エンド・ツー・エンド音声合成のための逆説的学習を用いた条件付き変分オートエンコーダ

Jaehyeon Kim [1] **Jungil Kong** [1] **Juhee Son** [1 2]

概要

近年、1段学習と並列サンプリングが可能なエンドツーエンド音声合成（TTS）モデルがいくつか提案されているが、そのサンプル品質は2段のTTSシステムに及ばない。本論文では、2段階音声合成方式よりも自然な 音声を生成できる並列エンドツーエンド音声合成方式を提案する。本手法は、正規化フローと敵対的学習過程を備えた変分推論を採用し、生成モデリングの表現力を向上させる。また、入力テキストから多様なリズムを持つ音声を合成するために、確率的な継続時間予測器を提案する。本手法は、潜在変数に対する不確実性モデリングと確率的継続時間予測器により、テキスト入力が異なるピッチとリズムで複数回発話される自然な一対多の関係を実現するものである。LJ Speechという単一話者によるデータセットを用いた人間の主観評価（平均意見スコア、MOS）により、本手法は一般に公開されている最高のTTSシステムを上回り、真実と同程度のMOSを達成することが示された。

1. はじめに

音声合成システムは、与えられたテキストからいくつかのコンポーネントを介して生の音声波形を合成する。ディープニューラルネットワークの急速な発展により、TTSシステムパイプラインは、テキストの正規化や音素化などのテキストの前処理とは別に、2段階のジェネレーティブモデリングに簡素化されている。第一段階は、メルスペクトログラム(Shen et al. , 2018)や言語的特徴(Oord et al. , 2016)を前処理したテキストから生成する。[1]を生成し、第二段階は中間表現を条件とした生波形を生成する（Oord et al., 2016; Kalchbrenner et al., 2018）。2段階のパイプラインのそれぞれにおけるモデルは、独立して開発されている。

ニューラルネットワークベースの自己回帰型TTSシステムは、リアルな音声を合成する能力を示しているが（Shen et al. , 2018; Li et al. , 2019）、その逐次生成プロセスは、最新の並列プロセサを十分に活用することを困難にしている。この制限を克服し、合成速度を向上させるために、いくつかの非自己回帰的な方法が提案されている。テキストからスペクトログラムへの生成ステップでは、テキストとスペクトログラム間のアライメント学習の難易度を下げるために、事前に訓練した自己回帰教師網からアテンションマップを抽出することが試みられている（Ren et al.、 2019; Peng et al.、2020）。最近では、尤度に基づく手法が、ターゲットメルスペクトログラムの尤度を最大化するアライメントを推定または学習することにより、外部アライナーへの依存をさらに排除している（Zeng et al.、2020；Miao et al.、2020；Kim et al.、2020）。一方、生成的副次的ネットワーク（GAN）（Goodfellow et al.、2014）は第二段階モデルで検討されてきた。それぞれが異なるスケールまたは周期のサンプルを区別する複数の識別器を有するGANベースのフィードフォワードネットワークは、高品質の生波形合成を達成する（Kumar et al.）

並列TTSシステムの進歩にもかかわらず、2ステージパイプラインは、後段のモデルが前段のモデルの生成サンプルで訓練される高品質の生産のための逐次訓練または微調整（Shenら、2018；Weissら、2020）を必要とするので、依然として問題がある。さらに、事前定義された仲介特徴への依存は、性能のさらなる改善を得るために学習された隠されたレプリケーションを適用することを排除する。近年、FastSpeech 2s (Ren et al. , 2021) や EATS (Donahue et al. , 2021) など、いくつかの研究により、波形全体ではなく短い音声クリップで学習し、メルスペクトルグラムデコ

[1]Kakao Enterprise, Seongnam-si, Gyeonggi-do, Repub- lic of Korea[2] School of Computing, KAIST, Daejeon, Re-.

ーダを活用してテキスト表現学習を支援するなど、効率的なエンドツーエンド学習方法が提案されている。

韓国での公開　連絡先：　Jaehyeon Kim（キム・　ジェヒョン　<jay.xyz@kakaoenterprise.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[1]TTSシステムにはテキストの前処理があるが、ここでは前処理されたテキストを「テキスト」という単語と同じように使う。

と、専用のスペクトログラムロスを設計して長さを緩和し

は、ターゲットと生成された音声の不一致を示します。しかし

しかし、学習した表現を利用することで性能が向上する可能性はあるものの、合成の質は2段式に劣る。

本研究では、現行の2ステージモデルよりも自然な音声を生成するパラレルエンド・ツー・エンドTTS手法を紹介する。変分オートエンコーダ（VAE）（Kingma & Welling, 2014）を用いて、TTSシステムの2つのモジュールを潜在変数で接続し、効率的なエンドツーエンドの学習を可能にします。高品質な音声波形を合成できるように本手法の表現力を向上させるため、条件付き事前分布に正規化フローを適用し、波形領域で敵対的学習を行う。TTSシステムでは、きめ細かい音声を生成することに加え、テキスト入力がピッチや長さなどのバリエーションを変えながら複数回話される一対多の関係を表現することが重要である。本論文では、この一対多の問題に対処するため、入力テキストから多様なリズムを持つ音声を合成するための確率的デュレーション予測器を提案する。本手法は、潜在変数に対する不確実性モデリングと確率的持続時間予測器により、テキストでは表現できない音声バリエーションを捉えることができる。

本手法は、一般に公開されている最良のTTSシステムであるGlow-TTS (Kim et al. , 2020) with HiFi-GAN (Kong et al. , 2020) よりも自然な音質の音声と高いサンプリング効率を得ることができる。デモページとソースコードの両方を公開している。[2]

2. 方法

本節では、提案手法とそのアーキテクチャを説明する。提案手法は、最初の3つのサブセクションで説明される：条件付きVAE定式化、変分推論に由来するアライメント推定、合成品質を向上させるための敵対的学習。全体的なアーキテクチャについては、このセクションの最後で説明する。図1a、1bはそれぞれ本手法の学習と推論の過程を示す。今後、本手法をVITS（Variational Inference with adversarial learning for end-to-end Text-to-Speech） と呼ぶことにする。

