吃音者音声合成のための音声認識を用いたテキストラベル修正*

☆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で，「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 つの非流暢症状がある。このような発話障害者のコミュニケーション支援技術として，テキスト音声合成 (Text-to-Speech: TTS) システムが注目されている。しかし，一般的な TTS ツールは流暢な音声を合成可能である一方で，使用者本人の声質ではないため，自身の声でコミュニケーションを取りたいという需要を満たせない。

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，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。これによって吃音者の声質は反映されるが，同時に吃音者音声の発話特徴も反映され流暢性が損なわれてしまう。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事前学習モデル，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。青枠部に示す部分について，入力テキストでは連続文章の一部に対応する部分であるにもかかわらず，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休止が発生してしまっている。さらに，赤枠部に示す部分について，休止が健常者事前学習モデルで合成したものと比較して間延びしてしまっている。

この問題の解決のために，我々は以前，テキスト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 つの手法を提案した [1]。しかし，前者の手法ではこの 2 つの問題を解決できたものの，

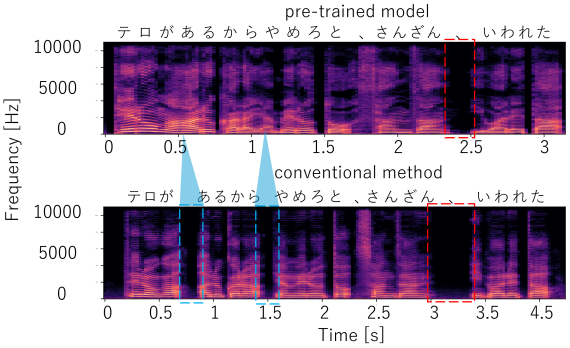

Fig. 1 Spectrograms of (upper) speech synthesized by pre-trained model and (lower) speech synthesized by fine-tuned model.

収録音声に対して人力での整形を行うため手間がかかる。後者の手法は連続文章中に休止が発生してしまう問題については完全な除去には至らなかった。そこで本研究では，人力ではなく音声認識を用いてテキストの修正を行う手法を提案する。

2 VITS

本研究で用いる VITS [2] について概要を示す。モデル構造を Fig. 2 に示す。Posterior Encoder 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 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 Decoder 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 [3] と GAN [4] をベースとした枠組みによって学習される。Flow [5] は入出力が可逆なモジュールで，学習時には 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 Encoder は音素列 c_{text} を入力として潜在表現 h_{text} を出力とするモジュールで，h_{text} から Projection でパラメータ μ_θ,σ_θ を生成する。Stochastic duration predictor は音素継続長 d の推定を行うためのモジュールで，学習時には，音声情報 $f_\theta(z)$ と音素情報 μ_θ,σ_θ から動的計画法を元にした手法である 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。

*Text Label Modification Using Speech Recognition for Speech Synthesis of Stuttered Speech by Ryo Nagakubo[1], Haruki Yamashita[1], Ryoichi Takashinma[1, 2], Misuzu Yasui[3], Tetsuya Takiguchi[1] ([1]Kobe University, [2]JST PRESTO, [3]Osaka University of Human Sciences.)

