# 吃音者向けTTSシステムのための健常者音素継続長を反映したVITSの学習手法の提案*

☆長久保 諒[1]，山下 陽生[1]，高島 遼一[1,2]，安井 美鈴[3]，滝口 哲也[1]
([1]神戸大学，[2]JST さきがけ，[3]大阪人間科学大学)

## 1 はじめに

吃音とは，滑らかな発話が困難になる発話障害の一種であり，特徴的な非流暢症状として「言葉の繰り返し (連発)」,「言葉の引き伸ばし (伸発)」,「言葉が出るのに時間がかかる (難発)」の 3 つがある。

このような発話障害者のコミュニケーション支援を行う技術として，テキスト音声合成 (Text-to-Speech: TTS) システムによる発話を代替したコミュニケーションが期待される。しかし，一般的に用いられている TTS ツールは流暢な音声を合成できる一方で，使用者ではない他者の声質で合成されてしまうため，自身の声質でコミュニケーションを取りたいという需要を満たすことができない。加えて，複数人が同一ツールを用いた際に誰が発言したのか分かりづらくなってしまうなどの課題がある。

本研究では前述の課題を解決するため，吃音者本人の声質でかつ流暢な音声を合成可能な TTS システムの作成を目的とする。まず，吃音者本人の声質で合成可能にするためには吃音者本人の音声でモデルの学習を行う必要がある。しかし，吃音者などの発話障害者にとって音声収録は健常者以上に大きな負担となるため，学習に必要とされる大量の音声データを収録することが難しい。そこで本研究ではその補完を目的として，大量の健常者音声を用いて学習した事前学習モデルに対して少量の吃音者音声を用いてファインチューニングを行う手法を用いる。これによって吃音者の声質での合成は実現されるが，同時に吃音者音声の発話特徴，特に難発に伴う発話特徴も反映されてしまう。そのため，合成される音声には本来無いはずの無音声区間が出現する問題（入力文章に無い休止）や，読点に伴う休止の無音声区間が長くなってしまう問題（極端に長い休止）が発生してしまう。具体例として「テロがあるからやめろと、さんざん、いわれた。」というテキストを入力として合成した音声のスペクトログラムを Fig. 1 に示す。上段が健常者音声で学習したモデル，下段が健常者モデルを吃音者音声でファインチューニングをしたモデルでそれぞれ合成した音声である。「テロが」と「あるから」と「やめろと」の間には入力文章に読点は存在せず，事前学習モデルで合成した音声では無音声区間が無い。しかし，吃音者音声でファインチューニングしたモデルで合成した音声では，青枠部に示すように無音声区間，すなわち「入力文章に無い休止」が発生してしまっている。また,「さんざん，いわれた。」の読点に伴う無音声区間について，赤枠部に示すように健常者事前学習モデルで合成したものと比較して長く，すなわち「極端に長い休止」が生まれてしまっている。

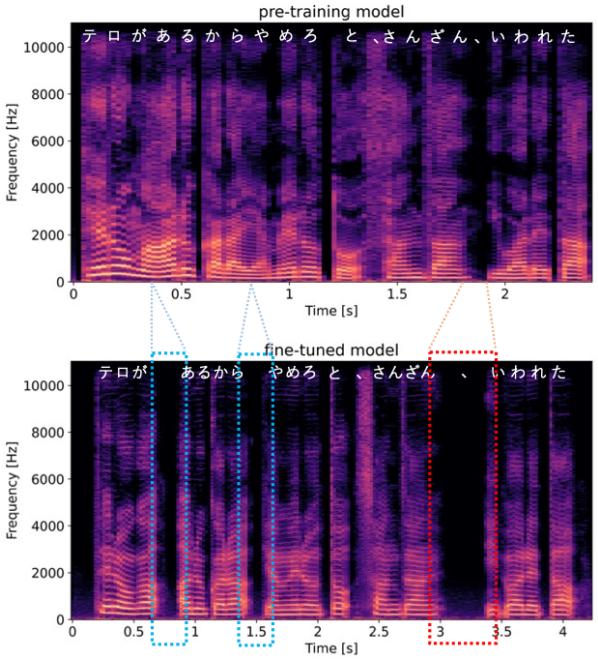


Fig. 1 Spectrograms of (upper) speech synthesized by pre-training model and (lower) speech synthesized by fine-tuned model.

