mステップ学習したモデルからスタートし、VCTKデータセットで200Kステップ学習を継続する。ただし、提案した変更により、重みの形状の不適合により、モデルの一部の層がランダムに初期化された。実験2、3では、前回の実験から続けて約140kステップの学習を行い、1言語ずつ学習する。また、各実験において、微調整を行った。

実験 3 では、SCL (Speaker Consistency Loss) を用いて、 $\alpha = 9$  で 50k ステップ学習した。最後に、実験 4 では、実験 3 のモデルを Speaker Consistency Loss で微調整したものを用いて学習を継続する。ZS-TTSの最新の研究[2, 3, 4]ではVCTKデータセットのみを用いているが、このデータセットは話者数が109人と少なく、録音条件のバリエーションも少ないことに注意する必要がある。このため、VCTKのみで学習した場合、録音条件や音声特性が学習時と大きく異なる新しい話者に対しては、一般にZS-TTSモデルはうまく汎化されない[13]。

モデルの学習には NVIDIA TESLA V100 32GB を使用し、バッチサイズは 64 である。TTSモデルの学習とボコーダHiFi-GANの識別には、AdamW optimizer [40]を使用し、ベータ0.8と0.99、重み減衰量 0.01、初期学習率は0.0002で、0.999875のガンマで指数関数的に減衰する[41]。多言語実験では、言語バランスのとれたバッチを保証するために、重み付きランダムサンプリング[41]を用いる。

## 4. 結果および考察

この論文では、[42] と同様に、Mean Opinion Score (MOS) 調査を用いて、合成音声の品質を評価する。合成音声と元の話者の類似性を比較するために、話者エンコーダから抽出された 2 つの音声の話者埋め込み間の話者エンコーダコサインシミュリティ (SECS) [4]を計算する。SECS は -1 から 1 の範囲で、値が大きいほど類似性が高いことを示す[2]。本稿では、先行研究[3, 4]に従い、Resemblyzer[43]パッケージの話者エンコーダを用いてSECSを計算し、先行研究との比較を可能にした。また、[1]、[3]、[4]の研究に従い、類似度MOS (Sim-MOS) を報告する。

この実験では、3 つの言語を用いているが、MOS メトリクスの計算コストが高いため、2 つの言語のみを用いて計算した。また、[4]に従い、MOS メトリクスの計算には2つの言語のみを使用した。また、[4]に従い、このようなメトリクスのみを提示した。

を、トレーニング時に未視聴のスピーカーに適用する。

MOSスコアは、厳密なクラウドソーシングによって取得されました。英語版のMOSとSim-MOSの計算には、それぞれ276人と200人の英語ネイティブの貢献者を使用した。ポルトガル語では、90人のポルトガル語ネイティブの協力者が、両方の指標に使用されました。

VCTKデータセットの第5文 (speakerID 005.txt) は、すべてのテスト話者が発話し、かつ長い文

(20ワード) であるため、評価時に話者埋め込みのための参照音声として使用しました。LibriTTSおよびMLSポルトガル語については、十分な長さの参照音声を確保するため、5秒以上のものだけを考慮して、話者ごとにランダムに1サンプル抽出しています。

英語でのMOS、SECS、Sim-MOSの計算には、LibriTTSデータセットのtest-cleanサブセットからランダムに55文を選び、20語以上の文のみを考慮した。ポルトガル語はこの55文の翻訳を使用した。推論の際、全ての話者をカバーし、十分な数の文を確保するため、話者ごとに5文を合成する。すべて

テスト文の

表1は、VCTKとLibriTTSのデータセットでは英語、MLSのデータセットではポルトガル語のサブセットで、すべての実験のMOSとSim-MOS、95%信頼区間とSECSを示しています。

### 4.1. VCTKデータセット

VCTKデータセットでは、実験1 (モノリンガル) と実験2+SCL (バイリンガル) で最良の類似性結果が得られた。両者とも同じSECSと同じSim-MOSを達成した。Sim-MOSによれば、SCLの使用は改善をもたらさなかったが、全ての実験の信頼区間が重なっており、この分析では結論が出ない。一方、SECSでは、3実験中2実験でSCLの使用により類似度が改善された。また、実験2については、両メトリクスともSCLの類似度への正の効果で一致している。