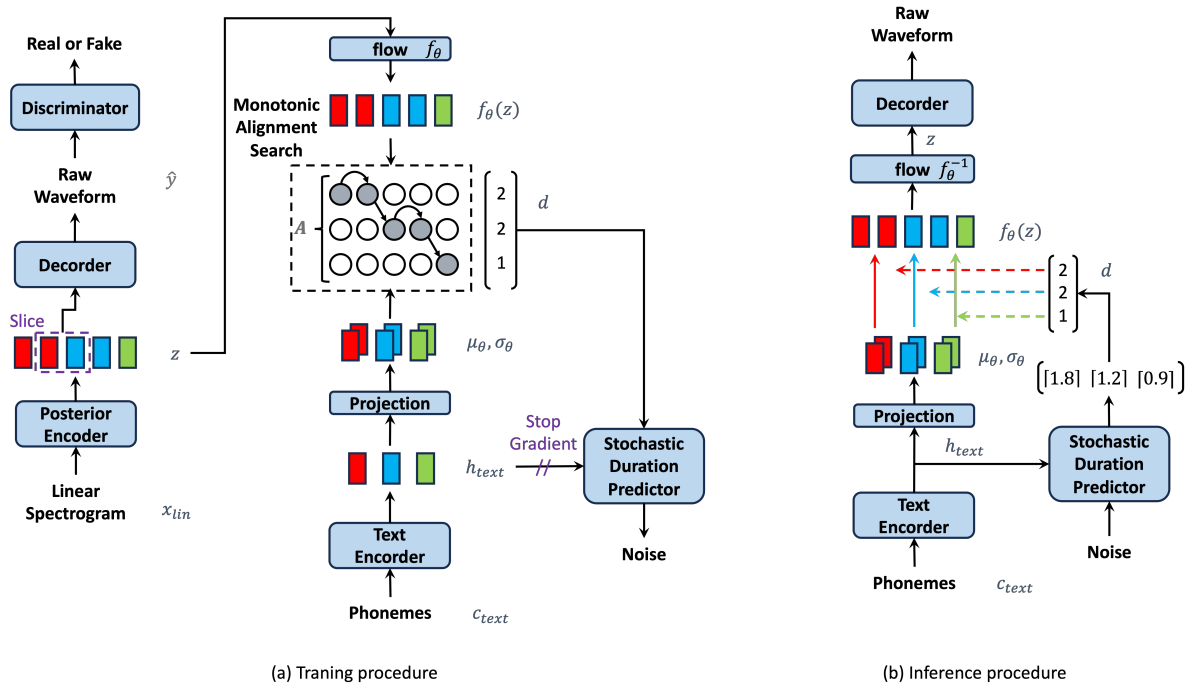

Fig. 2　The architectures of (a) VITS Traning procedure and (b) VITS Inference procedure (TTS model).

3　提案手法

吃音者音声合成における問題の原因は，以下の 2 つであると考える。まず，吃音者音声においては連続発話の途中に休止が頻繁に見られるが，学習テキストにおける該当部は連続文章として与えられるため吃音者音声と学習テキストとの間で不整合が起き，連続発話の一部として休止を学習してしまっている点である。加えて，学習テキストの休止部においても，吃音者音声では吃音症状によって間延びしてしまっており，本来よりも間延びした音素継続長の値を学習してしている点である。以前の研究 [1] で提案した健常者音素継続長を用いた学習手法では，後者が原因と考えられる休止の間延びを解消したものの，前者が原因と考えられる連続発話に区切れて聞こえる部分が存在するという問題が残った。よってこの問題の改善のため， 本研究では音声認識によってテキストラベルの整形を行う手法を提案する。これによって吃音者音声と学習テキストとの間での不整合の解消を目指し，健常者音素継続長を用いた学習手法と併用することでより流暢な吃音者音声合成を目指す。

3.1　健常者音素継続長を用いた学習

我々の先行研究 [1] では，音声合成モデルの学習時のモデル構造に対して変更を加え，健常者音声の値に近い音素継続長での音声合成を行う手法を提案した。VITS では，TTS の推論時には 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。学習については，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} との間の対数尤度 $logp_\theta(d|c_{text})$ を最大化することで，入力テキストの音素情報から音素継続長を推定できるように学習している。つまり，今回の場合は間延びしている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，この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 Encoder と Duration Predictor を学習時にのみ新たに追加し，これを用いて推定した健常者の音素継続長を学習対象として用いることで，健常者に近い値の音素継続長を出力できるようにする。