本研究では，先述したような合成音声における問題を解消し，吃音者の声質でかつ流暢な音声を合成する TTS システムを作成するために，テキスト音声合成モデルのファインチューニングの際に，テキストラベルの整形を行う手法と健常者音声の音素継続長を用いて学習する手法の 2 つの手法を検討する。

*Training of VITS model reflecting the duration of a physically unimpaired speaker for a text-to-speech system for a person with stutter. by Ryo Nagakubo[1], Haruki Yamashita[1], Ryoichi Takashima[1, 2], Mirei Yasui[3], Tetsuya Takiguchi[1] ([1]Kobe University, [2]JST PRESTO, [3]Osaka University of Human Sciences.)

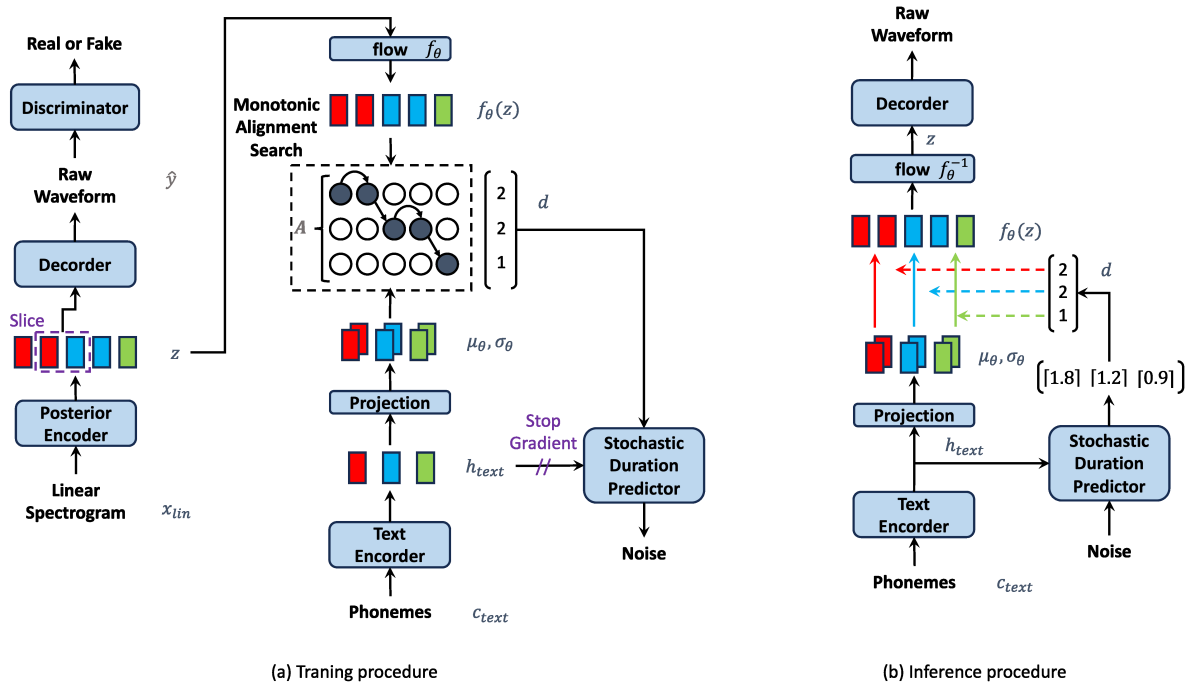


Fig. 2　(a) Traning procedure and (b) inference procedure for VITS.