もう一つの注目すべき結果は、VCTKデータセットの全ての実験において、SECSがグランドトゥールースより高いことである。これは、VCTKデータセット自体の特徴として、例えば、ほとんどの音声に大きな呼吸音が含まれるため、説明することができます。話者エンコーダはこれらの特徴を扱うことができず、その結果、グランドトゥールースのSECSを低下させる可能性があります。全体として、VCTKを用いた我々の最良の実験では、類似性 (SECSとSim-MOS) および品質 (MOS) の結果は、グランドトゥールースと同様であった。MOSに関する我々の結果は、VITSの論文[19]によって報告されたものと一致する。しかし、我々の修正により、このモデルは未知の話者に対して良好な品質と類似性を維持できることが示された。最後に、我々の最良の実験結果は、[3, 4]と比較して、類似度と品質において優れた結果を達成し、ゼロショット多言語TTSのためのVCTKデータセットにおけるSOTAを達成した。

### 4.2. LibriTTSデータセット

実験4では、LibriTTSの類似度が最も高くなった。この結果は、他の実験よりも多くの話者(~1.2k)を用いることで、より広い範囲の音声をカバーしたためと考えられる。と記録条件の多様性を実現します。一方、MOS

は、モノリンガルの場合、最良の結果を達成しました。

のテストサブセットのグランドトゥールースとして、各テストスピーカーの音声をランダムに5つ選択した。SECSとSim-MOSのグランドトゥールースとして、話者ごとに5つの音声をランダムに選択し、合成時に話者埋め込みの抽出に用いた参照音声と比較したところ、SECSとSim-MOSのグランドトゥールースとして、話者ごとに5つの音声をランダムに選択し、合成時に話者埋め込みの抽出に用いた参照音声と比較した。

<sup>7</sup> <https://www.definedcrowd.com/evaluation-of-experience/>

これは主に学習用データセットの品質によるものであることがわかった。実験1では、VCTKデータセットのみを使用し、他の実験で追加したデータセットと比較して、高い品質を実現した。

### 4.3. ポルトガル語MLSデータセット

ポルトガル語のMLSデータセットでは、信頼区間が他の実験と重複しているものの、実験3+SCLのMOS  $4.11 \pm 0.07$ が最高のMOS指標を達成しました。

メンツ。興味深いことに、中程度の品質の単一話者データセットでポル

トガル語を学習したモデルは、ゼロショット多人数話者合成では良好な品質に達することができます。

の論文。実験3はSim-MOSによると最良の実験である ( $3.19 \pm 0.10$ ) が、信頼区間を考慮すると他の実験と重複している。このデータセットでは、Sim-MOS

とSECSは一致しない。SECSの指標に基づくと、実験4+SCLでより高い類似度を持つモデルが得られた。これは、LibriTTSのデータセットが多様であることに起因すると考えられる。また、このデータセットはオーディオブックで構成されているため、MLSデータセットと録音特性や韻律が似ている傾向がある。SECSとSim-MOSのこの差は、Sim-

MOSの信頼区間によって説明できると考えている。最後に、このデータセットで達成されたSim-

MOSは、我々のモデルが1人の男性話者のみを用いて訓練されたことを考慮すると、関連性がある。



表1: すべての実験におけるSECS、MOS、Sim-MOSと95%信頼区間。

Exp.	バイシ ーティ ーケー			LibriTTS			MLS-PT		
	セクス	金属酸化 膜半導体	シムモス	セクス	金属酸化 膜半導体	シムモス	セクス	金属酸化 膜半導体	シムモス
グラウンド・ト ウルース	0.824	4.26±0.04	4.19±0.06	0.931	4.22±0.05	4.22±0.06	0.9018	4.61±0.05	4.41±0.05
アッテントロンZ S	(0.731)	(3.86±0.05)	(3.30 ±0.06)	-	-	-	-	-	-
SC-GLOWTTS	(0.804)	(3.78±0.07)	(3.99±0.07)	-	-	-	-	-	-
EXP.1	<b>0.864</b>	4.21±0.04	4.16±0.05	0.754	<b>4.25±0.05</b>	3.98±0.07	-	-	-
EXP.1 + SCL	0.861	4.20±0.05	4.13±0.06	0.765	4.21±0.04	4.05±0.07	-	-	-
EXP.2	0.857	<b>4.24±0.04</b>	4.15±0.06	0.762	4.22±0.05	4.01±0.07	0.740	3.96±0.08	3.02±0.1
EXP.2 + SCL	<b>0.864</b>	4.19±0.05	<b>4.17±0.06</b>	0.773	4.23±0.05	4.01±0.07	0.745	4.09±0.07	2.98±0.1
EXP.3	0.851	4.21±0.04	4.10±0.06	0.761	4.21±0.04	4.01±0.05	0.761	4.01±0.08	<b>3.19±0.1</b>
EXP.3 + SCL	0.855	4.22±0.05	4.06±0.06	0.778	4.17±0.05	3.98±0.07	0.766	<b>4.11±0.07</b>	3.17±0.1
EXP.4 + SCL	0.843	4.23±0.05	4.10±0.06	<b>0.856</b>	4.18±0.05	<b>4.07±0.07</b>	<b>0.798</b>	3.97±0.08	3.07±0.1