2.1. 変分推論

2.1.1. 概要

VITSは、変分下界を最大化することを目的とした条件付きVAEとして表現することができる。

データ$\log p_\theta(x|c)$の難解な周辺対数尤度の、証拠下限（ELBO）と呼

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\varphi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\varphi(z|x)}{p_\theta(z|c)}\right] \quad (1)$$

ばれるものである。

[2]ソースコード： https://github.com/jaywalnut310/vits
デモ： https://jaywalnut310.github.io/vits-demo/index.html

ここで、$p_\theta(z|c)$ は条件 c を与えられた潜在変数 z の事前分布、$p_\theta(x|z)$ はデータ点 x の尤度関数、$q_\varphi(z|x)$ は近似の後方分布であることを表す。このとき学習損失は負の

ELBOは、再構成損失 $-\log p_\theta(x|z)$ とKLダイバージェンス $\log q_\varphi(z|x) - \log p_\theta(z|c)$, ここで、$z \sim q_\varphi(z|x)$. の合計とみなすことができます。

2.1.2. リコンストラクションロス

再構成損失の対象データ点として、生の波形の代わりにメルスペクトログラムを用い、x_{mel} とする。潜在変数zをデコーダで波形領域$\hat{y}$にアップサンプリングし、$\hat{y}$をメルスペクトログラム領域$\hat{x}_{mel}$に変換する。そして、予測されたメルスペクトログラムとターゲットのメルスペクトログラムの間のL_1 ロスをrecon-struction lossとして使用する。

$$L_{recon} = \|x_{mel} - \hat{x}_{mel}\| \quad (2)$$

これは、データ分布に対してラプラス分布の和をとり、定数項を無視した最尤推定と見なすことができる。このように、メルスペクトラム領域での再構成損失を定義することで、人間の聴覚系の応答に近似したメルスケールを用いて、知覚的な品質を向上させることができる。生波形からのメルスペクトログラム推定は、STFTとメルスケールへの線形射影を用いるだけなので、学習可能なパラメータを必要としないことに注意する。さらに、この推定は学習時にのみ行われ、推論には使用されない。実際には、潜在変数z全体をアップサンプリングするのではなく、部分シーケンスをデコーダの入力として使用し、効率的なエンドツーエンド訓練に使用される窓付きジェネレータ訓練とする（Renら、2021；Donahueら、2021）。

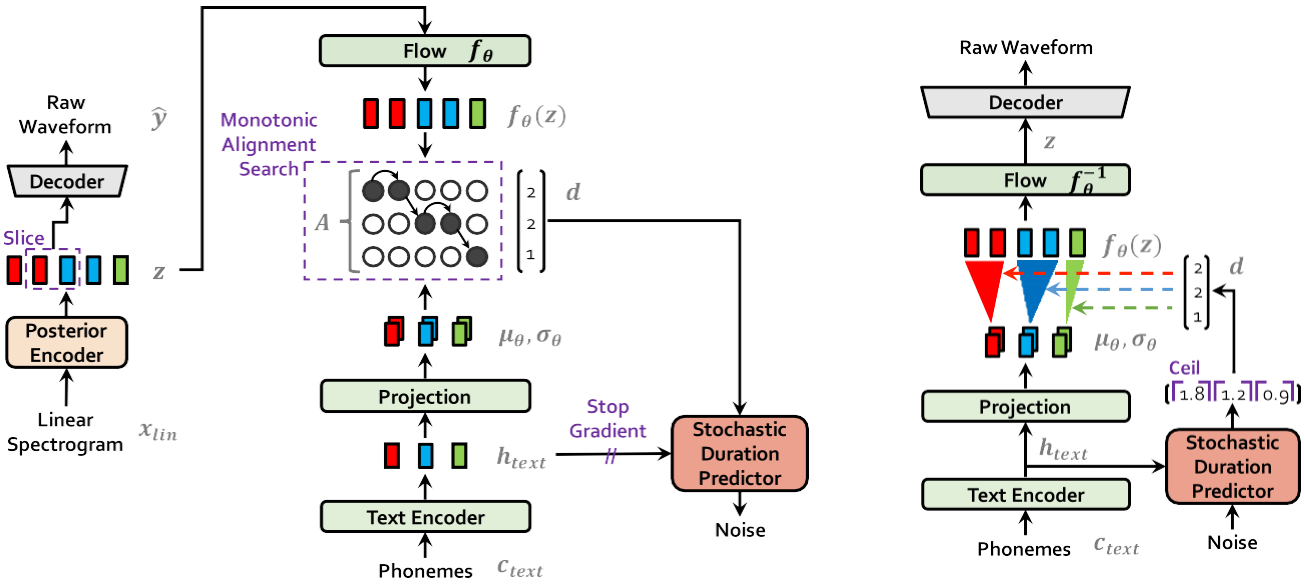
2.1.3. KL-ダイバージェンス

事前エンコーダの入力条件は、テキストから抽出された音素 c_{text} と、音素と潜在変数との間のアライメント A から構成される。アライメントはハードモノトニックアテンション行列は、$|c_{text}| \times |z|$ の次元で、各入力音素がどれくらいの長さに伸びるかを表す。はターゲット音声と時間的に整合している。アライメントのためのグランドトゥルースラベルが存在しないため、各トレーニングイテレーションでアライメントを推定しなければならないが、これについてはセクション2.2.1で説明する。我々の問題設定においては、事後エンコーダのために、より高分解能の情報を提供することを目的とする。そこで、入力としてメルスペクトログラムではなく、対象音声x_{lin} のリニアスケールスペクトログラムを使用する。この入力の変更は変分推論の性質に反しないことに注意。そして、KLダイバージェンスは

$$L_{kl} = \log q_\varphi(z|x_{lin}) - \log p_\theta(z|c_{text}, A) \quad (3)$$

となる。

$$z \sim q_\varphi(z|x_{lin}) = N(z; \mu_\varphi(x_{lin}), \sigma_\varphi(x_{lin}))$$



(a) 学習手順　　(b) 推論手順

図 1.(a)学習手順、(b)推論手順を示すシステム図。提案モデルは、条件付きVAE、事後エンコーダ、デコーダ、条件付き事前処理（緑のブロック：正規化フロー、線形射影層、テキストエンコーダ）、フローベースの確率的持続時間予測器として見ることができる。