## 2　VITS

本研究では，TTS モデルとして VITS [1] を用いる。VITS は Glow-TTS [2] にボコーダの HiFi-GAN [3] を組み合わせて改良を加え，End-to-End での学習を行うモデルである。モデル構造を Fig. 2 に示す。

VITS の構成の概要について述べる。Posterior Encorder は入力音声のスペクトログラムを入力として音響特徴量 $z$ を出力し，Decoder は音響特徴量 $z$ を入力として音声波形を出力する。また，Discriminator はデータセットの音声波形（Real）と Decorder の出力音声波形（Fake）の 2 種を入力としそれらの識別を行う。これらは VAE [4] と GAN [5] をベースとした構造で学習を行う。Flow [6] は入出力が可逆な構造で，学習時には音響特徴量 $z$ から話者情報を取り除いた $f_\theta(z)$ を出力するように学習し，推論時には逆変換によって発話情報に話者情報を付与する。Text Encorder は音素列 $c_{text}$ を入力として潜在表現 $h_{text}$ を出力とする構造で，$h_{text}$ から Projection でパラメータ $\mu_\theta, \sigma_\theta$ を生成する。Stochastic duration predictor は音素継続長 $d$ の推定を行うための構造で，学習時には，音声情報 $f_\theta(z)$ と音素情報 $\mu_\theta, \sigma_\theta$ から動的計画法ベースの手法である Monotonic Alignment Search (MAS) によって音素系列と音声フレーム系列の対応関係を求め，そこから計算した音素継続長を用いて学習する。

## 3　提案手法

1 章で述べた吃音者音声合成における 2 つの問題の原因は，音素継続長に問題がある吃音者音声に対して通常のテキストラベルを当てはめてファインチューニングを行っていることにあると考えられる。

例えば，「あらゆる現実を、すべて自分のほうへねじまげたのだ」という文章を読み上げた際に，吃音者音声では「あらゆる (pau) 現実を、すべて (pau) 自分の方へ (pau) ねじまげたのだ」といったような休止 (pau) 混じりの読み上げになってしまうが，この休止を連続する文章の一部として学習してしまうため，連続した文章の合成音声にもその休止が出現してしまう。これが入力文章に無い休止が合成音声に出現してしまう原因と考えられる。

また，発話に休止を挟む際に長くなってしまうことがある。そのため，読点に伴う休止の長さが安定しない。これが極端に長い休止が合成音声に出現してしまう原因と考えられる。こうした問題の改善のために，本研究では VITS のファインチューニング時において，テキストラベルの整形を行う手法と健常者音素継続長を用いて学習する手法の 2 手法を提案する。

### 3.1　テキストラベルの整形を行う手法

この手法では，学習データであるテキストラベルに対して処理を行い吃音者音声に合わせたものにする。音声合成モデルについては従来どおりの VITS を用いる。

まず，吃音者の発話中に読み上げ文章中に無い休止が出現した場合に，テキストラベルの対応する箇所に読点を挿入する。これによって，休止を連続する文章の一部として学習してしまうことを防ぐ。加えて，吃音者の音声データ全 503 発話から長期間の休止が含まれる発話を取り除き，比較的流暢な音声データ 151 発話のみを学習に用いる。これによって，読点に伴う休止の長さが不安定になってしまうことを防ぐ。

### 3.2　健常者音素継続長を用いて学習する手法

この手法では，音声合成モデルに対して変更を加え，健常者の音素継続長を学習するようにする。学習データについては従来手法と同じものを用いる。

VITS において TTS の推論時には，Stochastic Duration Predictor で入力テキストの音素情報 $c_{text}$ とホワイトノイズを用いて音素継続長 $d$ を推定している。本手法では，Duration Predictor の学習方法を変更することで，健常者の音素継続長に近い値を用いて音声合成ができるようにする。

Duration Predictor の学習について，音声特徴量から求めた音素継続長 $d$ と入力テキストの音素情報 $c_{text}$ を用いて，$c_{text}$ に対する $d$ の対数尤度 $\log p_\theta(d|c_{text})$ を最大化することで，音素情報から音素継続長を推定できるように学習する。

ここで，従来どおりの VITS では，Monotonic Alignment Search (MAS) によって動的計画法でアライメントを計算し，そこから求めた音素継続長 $d$ を教師として用いる。つまり，この場合は吃音者音声の音素継続長を学習に用いていることになる。これを健常者の音素継続長を学習するように変更するために，本手法では健常者音声で学習を行った事前学習モデルと同値の重みを保ち続ける Text encorder と Duration Predictor を学習時に追加し，これを用いてテキストの音素情報 $c_{text}$ から推定した健常者の音素継続長 $d$ を教師として学習に用いる。