ポルトガル語

男女別のメトリクスを分析すると、男性話者と女性話者のみを考慮した実験4のMOSは、それぞれ4.14±0.11と3.79±0.12である。また、男性話者と女性話者のSim-MOSは、それぞれ3.29±0.14と2.84±0.12である。0.14.したがって、我々のモデルのポルトガル語における性能は

は性別に影響される。これは、我々のモデルがポルトガル語の女性話者で訓練されていなかったために起こったことだと考えています。それにもかかわらず、我々のモデルはポルトガル語の女性の音声を生成することができました。アッテントロンモデルでは、Sim-Simを達成しました。

約100人の話者と英語学習を行った結果、MOSは3.30±0.06となった。信頼度インター

また、ターゲット言語が男性1名の場合でも、Sim-MOSを達成することができました。したがって、我々のアプローチは、低リソース言語におけるゼロショット多言語TTSモデル開発のためのソリューションになり得ると考えている。

フランス語も含めると（つまり実験3）、ポルトガル語の品質と類似度（SECSによる）の両方が向上したように見える。これは、M-

AILABSのフランス語データセットがポルトガル語コーパスよりも高品質であること、また、言語ごとにバッチをバランスさせることにより、モデル学習時にバッチの中の低品質音声が増加するためと考えられます。また、TTS-

Portugueseは単一話者のデータセットであり、実験2で言語ごとにバッチをバランスさせると、バッチの半分が男性1人のみで構成されるため、類似度の増加が説明できる。フランス語が追加された場合、ポルトガル語話者の音声で構成されるバッチは3分の1になります。

#### 4.4. スピーカーの整合性喪失

話者整合性損失（SCL）を用いることで、SECSで測定した類似度が改善された。一方、Sim-MOSでは、実験間の信頼区間は、SCLが類似性を向上させたとは断言することはできない。しかし、SECSは訓練時に見られなかった特性を記録することで、汎化することができると考えている。例えば、実験1では、LibriTTSデータセットの録音特性を学習で見えていないモデルが、このデータセットに対するテストでは、SECSとSim-MOSの両方のメトリクスがSCLのおかげで類似度の向上を示しました。一方、SCLを用いると、生成される音声

の品質が若干低下するようである。これは、SCLを使用することで、モデルが基準オーディオに存在する録音特性を生成するように学習し、より多くの歪みやノイズを生成するためであると考えられる。ただし

## 5. ゼロショット音声変換

SC-

GlowTTS[4]モデルと同様に、エンコーダには話者の身元に関する情報を与えないので、エンコーダが予測する分散は、強制的に話者非依存になる。そこで、YourTTSでは、モデルの後置エンコーダ、デコーダ、HiFi-GANジェネレータを用いて音声変換を行うことができる。また、YourTTSに外部の話者埋め込みを条件とすることで、ゼロショット音声変換の設定において、未視聴話者の音声を模倣することが可能となりました。

また、[44]では、AutoVC[45]とNoiseVC[44]のMOSとSim-MOSを、VCTKの10話者について、学習中に見かけなかったモデルで報告されています。この結果を比較するために、VCTKのテストサブセットから8話者（4M/4F）を選択した。また、[44]では10人の話者を用いているが、男女比の関係から8人のみとした。

さらに、ポルトガル語に対するモデルの汎化性を分析し、モデルが1人の話者のみで学習された言語において我々のモデルが達成した結果を検証するために、MLSポルトガル語データセットのテストサブセットから8人の話者（4M/4F）を使用しました。したがって、どちらの言語でも、学習で使用されなかった話者を使用しています。より深い分析のために、[45]に従って、男性ただし、高品質なリファレンスサンプルを用いたテストでは、を使用することで、高品質な音声を生成することができます。