因子化正規分布は、我々の事前および事後エンコーダのパラメータ化に使用される。我々は、事前分布の表現力を高めることが、現実的なサンプルを生成するために重要であることを発見した。したがって，我々は，因数分解された正規事前分布の上に，変数の変化の法則に従って，単純な分布からより複雑な分布への可逆変換を可能にする正規化フローf_θ (Rezende & Mohamed, 2015)を適用する．

$$p_\theta(z|c) = \mathcal{N}(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|, \quad (4)$$

$c = [c_{text}, A]$ です。

2.2. アライメント推定

2.2.1. モノトニックアライメント探索

入力テキストとターゲット音声の間のアライメントAを推定するために、正規化フローfでパラメータ化したデータの尤度を最大化するアライメントを探索する手法であるMonotonic Alignment Search (MAS) (Kim et al. , 2020)を採用した。

$$\hat{A} = \arg\max_{A} \log p_\theta(x|c_{text}, A) \quad (5)$$

$$= \arg\max_{A} \log \frac{p_\theta(x, z|c_{text}, A)}{p_\theta(z|c_{text}, A)}$$

$$= \arg\max_{A} \log p_\theta(z|c_{text}, A)$$

$$= \log \mathcal{N}(f_\theta(z); \mu_\theta(c_{text}, A), \sigma_\theta(c_{text}, A)) \quad (6)$$

のテキストを、単語を飛ばすことなく順番に並べる。最適なアライメントを見つけるために、Kimら(2020)はダイナミック・プログラミングを使用している。我々の目的は厳密な対数尤度ではなくELBOであるため、我々の設定にMASを直接適用することは困難である。したがって、我々は、ELBOを最大化するアライメントを見つけるためにMASを再定義し、これは潜在変数zの対数尤度を最大化するアライメントを見つけることに帰着する。

式5と式6が似ているため、オリジナルのMASの実装をそのまま使用することができる。付録Aには、MASの疑似コードが含まれている。

2.2.2. テキストからの期間予測

各入力トークンd_iの持続時間は、推定されたアライメントの各行の全列を合計することで計算することができる。

$$d_i = \sum_j A_{i,j}.$$

この期間は、determinis -のトレーニングに使用することができます。

$$= \arg\max_{\hat{A}} \log N(f(x); \mu(c_{テキスト}, \hat{A}), \sigma(c_{テキスト}, \hat{A})) \quad (5)$$

ここで、候補となるアラインメントは、ヒトがチック持続時間予測ツールは、以前の研究で提案されたものである（Kim et al. , 2020)が、人が毎回異なる発話速度で発話する様子を表現することはできない。人間のような音声を生成するために

の持続時間分布に従うように確率的持続時間予測器を設計する。

音素確率的継続時間予測は、フローベースの生成モデルであり、一般に最大類似度推定によって学習される。しかし、入力音素の継続時間は1)離散整数であり、連続正規化フローを使用するために量子化する必要があるため、最大類似度推定を直接適用することは困難であり、また、2)連続正規化フローを使用するために量子化する必要があるため、最大類似度推定を使用することは困難である。
2)
スカラーであるため、反転可能性による高次元変換を防ぐことができる。これらの問題を解決するために、変分的脱量子化（Ho et al.，2019）と変分的データ増大（Chen et al.，2020）を適用する。具体的には、継続時間列dと同じ時間分解能と次元を持つ2つの確率変数uとvを導入し、
変分的デクワンタイズと変分的データオーグメンテーションを行う。
ションをそれぞれ行う。差分$d-u$が正の実数列となるようにuのサポートを[0, 1]に制限し、vとdをチャンネルごとに連結して
を高次元潜在表現とする。我々は、近似的な事後分布 $q_\varphi(u, v|d, c_{text})$ を介して、2つの変数をサンプリングする。結果として得られる目的は、変分法
音素の継続時間の対数尤度の下限値。

$$\log p_\theta(d|c_{text}) \geqq \mathbb{E}_{q_\varphi(u,v|d,c_{text})}\left[\log \frac{p_\theta(d-u, v|c_{text})}{q_\varphi(u, v|d, c_{text})}\right] \quad (7)$$

そのとき、学習損失L_{dur}
は負の変分下限となる。持続時間予測器の学習が他のモジュールの学習に影響を与えないように、入力の勾配の逆伝播を防ぐstop gradient operator (van den Oord et al. , 2017)を入力条件に適用する。

サンプリングの手順は比較的簡単で、ランダムなノイズから確率的継続時間予測器の逆変換によって音素継続時間をサンプリングし、それを整数に変換するものである。

2.3. アドバーサリー・トレーニング

本研究では、音声合成で成功した2種類の損失、すなわち、敵対的訓練には最小二乗損失関数（Mao et al.、2017）、生成器の訓練には追加的特徴一致損失（Larsen et al.、2016）を用いる。

は、N_l
個の特徴を持つ識別器である。注目すべきは、特徴マッチング損失は、VAEの要素ごとの再構成損失の代替としてsug-
gested識別器の隠れ層で測定される再構成損失とみなすことができる（Larsen et al.、2016）。

2.4. 最終損失

VAEとGANの学習を組み合わせると、条件付きVAEの学習にかかる総損失は次のように表されます。

$$l_{vae} = l_{recon} + l_{kl} + l_{dur} + l_{adv}(g) + l_{fm}(g)$$

(11)である。

2.5. モデル・アーキテクチャ

提案モデルの全体構成は、事後エンコーダ、事前エンコーダ、デコーダ、識別器、確率的持続時間予測器からなる。事後エンコーダと識別器は学習にのみ使用され、推論には使用されない。アーキテクチャの詳細は付録Bを参照されたい。

2.5.1. 後置エンコーダ

事後エンコーダには、WaveGlow (Prenger et al. , 2019) と Glow-TTS (Kim et al. , 2020) で使用されている非因果的なWaveNet残差ブロックを使用します。WaveNetの残差
ブロックはゲート活性化ユニットとスキップ接続を持つ拡張畳み込みの層で構成される。ブロックの上の線形射影層は、正規事後分布の平均と分散を生成する。多話者の場合、残差ブロックでグローバルコンディショニング(Oord et al. , 2016)を使用して話者埋め込みを追加する。