変更を加えた学習時のモデルのうち音素継続長推定に関する部分のみを抜き出して Fig. 3 に示す。推論時のモデル構造については従来手法と同様である。

## 4　評価実験

### 4.1　実験条件

本実験で用いる吃音者の音声データには吃音者女性 1 名が ATR コーパス [7] に含まれる音素バランス文 503 文を読み上げた収録音声を使用する。サンプリング周波数は 22.05kHz である。テキストラベルの整形を行う手法では，この 503 発話のうち比較的流暢であった 151 発話を抜き出し，それぞれ学習データに 121 発話，評価データに 13 発話を用いた。また，

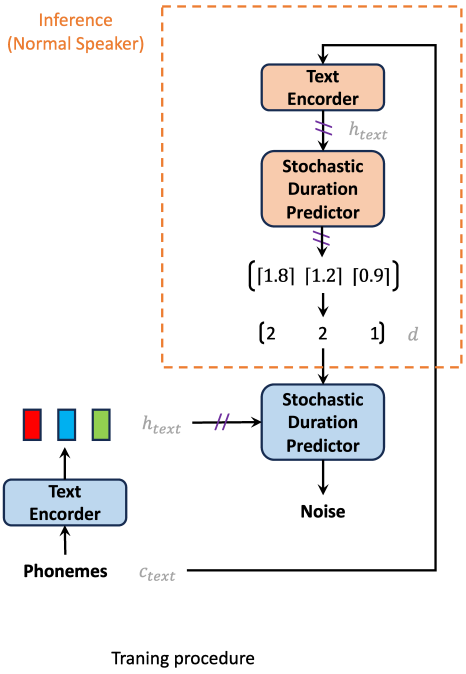


Fig. 3　Proposed training procedure for the stochastic duration predictor of VITS.

健常者音素継続長を用いて学習する手法については 503 発話すべてを用いて，それぞれ学習データに 400 発話，評価データに 53 発話を用いた。

TTS モデルの学習の際には，事前学習モデルとして ESPnet2 [8] の JSUT コーパス [9] を用いた VITS 学習済みモデルを使用し，吃音者音声でファインチューニングを行う。この際に，2 つの提案手法それぞれについて評価する。

### 4.2　実験結果

合成音声の例として,「テロがあるからやめろと、さんざん、いわれた。」という文章について，従来手法による合成音声及び提案手法 2 種による合成音声の合わせて 3 つのスペクトログラムを Fig. 4 に示す。上段には従来手法でファインチューニングを行ったモデル，中段には提案手法のうちテキストラベルの整形を行う手法のモデル，下段には提案手法のうち健常者音素継続長を用いて学習する手法のモデルでそれぞれ合成した音声のスペクトログラムを示している。

本実験では 1 章で述べたような「入力文章に無い休止」と「極端に長い休止」の 2 つの問題が改善できたかどうかに注目する。まず，テキストラベルの整形を行う手法について，従来手法と提案手法のスペクトログラムを比較すると，青枠部に示すように従来手法において存在していた入力文章に無い無音声区間が無くなっていることがわかる。また，赤枠部