、女性、男女混合の話者間の伝達を個別に比較しました。この分析では、各話者について、3秒以上のサンプルだけを考慮し、ランダムに参照サンプルを選択し、他の各話者の音声で転送を生成しました。また、英語話者とポルトガル語話者の間の音声の伝達を分析した。MOSとSim-MOSは4章で述べたように計算する。ただし、英語とポルトガル語（pt-en、en-pt）間の音声送受信を行う場合のSim-MOSの計算では、参照サンプルが一方の言語であり、他方の言語で送受信を行うため、両言語の評価者（英語とポルトガル語でそれぞれ58名、40名）を使用した。

表2は、これらの実験におけるMOSとSim-MOSを示したものです。ゼロショット音声変換のサンプルは、デモページにあります。<sup>8</sup>

### 5.1. 言語内結果

英語話者から他の英語話者へのゼロショット音声変換（en-en）において、我々のモデルはMOSが

4.20±0.05、Sim-

MOSは4.07±0.06となった。での比較のためのMOSとSim-MOSの結果を報告した[44]。

は、AutoVC [45]とNoiseVC

[44]モデルである。学習時に見かけなかった10人のVCTK話者に対して、AutoVCモデルは以下を達成した。

そのようなことはありません。

<sup>8</sup><https://edresson.github.io/YourTTS/>

表2: ゼロショット音声変換実験のMOSとSim-MOS(95%信頼区間付き)。

レフ/ター	M-M		M-F		F-F		F-M		A-L	
	金属酸化膜半導体	シムモス	金属酸化膜半導体	シムモス	金属酸化膜半導体	シムモス	金属酸化膜半導体	シムモス	金属酸化膜半導体	シムモス
ja-ja	4.22±0.10	4.15±0.12	4.14±0.09	4.11±0.12	4.16±0.12	3.96±0.15	4.26±0.09	4.05±0.11	4.20±0.05	4.07±0.06
ピーターイーピーター	3.84 ± 0.18	3.80 ± 0.15	3.46 ± 0.10	3.12 ± 0.17	3.66 ± 0.2	3.35 ± 0.19	3.67 ± 0.16	3.54 ± 0.16	3.64 ± 0.09	3.43 ± 0.09
EN-PT	4.17±0.09	3.68 ± 0.10	4.24±0.08	3.54 ± 0.11	4.14±0.09	3.58 ± 0.12	4.12±0.10	3.58 ± 0.11	4.17±0.04	3.59 ± 0.05
ピーターイーエン	3.62 ± 0.16	3.8 ± 0.10	2.95 ± 0.2	3.67 ± 0.11	3.51 ± 0.18	3.63 ± 0.11	3.47 ± 0.18	3.57 ± 0.11	3.40 ± 0.09	3.67 ± 0.05

MOSは3.54 ± 1.08となり、Sim-MOSは1.91±1.34となりました。一方、NoiseVCモデルは、MOSが3.38±1.35、Sim-MOSが3.05±1.25となりました。したがって、本モデルはゼロショット音声変換において、SOTAと同等の結果を得ることができた

をVCTKデータセットで学習させた。このモデルはより多くのデータと話者を用いて学習させたが、セクション4におけるVCTKデータセットの類似度の結果は、VCTKデータセットのみを用いて学習させたモデル（実験1）が、本セクションで検討したモデル（実験4）よりも優れた類似度を示していることを示唆している。したがって、YourTTSは、VCTKデータセットのみを用いて学習・評価した場合、ゼロショット音声変換において、非常に近い結果、あるいは、優れた結果を得ることができると考えている。

ポルトガル語話者から別のポルトガル語話者へのゼロショット音声変換において、我々のモデルはMOS 3.64 ± 0.09、Sim-MOS 3.43 ± 0.09を達成した。また音声転送において、我々のモデルの性能が著しく低いこ

女性話者間の類似度 (3.35 ± 0.19) は、男性話者間の移動度 (3.80 ± 0.15) と比較して高い。これは、ポルトガル語には女性話者が少ないためと考えられる。