2.5.2. 先行エンコーダ

事前エンコーダは、入力音素を処理するテキストエンコーダc_{text}
と、事前分布の柔軟性を向上させる正規化フローf_θ
で構成されている。テキストエンコーダは、絶対位置エンコーディングの代わりに相対位置表現（Shaw et al.，2018）を用いる変換エンコーダ（Vaswani et al.テキストエンコーダーと、事前分布の構築に用いる平均と分散を生成するテキストエンコーダーの上の線形射影層を通して、c_{text} から隠れ表現h_{text}を得ることができる。正規化フローは、アフィン結合層(Dinh et al. , 2017)からなるスタックである。

$$L_{adv}^{\text{ラドバ}}(D) = \mathbb{E}_{(y,z)}\left[(D(y) - 1)^2 + (D(G(z)))^2\right], \quad (8)$$

$$L_{adv}(G) = \mathbb{E}_z\left[(D(G(z)) - 1)^2\right], \quad (9)$$

$$L_{fm}(G) = \mathbb{E}_{(y,z)}\left[\sum_{l=1}^{T} \frac{1}{N_l} \left\| D^l(y) - D^l(G(z)) \right\|_1\right] \quad (10)$$

ここで、T は識別器の総レイヤー数、D^l は l 番目のレイヤーの特徴マップを出力する。

WaveNet
残留ブロックのスタック。単純化するために、正規化フローは、体積保存変換とする。
をヤコビアンの行列式が1であるようにする。多人数の場合、残差に話者埋め込みを追加する。
をブロックし、グローバルコンディショニングによる正規化フローを実現します。

2.5.3. デコーダ

デコーダは基本的にHiFi-GAN V1 genera- tor (Kong et al. , 2020)である。トランスのスタックで構成されている。

のポーズ付き畳み込みを行い、それぞれの畳み込みの後にマルチ受容野融合モジュール（MRF）を行う。MRFの出力は、異なる受容野サイズを持つ残差ブロックの出力の和である。多話者設定の場合、話者埋め込みを変換する線形層を追加し、入力潜在変数zに追加する。

2.5.4. ディスクリミネーター

HiFi-GANで提案された多周期識別器の識別器アーキテクチャを踏襲している（Kong et al.）多周期識別器は、マルコフ窓ベースのサブ識別器の混合物であり(Kumar et al. , 2019)、それぞれが入力波形の異なる周期パターンに作用する。

2.5.5. ストキャスティックス・デュレーション・プレディクター

確率的持続時間予測器は、条件付き入力h_{text}から音素持続時間の分布を推定する。確率的持続時間予測器の効率的なパラメータ化のために、拡張および深さ分離された畳み込み層で残差ブロックを積み重ねる。また、単調な有理数-2次スプラインを用いた可逆的な非線形変換の形をとるニューラル・スプラインフロー（Durkan et al.ニューラル・スプラインフローは、一般的に用いられるアフィン結合層と比較して、同程度のパラメータ数で変換表現力を向上させることができる。多人数話者設定の場合、話者埋め込みを変換する線形層を追加し、入力h_{text} .

3. 実験風景

3.1. データセット

我々は、2つの異なるデータセットで実験を行った。他の公開モデルとの比較のためにLJ Speechデータセット（Ito, 2017）を、我々のモデルが多様な音声特性を学習し表現できるかどうかを検証するためにVCTKデータセット（Veaux et al.LJ Speechデータセットは、総長約24時間の1人の話者の短い音声クリップ13,100個から構成されています。音声フォーマットは16ビットPCM、サンプルレートは22kHzで、何も加工せずにそのまま使用した。このデータセットをトレーニングセット（12,500サンプル），バリデーションセット（100サンプル），テストセット（500サンプル）に分割して実行した．VCTKデータセットは，様々なアクセントを持つ109人の英語母語話者による約44,000の短い音声クリップから構成される．音声クリップの総時間は約44時間である．音声フォーマットは16ビットPCMで、サンプル　レートは44 kHzである。我々はサンプルレートを22 kHzに下げた。データセットをトレーニングセット（43,470サンプル）、バリデーションセット（100サンプル）、テストセット（500サンプル）にランダムに分割した。

3.2. 前処理

後置エンコーダの入力として、生波形から短時間フーリエ変換（STFT）により得られる線形スペクトログラムを使用する。FFTサイズ、ウィンドウサイズ、変換のホップサイズはそれぞれ1024、1024、256に設定されている。再構成損失には80バンドのメルスケールスペクトログラムを使用し、これは線形スペクトログラムにメルフィルタバンクを適用して得られるものである。

先行エンコーダの入力として国際音声記号（IPA）列を用いる。オープンソースのソフトウェア（Bernard, 2021）を用いてテキスト列をIPA音素列に変換し、変換後の列にはGlow-TTSの実装に従って空白トークンを挟み込んでいる。

3.3. トレーニング

ネットワークはAdamW opti-mizer (Loshchilov & Hutter, 2019)を用いて、$\beta_1 = 0.8, \beta_2 = 0.99$で学習する。重み減衰 $\lambda = 0.01$ である。学習率の減衰は、エポック毎に $0.999^{1/8}$ の係数でスケジュールされ、初期学習率は 2×10^{-4} である。先行研究(Ren et al. , 2021; Donahue et al. , 2021)に従い、窓付き(windowed)を採用する。ジェネレータ学習とは、学習時間や学習時のメモリ使用量を減らすために、生波形の一部分のみを生成する方法である。潜在表現全体を与えるのではなく、窓サイズ32で潜在表現のセグメントをランダムに抽出してデコーダに与え、さらにグランドトゥルース生波形から対応する音声セグメントを抽出して学習対象としている。4台のNVIDIA V100GPUを用い、混合精度学習を行う。バッチサイズはGPUあたり64に設定され、モデルは800kステップまで学習される。