3.2　音声認識モデルによるテキストラベル修正

本研究では，吃音者音声において頻繁にみられる連続発話中の休止について，学習テキストにおける該当部は連続文章として与えられるという不整合を解消することを目的とする。具体的には，吃音者音声で読み上げテキストにない休止がみられる箇所について，学習テキストに読点 “、” を挿入することで学習テキストを実際の吃音者の発話に近づけ，吃音者音声と学習テキストとの間の不整合を解消する。これによって収録音声ごとの個別調整無しに区切れて聞こえる部分を除去し，より流暢な音声の合成を目指す。この手法の流れを Fig. 3 に示す。まず，学習に用いる吃音者音声を音声認識モデルに入力する。ここで，休止をより検出しやすくし，音声認識の精度を向上するために，学習テキストに対して読点をランダムに挿入したテキストを用いて学習を行った言語モデルを併用す

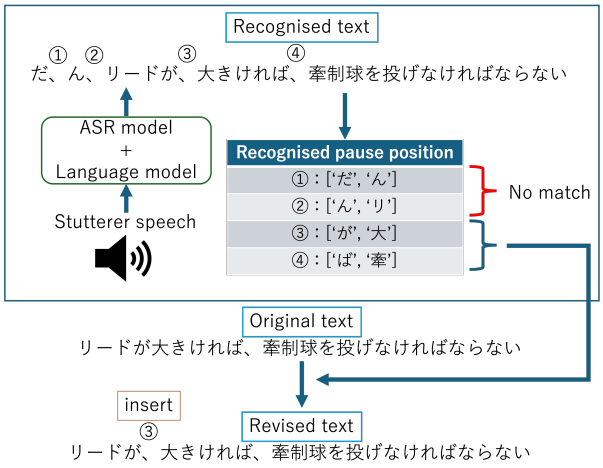

Fig. 3　Overview of the proposed text modification system with automatic speech recognition.

る。このようにして吃音者音声に含まれる休止を読点として認識したテキスト（Recognised text）を得る。得られたテキストは認識内容に誤りが見られるので，これを学習に用いることはできない。よって，音声認識テキストから読点を検出してその前後の文字を記録し，元の学習テキスト (Original text) におけるその前後の文字が一致する箇所に読点を挿入する。この際，一致する箇所の無い読点は認識誤りを含む可能性があるため挿入は行わない。こうして読点を反映した元の学習テキスト（Modified text）を音声合成モデルの学習に用いる。

4　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 [6] に含まれる音素バランス文 503 文を朗読した録音音声を使用する。音声認識モデルは Conformer [7] モデルを日本語話し言葉コーパス（CSJ） [8] を用いて学習した。言語モデルは Transformer [9] ベースのモデルを用い，ATR503 文にランダムで読点を追加したテキストを 100 セット用意し学習させた。音声合成モデルは JSUT コーパス [10] を用いて学習した VITS モデルを事前学習モデルとして使用し，このモデルを吃音者音声を用いて従来手法と提案手法でそれぞれファインチューニングを行い合成した音声および修正テキストについて比較する。これらの実験は ESPnet2 [11] 上で行った。

4.2　読点挿入の精度比較

音声認識モデルによる読点検出精度の検証を行う。吃音者音声を聴取し，元の学習テキストに対して休止が発生した部分に手動で読点の付与を行ったテキストを作成し，これとの間の読点の位置に関する適合率 (Precision) と再現率 (Recall) を元の学習テキスト (Baseline) と提案手法による修正テキスト (Proposed) の 2 つについてそれぞれ求め，提案手法の精度を検証する。結果を Table 1 に示す。

Table 1　Precision, Recall, and F-measure for the Modified text and Original text.

	Precision	Recall
Original text	0.90	0.42
Modified text	0.90	0.67
Modified text (w/o language model)	0.90	0.58

結果では，提案手法によって再現率が向上していることがわかる。このことから，提案手法によって読み上げテキストに含まれない休止を検出し，修正テキストに反映できていると言える。また，適合率が低下していないことから，誤った箇所への読点挿入も多くはみられない事がわかる。