を学習させた。また、このモデルでは、ポルトガル語の女性の声を見たことがなくても、ポルトガル語の女性の声を近似的に再現することができる。

## 5.2. クロスリンガル結果

どうやら、英語話者とポルトガル語話者の間の転送は、ポルトガル語話者間の転送と同じようにうまくいくようです。しかし、ポルトガル語話者から英語話者への転送（pt-en）では、MOSスコアの品質が低下しています。これは、特に、ポルトガル語話者から英語話者への音声変換の品質が低いことに起因しています。一般に、上述したように、モデルの学習において女性話者が不足しているため、女性話者への転送は悪い結果をもたらす。この場合、ポルトガル語の男性話者の音声を英語の女性話者の音声に変換する必要があるため、課題はさらに大きくなります。

英語では、変換の際、話者の性別はモデルの性能に大きな影響を与えなかった。しかし、ポルトガル語を含む変換では、モデルの学習に女性の声がないため、汎化には支障があった。

## 6. スピーカーの適合性

このような録音条件の違いは、ゼロショット多人数音声合成の課題である。また、学習時の音声と大きく異なる音声を持つ話者も課題となる[13]。しかし、新しい話者や録音条件への適応の可能性を示すために、我々は20秒から61秒の音声サンプルを選択した。

を、Com-mon

Voice

[38]データセットのポルトガル話者2名と英語話者2名 (1M/1F) に対して実施した。この4人の話者を用いて、実験4のチェックポイントに対して、話者整合性損失を用いて、各話者個別に微調整を行う。

微調整の間、多言語合成が損なわれないように、実験

4

で使用したすべてのデータセットを使用した。ただし、適応された話者からのサンプルがバッチの4分の1になるように、重み付きランダムサンプリング[41]を使用する。この方法で1500ステップの学習を行う。評価には、セクション4で説明したのと同じアプローチを用いる。

表3は、各話者の性別、総時間（秒）、学習時に使用したサンプル数、および、グランドトゥルース（GT）、ゼロショット多人数TTSモード（ZS）、話者サンプルによる微調整（FT）のSECS、MOS、Sim-MOSの指標を示したものである。

一般に、学習時に見られなかった録音特性を持つ話者の1分未満の音声を用いたモデルの微調整は、すべての実験で類似度を有意に改善し、非常に有望な結果を得た。

英語では、ゼロショット多人数TTSモードでの本モデルの結果は既に良好で、微調整後は男性話者、女性話者ともにグランドトゥルースに匹敵するSim-MOSを達成しました。また、微調整後のモデルは、グランドトゥルースよりも大きなSECSを達成しており、これは既に過去の実験でも確認されています。この現象は、モデルが録音特性や参照サンプルの歪みをコピーするように学習し、他の実スピーカースAMPLEに対して優位に立つことで説明できると考えています。

ポルトガル語では、ゼロショットと比較すると、微調整は自然さを少し犠牲にすることで、より良い類似性を得ることができるようです。については

男性スピーカースAMPLEでは、Sim-MOSが $3.35 \pm 0.12$ からその話者については、わずか31秒の発話で微調整を行った結果、 $4.19 \pm 0.07$ となった。女性話者の場合、類似度の向上は

さらに、"ent"については、ゼロショット時の $2.77 \pm 0.15$ から、わずか20秒の発話で $4.43 \pm 0.06$ に向上させることができました。

しかし、表3は、使用する音声の量と音声の自然さ（MOS）の間に直接的な関係があることを示しているようです。話者の音声を約1分間使用した場合、我々のモデルは話者の音声特性をコピーすることができ、ゼロショットモードと比較して自然度を高めることも可能です。一方、44秒以下の音声を使用すると、ゼロショットやグランドトゥルースモデルと比較して、生成される音声の品質/自然さが低下する。したがって、我々のモデルはわずか20秒の発話で話者の発話特性をコピーする良好な結果を示していますが、より高い品質を可能にするためには45秒以上の発話がより適切であると言えます。最後に、モデルを微調整すると、主に学習で使用する話者の少ないポルトガル語やフランス語で、音声変換が大幅に改善されることにも気づきました。