3.4. 比較のための実験セットアップ

我々は、我々のモデルを一般に公開されている最良のモデルと比較した。第一段階モデルとして自己回帰モデルであるTacotron 2とフローベースの非自己回帰モデルであるGlow-TTSを、第二段階モデルとしてHiFi-GANを用いた。それぞれ公開されている実装を用い、あらかじめ学習させた重みを用いている。[3]2段階のTTSシステムは、理論的には逐次学習によってより高い合成品質を達成できるため、第1段階モデルからの予測出力を用いて、100kステップまで微調整したHiFi-GANを組み込んだ。経験的に、Tacotron 2で生成されたメルスペクトログラムでHiFi-GANを教師強制モードで微調整した方が、Glow-TTSで生成されたメルスペクトログラムで微調整するよりもTacotron 2、Glow-TTSともに良い品質となることが分かったため、より良い微調整を行ったHiFi-GANを付加することにした。

[3]実装は以下の通りです。
タコトロン2 : https://github.com/NVIDIA/tacotron2
Glow-TTS : https://github.com/jaywalnut310/glow-tts
HiFi-GAN : https://github.com/jik876/hifi-gan

Tacotron 2とGlow-TTSの両者にGAN。

各モデルはサンプリング時にある程度のランダム性を持つため、実験を通して各モデルのランダム性を制御するハイパーパラメータを固定した。タクトロン2のプリネットにおける脱落確率は0.5とした。Glow-TTSでは、事前分布の標準偏差を0.333に設定した。VITSでは、確率的持続時間予測器の入力ノイズの標準偏差を0.8とし、
、事前分布の標準偏差にスケールファクタ0.667を乗じた。

4. 結果

4.1. 音声合成の品質

品質を評価するために、クラウドソーシングによるMOSテストを実施しました。評価者はランダムに選んだ音声サンプルを聴き，その自然さを1〜5の5段階で評価した．評価者は各オーディオサンプルを1回ずつ評価することができ，振幅の違いがスコアに影響しないように，すべてのオーディオクリップを正規化した．本研究における品質評価は，すべてこの方法で行った．

評価結果を表1に示す。VITSは他のTTSシステムより優れており、グランドトゥルースと同程度のMOSを達成している。また、Glow-TTSで採用されている確率的持続時間予測器ではなく、決定論的持続時間予測器を採用したVITS（DDP）は、MOS評価においてTTSシステムの中で2番目に高いスコアを獲得した。これらの結果は、1)確率的継続時間予測器は決定論的継続時間予測器よりも現実的な音素継続時間を生成すること、2)我々のエンドツーエンド学習法は、継続時間予測器のアーキテクチャが似ていても、他のTTSモデルよりも優れたサンプルを作成する有効な方法であることを示唆している。

表 1. LJ Speechデータセットにおける評価済みMOSと95%信頼区間との比較。

モデル	MOS (CI)
グラウンド・トゥルース	4.46 (±0.06)
タコトロン2＋HiFi-GAN	3.77 (±0.08)
タコトロン2＋HiFi-GAN（ファインチューニング済み）	4.25 (±0.07)
グロ-TTS＋HiFi-GAN	4.14 (±0.07)
Glow-TTS＋HiFi-GAN（ファインチューニング済み）	4.32 (±0.07)
VITS (DDP)	4.39 (±0.06)
ブイティーエス	**4.43 (±0.06)**

我々は、事前エンコーダの正規化フローと線形スケールスペクトログラムの後置を含む我々の方法の有効性を示すためにアブレーション研究を実施した。アブレーション研究では、すべてのモデルが300kステップまで学習された。その結果を表2に示す。除去

事前エンコーダの正規化フローを変更すると、ベースラインから1.52MOS減少し、事前分布の柔軟性が合成品質に大きく影響することが示された。後段入力のリニアスケールスペクトログラムをメルスペクトログラムに置き換えると、品質劣化（-0.19MOS）となり、VITSにとって高解像度情報が合成品質向上に有効であることが示された。

*表2.*アブレーション試験におけるMOSの比較。

	ModelMOS (CI)
グランドトゥルース4	.50 (±0.06)
ベースライン4	.50 (±0.06)
ノーマライジングフローなし 2	.98 (±0.08)
メルスペクトログラムによる4	.31 (±0.08)

4.2. 多人数音声合成への一般化

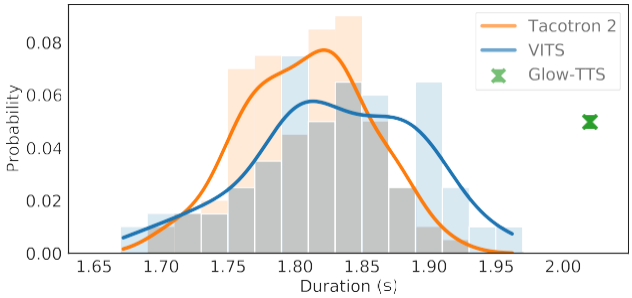
本モデルが多様な音声特性を学習し表現できることを検証するため、多人数音声合成への拡張能力を示したTacotron 2、Glow-TTS、HiFi-GANと比較した（Jia et al. , 2018; Kim et al. , 2020; Kong et al. , 2020）。我々は、VCTKデータセットでmodelを学習させた。セクション 2.5 で説明したように、モデルに話者埋め込みを追加した。Tacotron 2では話者埋め込みをブロードキャストしてエンコーダ出力と連結し、Glow-TTSでは先行研究に従いグローバルコンディショニングを適用した。評価方法は4.1節で述べたものと同じである。表 3 に示すように、我々のモデルは他のモデルよりも高い MOS を達成している。これは、本モデルが様々な音声特性をエンドツーエンドで学習し、表現していることを示している。

*表3.*VCTKデータセットにおける評価済みMOSと95%信頼区間との比較。

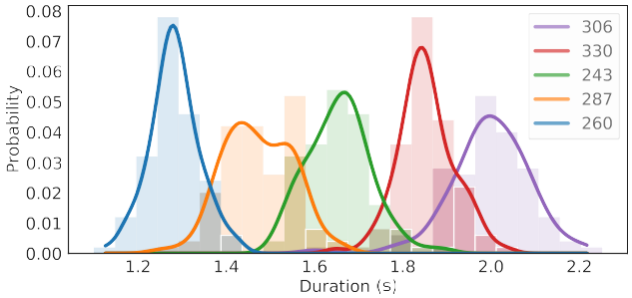
モデル	MOS (CI)
グラウンド・トゥルース	4.38 (±0.07)
タコトロン2＋HiFi-GAN	3.14 (±0.09)
タコトロン2＋HiFi-GAN（ファインチューニング済み）	3.19 (±0.09)
グロ-TTS＋HiFi-GAN	3.76 (±0.07)
Glow-TTS＋HiFi-GAN（ファインチューニング済み）	3.82 (±0.07)
VITS	**4.38 (±0.06)**