4.3　合成音声の休止出現回数

合成音声の休止出現回数を比較することで，提案手法によって連続文章中の休止がどれほど除去できたかを検証する。元のテキストを用いて学習したモデル (Baseline)，提案手法で修正を行ったテキストで学習したモデル（Proposed method），手動での修正を行ったテキストで学習を行ったモデル（Human modification）で ATR コーパスに含まれる学習に用いていない 50 発話を合成し，そこから検出された休止の合計数を比較する。休止の検出手法としては，音声中最も音量が高い部分から $20dB$ 音量が低くなった部分を休止とみなして検出を行う。ここで，テキストから推定できる妥当な休止の検出数の参考値として，合成に用いたテキストに含まれる“、”と“っ”の合計数も示す（“、” and “っ” in text)。以上の結果を Table 2 に示す。

元のテキストを用いて学習したモデルと比較して提案手法で修正を行ったテキストで学習したモデルは休止の数が減少しており，妥当な休止の検出数に近づいている。このことから，提案手法によって連続文章中の休止がある程度除去できていることがわかる。

4.4　合成音声のスペクトログラム比較

音声の例として，「検札に来た車掌も，見て見ぬふりである」というテキストから合成した音声のスペクトログラムを Fig. 4 に示す。青枠部に示すように元のテキストを用いて学習したモデル（Baseline) では連続文章中に休止が出現しており，健常者音素継続長を用いた手法で学習したモデル（Modified duration) でも休止は完全には取り除けていない。しかし，提

Table 2　Comparison of the number of pause appearances.

	Original text	Modified text	Modified text (w/o language model)	Human modification	“、” and “っ” in text
Number of pause appearances	211	152	174	113	111

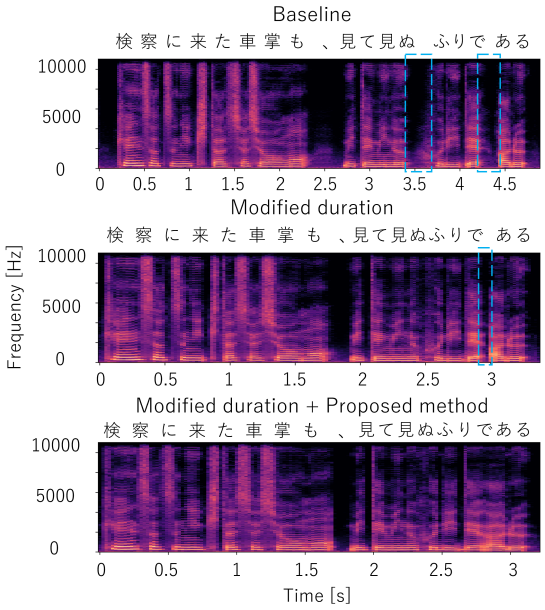

Fig. 4　Spectrograms of speech synthesized by a fine-tuned model using the conventional method (top), by a fine-tuned model using the method based on the nomal speaker’s duration (middle), and by a fine-tuned model using the proposed method (bottom).

案手法と健常者音素継続長を用いた手法を併用したモデル（Modified duration + Proposed method）では，それを取り除き流暢な音声を合成できた。

5　まとめ

本研究では，吃音者音声合成における問題について，以前提案した健常者音素継続長を用いて学習する手法を改善するために，音声認識を用いたテキスト修正手法を提案した．結果について，Fig. 4 に示すように連続文章中の休止発生を抑制することに成功したものの，Table 1 や Table 2 からわかるように，まだ人手による修正の精度には及ばない。今後は休止をより多く検出できる手法を検討する．

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

参考文献

[1] 長久保 諒 他，“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”，音講論春，pp. 919–922，2024.

[2] J. Kim *et al.*, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[3] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[4] I. Goodfellow *et al.*, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[5] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 1530–1538.

[6] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[7] A. Gulati *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[8] K. Maekawa, “Corpus of spontaneous japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[9] A. Vaswani *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[10] R. Sonobe *et al.*, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” *arXiv preprint arXiv:1711.00354*, 2017.

[11] T. Hayashi *et al.*, “ESPnet2-TTS: Extending the edge of TTS research,” *arXiv preprint arXiv:2110.07840*, 2021.