表3：話者適応実験のSECS、MOS、Sim-MOSと95%信頼区間。

	セックス	DUR.(SAM.)	モード	セクス	金属酸化膜半導体	シムモス
エン	M	61S (15)	ジーティー	0.875	4.17±0.09	<b>4.08±0.13</b>
			ゼットエス	0.851	4.11±0.07	4.04±0.09
			エフティー	<b>0.880</b>	4.17±0.07	<b>4.08±0.09</b>
	F	44S (11)	ジーティー	0.894	4.25±0.11	<b>4.17±0.13</b>
			ゼットエス	0.814	4.12±0.08	4.11±0.08
			エフティー	<b>0.896</b>	4.10±0.08	<b>4.17±0.08</b>
ピーティー	M	31S (7)	ジーティー	0.880	4.76±0.12	<b>4.31±0.14</b>
			ゼットエス	0.817	4.03±0.11	3.35±0.12
			エフティー	<b>0.915</b>	3.74±0.12	4.19±0.07
	F	20S (5)	ジーティー	0.873	4.62±0.19	<b>4.65±0.14</b>
			ゼットエス	0.743	3.59±0.13	2.77±0.15
			エフティー	<b>0.930</b>	3.48±0.13	4.43±0.06

## 7. 結論、限界、今後の課題

本研究では、VCTKデータセットにおいて、ゼロショット多言語TTSとゼロショット音声合成でSOTAを達成したYourTTSを発表した。さらに、単一話者のデータセットのみを用いて、我々のモデルがターゲット言語において有望な結果を得ることができることを示す。さらに、学習時の音声と録音条件が大きく異なる話者に対して、1分以内の発話で新しい音声に適応できることを示す。

しかし、我々のモデルにはいくつかの限界がある。すべての言語のTTS実験において、我々のモデルは確率的持続時間予測器において不安定であり、一部の話者と文において、不自然な持続時間を生成している。また、特にポルトガル語では、いくつかの単語で誤った発音が発生することがあります。35, 46, 19]とは異なり、我々は音写を使用していないため、このような問題が発生しやすくなっています。ポルトガル語の音声変換では、女性の音声を学習していないため、話者の性別がモデルの性能に大きく影響します。話者適応では、20秒間の発話で話者の発話特性をコピーすることに成功したが、45秒以上の発話があれば、より高品質に変換できる。

今後は、YourTTSモデルの継続時間予測機能の改善や、より多くの言語での学習を行う予定である。また、低リソース環境における自動音声認識モデルの学習におけるデータ補強への応用も検討する予定です。

## 8. 謝辞

本研究は、CAPES (Coordenação de Aperfeiçoamento de

Pessoal de Nível Superior - Brasil) のファイナンスコード001、およびCNPq (National Council of Technological and Scientific Development) のグラント304266/2020-

5の一部によって資金提供された。さらに、本研究の一部は、人工知能エクセレンスセンター (CEIA) より資金提供を受けています。<sup>10</sup>文部省高等教育局 (SESU/MEC) およびサイバーラボグループの助成によるプロジェクトを通じて行われました。<sup>11</sup>また、産業レベルのMOSテストを可能にしたDefined.aiに感謝したい。<sup>12</sup>のおかげで、産業レベルのMOSテストが簡単に利用できるようになりました。最後に、Coqui TTSレポジのすべての寄稿者に感謝します。

<sup>10</sup><http://centrodeia.org> <sup>11</sup><https://cyberlabs.ai>  
<sup>12</sup><https://www.defined.ai>

トリー<sup>13</sup>この仕事は、すべての人の献身的な努力によってのみ可能だったのです。

## 9. 参考文献

- [1] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp.4480-4490.
- [2] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp.6184-6188.
- [3] S. Choi, S. Han, D. Kim, and S. Ha, "Attention:このような場合、「アテンション・ベースの可変長埋め込みを利用した数発のテキスト音声合成」 *arXiv preprint arXiv:2005.08484*, 2020.
- [4] E. カサノバ、C. シュルビー、E. ゴエルジ、N.M. ムラー、F.S. デオリベイラ、A. Candido Jr., A. da Silva Soares, S. M. Aluisio, and M. A. Ponti, "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," in *Proc. Interspeech 2021*, 2021, pp.3645-3649.
- [5] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019-10 029.
- [6] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.
- [7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *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] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp.4879-4883.
- [9] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.
- [10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp.5329-5333.