4.3. 音声のバリエーション

確率的持続時間予測器が何種類の長さの音声を生成し、合成されたサンプルが何種類の音声特性を持つか検証した。



(a) サンプルの持続時間の比較。Glow-TTSは決定論的持続時間予測器のため、単一の値しか提供しません。



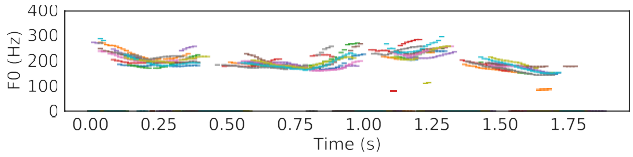
(b) 異なるスピーカーにおけるサンプル時間の比較。

図 2.(a) LJ Speechデータセットと(b) VCTKデータセットにおけるサンプル時間(秒)。

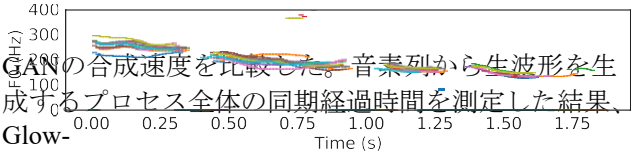
Valle et al. (2021)と同様、ここでのサンプルはすべて、"*How much variation is there? "という文章から生成されたものである。*".図2aは、各モデルから生成された100個の発話の長さのヒストグラムである。Glow-TTSでは決定論的な発話率予測により固定長の発話のみが生成されるのに対し、本モデルではTacotron 2と同様の長さ分布が得られています。図2bは、多人数話者設定において、5人の話者ごとに生成された100個の発話の長さを示しており、モデルが話者依存の音素長を学習していることを示しています。図3は、YINアルゴリズム（De Cheveigne´ & Kawahara, 2002）を用いて抽出した10音声のF0等値から、多様なピッチとリズムを持つ音声を生成していることを示し、図3dは、異なる話者IDで生成した5音声から、話者IDごとに非常に異なる音声長やピッチを表現していることを表しています。Glow-TTSは、事前分布の標準偏差を大きくすることでピッチの多様性を高めることができますが、逆に合成品質を低下させる可能性があることに注意してください。

4.4. 合成速度

本モデルとパラレル2段TTSシステムGlow-TTS、HiFi-



(a) ブイティーエス



GANの合成速度を比較した。音素列から生波形を生成するプロセス全体の同期経過時間を測定した結果、Glow-TTSでは、音素列から生波形を生成するプロセス全体の同期経過時間は1.5秒であった。

(b) タコトロン2

(c) グロ-TTS

(d) VITS（マルチスピーカー）

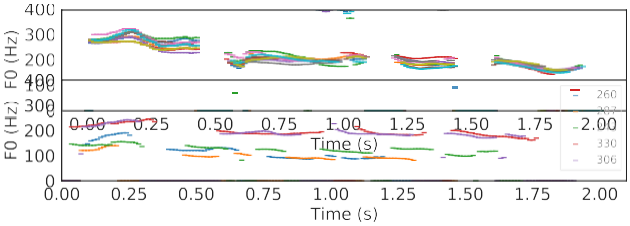


*図3.*How much variation is there? "という発話に対するピッチトラック。サンプルは(a)VITS、(b)Tacotron 2、(c)Tacotron 2から生成されたものである。(c) Glow-TTSは1話者設定、(d)VITSは多人数話者設定から。

のテストセットから無作為に選んだ100文で行った．その結果を表4に示す．本モデルでは、あらかじめ定義された中間表現を生成するモジュールを必要としないため、サンプリングの効率と速度が大幅に改善された。

5. 関連作品

5.1. エンドツーエンド音声合成

現在、2段階パイプラインを持つニューラルTTSモデルは、人間のような音声を合成することができる（Oord et al.）しかし、それらは一般的に第1ステージのモデル出力で訓練または微調整されたボコーダを必要とし、訓練と展開の非効率性を引き起こします。また、あらかじめ定義された中間特徴ではなく、学習された隠れ表現を使用できるエンドツーエンドアプローチの潜在的な利点を享受することができない。

*表4.*合成速度の比較。*n* kHzの速度は、このモデルが1秒間に*n*×*1000*個の生のオーディオサンプルを生成できることを意味します。リアルタイムとは、リアルタイム以上の合成速度を意味する。

機種名回転数	（kHz）	
	リアルタイム	Glow-
TTS＋HiFi-GAN606	.05	×*27*.48
VITS	1480.15	×*67*.12
VITS (DDP)	2005.03	×*90*.93

近年、メルスペクトラムよりも豊富な情報（高周波応答や位相など）を含む生波形をテキストから直接生成するという、より困難なタスクに取り組むために、1ステージのエンドツーエンドTTSモデルが提案されています。FastSpeech 2s (Ren et al. , 2021) はFastSpeech 2の拡張版で、敵対的学習とテキスト再送信の学習を助ける補助的なメルスペクトログラムデコーダを採用し、エンドツーエンドの並列生成を可能にしたものである。しかし、1対多の問題を解決するためには、FastSpeech 2sは学習時の入力条件として、音声から音素の継続時間、ピッチ、エネルギーを抽出する必要があります。EATS (Donahue et al. , 2021) は、同様に敵対的なトレーニングを採用し、微分可能なアライメント方式を採用している。生成された音声とターゲット音声の間の長さのミスマッチ問題を解決するために、EATSは動的プログラミングによって計算されるソフト動的時間ワーピング損失を採用する。Wave Tacotron (Weiss et al. , 2020) は正規化フローとTacotron 2を組み合わせてエンドツーエンド構造を実現しているが、自己回帰的なままである。前述のすべてのエンドツーエンドTTSモデルの音声品質は、2段階モデルより劣る。