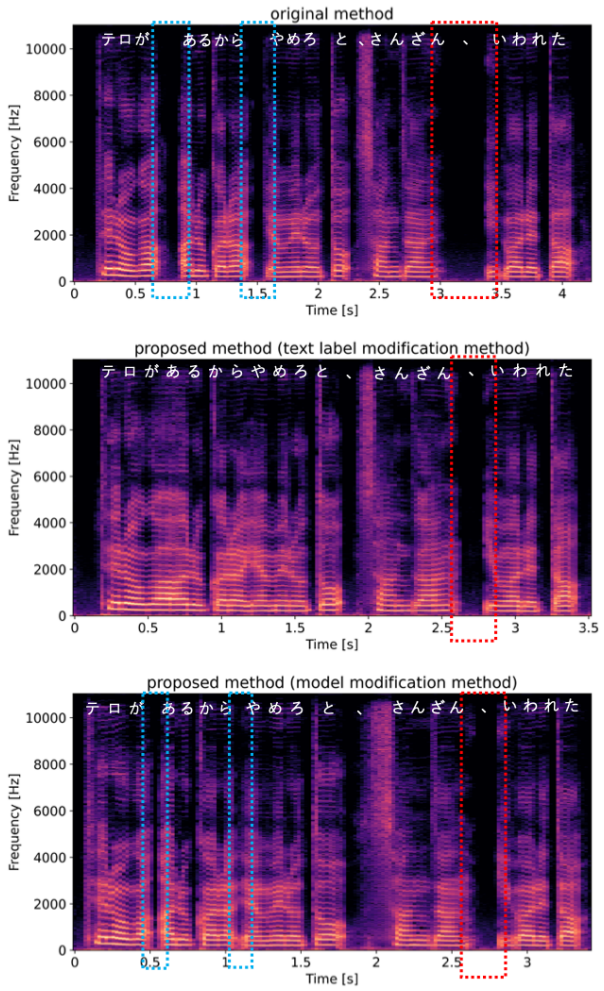


Fig. 4 Spectrograms of speech synthesized using the conventional method (top), the proposed method based on the text label modification (middle) , and the proposed method based on the nomal speaker's duration (bottom).

で示した極端に長い読点に伴う無音声区間について，他の読点の部分と同等の長さになっている。次に，健常者音素継続長を用いて学習する手法について，従来手法と提案手法のスペクトログラムを比較すると，青枠部に示すように従来手法において存在していた無音声区間が，提案手法では短くなっている。これによって，従来手法と比較してある程度滑らかに接続して聞こえるようになった。また，赤枠部で示した読点に伴う無音声区間について，他の読点の部分と同等の長さになっている。

提案手法 2 種を比較すると，「極端に長い休止」の問題はどちらも同等に改善できている一方で，「入力文章に無い休止」の問題はテキスト整形を行う手法では除去できているのに対して，健常者音素継続長を用いて学習する手法では短縮はできたものの除去することはできず解決には至らなかった。

## 5 おわりに

本研究では，吃音者音声合成における問題の改善のために，テキストラベルの整形を行う手法と，モデルに変更を加えて健常者音素継続長を用いて学習する手法の 2 種類の手法をそれぞれ実験した。前者の手法では 2 つの問題は解決できた一方，後者の手法では「極端に長い休止」の補正はできたものの，「入力文章に無い休止」の完全な除去には至らなかった。今後は後者の手法のような整形の手間のかからない手法で前者の手法と同様の成果を得る手法を検討する。

**謝辞** 本研究の一部は，JST さきがけ JPMJPR23I7 および JSPS 科研費 JP21H00906，JP22K12168 の支援を受けたものである。

## 参考文献

[1] J. Kim, *et al.* "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech." International Conference on Machine Learning. PMLR, 2021.

[2] J. Kim, *et al.* "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search." Advances in Neural Information Processing Systems 33 (2020): 8067-8077.

[3] K. Jungil, *et al.* "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." Advances in Neural Information Processing Systems 33 (2020): 17022-17033.

[4] D. P. Kingma, and M. Welling "Auto-encoding Variational Bayes." arXiv preprint arXiv:1312.6114 (2013).

[5] I. Goodfellow, *et al.* "Generative Adversarial Nets." Advances in neural information processing systems 27 (2014).

[6] D. Rezende, and S. Mohamed "Variational Inference with Normalizing Flows." International conference on machine learning. PMLR, 2015.

[7] A. Kurematsu, *et al.* "ATR Japanese speech database as a tool of speech recognition and synthesis." Speech communication 9.4 (1990): 357-363.

[8] T. Hayashi, *et al.* "ESPnet2-TTS: Extending the Edge of TTS Research." arXiv preprint arXiv:2110.07840 (2021).

[9] R. Sonobe, *et al.* "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis." arXiv preprint arXiv:1711.00354 (2017).