---

<sup>13</sup> <https://github.com/coqui-ai/TTS>

- [11] N.このような場合、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」、「震災の影響」。
- [12] A.Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0:また、このような場合、「音声の自己教師付き学習のためのフレームワーク」、*Advances in Neural Information Processing Systems*, vol.33, 2020.
- [13] X.Tan, T. Qin, F. Soong, and T.-Y.Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.
- [14] C.ヴェオ、山岸純一、マクドナルドラ、「Supersed-ctr vctk corpus:cstr voice cloning toolkit 用英語多人数話者コーパス、"*University of Edinburgh.The Centre for Speech Technology Research (CSTR)*, 2016.
- [15] Y.Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, 2019, pp.6935-6939.
- [16] Y.Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y.Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language:多言語音声合成と異言語音声クロニング," *Proc. Interspeech 2019*, pp.2080- 2084, 2019.
- [17] T.Nekvinda and O. Dušek, "One model, many languages:多言語音声合成のためのメタ学習『*Interspeech 2020*』。pp.2972-2976, 2020.
- [18] S.Li, B. Ouyang, L. Li, and Q. Hong, "Light-tts:Lightweight multi-speaker multi-lingual text-to-speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, 2021, pp.8383-8387.
- [19] J.Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *arXiv preprint arXiv:2106.06103*, 2021.
- [20] J.Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts:単調アライメント探索によるテキスト音声合成のための生成フロー" *arXiv preprint arXiv:2005.11129*, 2020.
- [21] L.Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.OpenReview.net, 2017.[オンライン].利用可能 : <https://openreview.net/forum?id=HkpbnH9lx>
- [22] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A.Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A Generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
- [23] J.Kong, J. Kim, and J. Bae, "Hifi-gan:Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.
- [24] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [25] R.Prenger, R. Valle, and B. Catanzaro, "Waveglow:A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, 2019, pp.3617-3621.
- [26] D.Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-Lingual Speaker Adaptation Using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis," in *Proc. Interspeech 2021*, 2021, pp.1614-1618.
- [27] M.Binˆkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N.Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2019.
- [28] Y.Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu.Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.
- [29] H.H.S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline sys- tem for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.
- [30] J.S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S.Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.
- [31] J.S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp.1086-1090.[オンライン].利用可能: <http://dx.doi.org/10.21437/Interspeech.2018-1929>
- [32] A.Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.
- [33] V.Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS:A large-scale multilingual dataset for speech research," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds.ISCA, 2020, pp.2757-2761.[オンライン].利用可能: <https://doi.org/10.21437/Interspeech.2020-2826>
- [34] H.Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts:A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.
- [35] E.Casanova, A. C. Junior, C. Shulby, F. S. de Oliveira, J. P. Teixeira, M. A. Ponti, and S. M. Aluisio, "Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese," 2020.
- [36] X.Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet:A full-band and sub-band fusion model for real-time single channel speech enhancement," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun 2021.[オンライン].利用可能: <http://dx.doi.org/10.1109/ICASSP39728.2021.9414177>
- [37] Munich Artificial Intelligence Laboratories GmbH, "The m-a-ilabs speech dataset - caito," 2017.[オンライン].利用可能: <https://www.caito.de/2019/01/the-m-a-ilabs-speech-dataset/>
- [38] R.このような状況下において、「震災」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」。「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」「原発事故」。また、このようなコーパスを利用することで、より効果的かつ効率的な情報収集が可能となる。
- [39] K.伊藤ら、「lj speech dataset」、2017年。
- [40] I.Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017.
- [41] A.バシユケ、S.グロス、F.マサ、A.レラー、J.ブラッドベリー、G.チャナン  
T.Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch:An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol.32, pp.8026-8037, 2019.
- [42] F.Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "Crowdmos:また、このような場合、「クラウドソーシングの平均的な意見スコア調査のためのアプローチ」、「音響・音声・信号処理 (ICASSP)、2011 IEEE International Conference on.IEEE, 2011, pp.2416-2419.
- [43] C.ジェミン、「修士論文。リアルタイムボイスクロニング」、2019年
- [44] S.Wang and D. Borth, "Noisevc:高品質なゼロショット音声変換に向けて," *arXiv preprint arXiv:2104.06074*, 2021.
- [45] K.Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc:autoencoderの損失のみによるゼロショット音声スタイル転送、" in *International Conference on Machine Learning*.PMLR, 2019, pp.5210-5219.
- [46] E.カサノバ、A・C・ジュニア、F・S・デ・オリベイラ、C・シュルビー。  
J.P. Teixeira, M. A. Ponti, and S. M. Aluisio, "End-to-end speech synthesis applied to brazilian portuguese," *arXiv preprint*