前述のエンドツーエンドモデルと異なり、条件付きVAEを利用することで、1）入力条件を追加することなくテキストから直接生波形を合成することを学習し、2）損失計算ではなく動的計画法（MAS）を用いて最適配置を探索し、3）サンプルを並行して生成し、4）公開されている最良の2ステージモデルを上回る性能を実現しています。

5.2. 変分オートエンコーダ

VAE (Kingma & Welling, 2014) は最も広く用いられている尤度ベースの深層生成モデルの一つである．我々は、TTSシステムに条件付きVAEを採用する。条件付きVAEは、観測された条件が、出力を生成するために使用される潜在変数の事前分布を調節する条件付き生成モデルである。音声合成では、Hsuら（2019）、Zhangら（2019）がTacotron 2とVAEを組み合わせて、発話スタイルと韻律を学習しています。BVAE-TTS (Lee et al. , 2021) は、双方向VAE (Kingma et al. , 2016) に基づいてメルスペクトログラムを並列に生成している。前段のモデルにVAEを適用した先行研究とは異なり、我々は並列のエンドツーエンドTTSシステムにVAEを採用する。

Rezende & Mohamed (2015)、Chen et al. (2017)、Ziegler & Rush (2019) は、正規化フローを用いて事前・事後分布の表現力を向上させることでVAE性能を向上させています。事前分布の表現力を向上させるために、条件付き事前ネットワークに正規化フローを追加し、より現実的なサンプルの生成につなげます。

我々の仕事と同様に、Maら（2019）は、非自己回帰型ニューラル機械翻訳のための条件付き事前ネットワークにおける正規化フローを用いた条件付きVAE、FlowSeqを提案した。しかし、我々のモデルが潜在配列と原配列とを明示的にアライメントできる点は、注意メカニズムによって暗黙のアライメントを学習する必要があるFlowSeqとは異なる。本モデルでは、MASを介して潜在配列と時間的に整列した原配列をマッチングさせることで、潜在配列を標準的な正規のrandom変数に変換する負担を取り除き、正規化フローをよりシンプルなアーキテクチャで実現することが可能である。

5.3. 非自己回帰的音声合成における継続時間の予測

自己回帰型TTSモデル（Taigman et al. , 2018; Shen et al. , 2018; Valle et al. , 2021）は、その自己回帰構造と、推論やプライミング時のドロップアウト確率の維持などのいくつかのトリックによって、異なるリズムの多様な音声を生成する（Graves, 2013）。一方、並列TTSモデル（Ren et al. , 2019; Peng et al. , 2020; Kim et al. , 2020; Ren et al. , 2021; Lee et al. , 2021）は、決定論的継続時間予測に依存してきた。これは、パラレルモデルでは、1つのフィードフォワードパスでターゲット音素の継続時間またはターゲット音声の全長を予測しなければならず、音声リズムの相関するジョイント分布を捉えることが困難であるためである。本研究では、推定された音素継続時間の結合分布を学習し、その結果、多様な音声リズムを並列に生成するフローベースの確率的継続時間予測器を提案する。

6. 結論

本研究では、学習と生成をエンドツーエンドで行うことができる並列TTSシステム、VITSを提案した。さらに、音声の2節リズムを表現するために、確率的な継続時間予測器を導入した。その結果、あらかじめ定義された中間音声表現を介することなく、テキストから直接自然な音声波形を合成することが可能となった。実験の結果、本手法は2段階のTTSシステムよりも優れており、人間に近い品質を達成することができた。本手法は、従来二段式音声合成システムが用いられてきた多くの音声合成タスクにおいて、性能向上と学習手順の簡略化のために利用されることが期待される。また、本手法は2つの分離した音声合成パイプラインを統合するものであるが、2つの分離したパイプラインを統合することで、より高い音声合成性能を得ることができることを指摘したい。

は、テキストの前処理という問題が残っている。言語表現の自己教師あり学習の研究は、
、テキスト前処理のステップを削除するための可能な方向性である可能性があります。今後、ソースコードと学習済みモデルを公開し、多くの研究を進めていく予定である。

謝辞

Sungwon Lyu, Bokyung Son, Sunghyo Chung, and Jonghoon Mo
には、有益な議論と助言をいただいた。

参考文献

Bernard, M. Phonemizer. https://github.com/bootphon/phonemizer, 2021.

Bin´kowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K. High Fidelity Speech Synthesis with Adversarial Networks.（逆境ネットワークを用いた高忠実度の音声合成）。In *International Conference on Learning Representations*, 2019.

このような場合、「李舜臣」は、「李舜臣」を「李舜臣」と呼ぶことにする。Vflow: Variational Data Augmentation
を用いたより表現力豊かな生成フロー(Generative Flow)。*機械学習国際会議*, pp.1660-1669.PMLR, 2020.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational lossy autoencoder. 2017. URL https://openreview.net/forum?id=BysvGP5ee。

De Cheveigne´, A. and Kawahara, H. Yin, a fundamental frequency estimator for speech and music.*The Journal of the Acoustical Society of America*, 111(4):1917-1930, 2002.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using Real NVP.In *International Conference on Learning Representations*, 2017.

Donahue, J., Dieleman, S., Binkowski, M., Elsen, E., and Simonyan, K. End-to-end Adversarial Text-to-Speech.In *International Conference on Learning Representations*, 2021.URL https://openreview.net/forum?id=rsf1z-JSj87.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural Spline Flows（ニューラル・スプライン・フロー）.In *Advances in Neural Information Processing Systems*, pp.7509-7520, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.Generative Adversarial Nets.*神経情報処理システムにおける進歩*, 27:2672-2680, 2014.

Graves, A. Generating sequences with recurrent neural networks（リカレントニューラルネットワークによるシーケンスの生成）, *arXiv preprint arXiv:1308.0850*, 2013.

本論文では、Flow++を用いた、フローベースの生成モデルの改良について述べる。In *International Conference on Machine Learning*, pp.2722- 2730.PMLR, 2019.

Hsu, W.-N., Zhang, Y., Weiss, R., Zen, H., Wu, Y., Cao, Y., and Wang, Y. Hierarchical Generative Modeling for Controllable Speech Synthesis.（制御可能な音声合成のための階層的生成モデリング）。In *International Confer- ence on Learning Representations*, 2019.URL https://openreview.net/forum?id=rygkk305YQ。

Ito, K. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez-Moreno, I., et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis（話者検証から多言語音声合成への学習移行）.In *Advances in Neural Information Processing Systems*, 2018.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. Efficient neural au- dio synthesis（効率的なニューラルオディオ合成）.In *International Conference on Machine Learning*, pp.2410-2419.PMLR, 2018.

このような場合、「音声合成のためのフロー」は、「音声合成のためのフロー」と「音声合成のためのフロー」とに分けられる。*神経情報処理システムにおける進歩*, 33, 2020.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes.In *International Conference on Learning Repre- sentations*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational in- ference with inverse autoregressive flow.（逆自己回帰フローによる改良型変分法）. *神経情報処理システムにおける進歩*, 29:4743-4751, 2016.

HiFi-GAN: Generative Ad-versarial networks for Efficient and High Fidelity Speech Synthesis（HiFi-GAN：効率的で忠実な音声合成のための生成アドバーサリーネットワーク）.*神経情報処理システムの進歩*, 33, 2020.

クマール、K、クマール、R、ド・ボワジエール、T、ゲスティン、L、テオ。W.MelGAN: Generative Adversarial Net-works for Conditional Waveform Synthesis（メルガン：条件付き波形合成のための生成的逆行列ネットワーク）」 第 32 巻, pp.14910-14921, 2019.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. 学習したものを用いた画素を超えた自動エンコーディング。

類似性メトリック。In *International Conference on Ma- chine Learning*, pp.1558-1566.PMLR, 2016.

Lee, Y., Shin, J., and Jung, K. Bidirectional Variational Inference for Non-Autoregressive Text-to-speech.In *International Conference on Learning Representations*, 2021.URL https://openreview.net/forum?id=o3iritJHLfO.

李娜、劉斯、劉耀、趙斯、劉茂、変換ネットワークを用いたニューラル音声合成。
pp.6706-6713, 2019.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization.In *International Conference on Learning Representations*, 2019.URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ma, X., Zhou, C., Li, X., Neubig, G., and Hovy, E. Flowseq: Generative Flowによる非自己回帰的条件付きシーケンス生成.*2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.4273-4283, 2019 *に収録されています。*

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks.In *Proceedings of the IEEE international conference on computer vision*, pp.2794-2802, 2017.

Miao, C., Liang, S., Chen, M., Ma, J., Wang, S., and Xiao.（ミャオ、C.、リャン、S.、チェン、マ、J.、ワン、S.、シャオ
J.Flow-TTS: A non-autoregressive network for text to speech based on flow.*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.7209-7213 に掲載。IEEE, 2020.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Peng, K., Ping, W., Song, Z., and Zhao, K. Non-autoregressive neural text-to-speech.In *International Conference on Machine Learning*, pp.7586-7598.PMLR, 2020.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. Deep Voice 3: 2000- Speaker Neural Text-to-Speech.（英語）。In *International Confer- ence on Learning Representations*, 2018.URL https://openreview.net/forum?id=HJtEm4p6Z。

Prenger, R., Valle, R., and Catanzaro, B. Waveglow:A flow-based generative network for speech synthesis（音声合成のためのフローベースの生成ネットワーク）。In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3617-3621.IEEE, 2019.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, and Liu, T.-Y.（任洋洋、阮洋、譚暁、趙洙、趙洙、劉垠）。FastSpeech:Fast, Robust and Controllable Text to Speech. volume 32, pp.3171-3180, 2019.

秦、T、趙、S、Zhao、Liu、T.Y.。また、このような場合、「李舜臣」氏は、「李舜臣」氏と同じように、「李舜臣」氏を「李舜臣」氏と呼ぶことにする。In *International Conference on Learning Representations*, 2021.URL https://openreview.net/forum?id=piLPYqxtWuA.

このような場合、「曖昧模糊」と呼ばれる。In *International Conference on Ma- chine Learning*, pp.1530-1538.PMLR, 2015.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-Attention with Relative Position Representations（相対位置表現を用いた自己注意）.*計算言語学会の北米支部の2018年大会の議事録*にて。*Human Language Technologies, Volume 2 (Short Papers)*, pp.464-468, 2018.

シェン、J、パン、R、ワイス、R.J、シュスター、M、ジャイトリー、N,
Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.4779-4783.IEEE, 2018.

Taigman, Y., Wolf, L., Polyak, A., and Nachmani, E. Voiceloop:Phono-logical LoopによるVoice FittingとSynthesis。In *International Conference on Learning Representations*, 2018.URL https://openreview.net/forum?id=SkFAWax0-.

Valle, R., Shih, K. J., Prenger, R., and Catanzaro, B.Flowtron: an Autoregressive Flow-based Gener-ative Network for Text-to-Speech Synthesis (フロートロン：音声合成のための自己回帰フローに基づく生成ネットワーク).*In-ternational Conference on Learning Representations*, 2021.URL https://openreview.net/forum?id=Ig53hpHxS4.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning.In *Advances in Neural Information Processing Systems*, pp.6309-6318, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Atten- tion is All you Need.（アテンションはすべて必要である）。*Adv ances in Neural Information Processing Systems*, 30:5998-6008, 2017.

Vaux, C., Yamagishi, J., MacDonald, K., et al. CSTR VCTKコーパス:CSTRボイスクローニングツールキット用英語多人数話者コーパス.エジンバラ大学。*The Cen- tre for Speech Technology Research (CSTR)*, 2017.

Weiss, R. J., Skerry-Ryan, R., Battenberg, E., Mariooryad, S., and Kingma, D. P. Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis. *arXiv preprint arXiv:2011.03568*, 2020.

曽, Z., 王, J., 程, N., 夏, T., 蕭, J. Alignttts:を使用した効率的なフィードフォワード音声合成システム。*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Pro- cessing (ICASSP)*, pp.6714-6718 に掲載。IEEE, 2020.

Zhang, Y.-J., Pan, S., He, L., and Ling, Z.-H.エンドツーエンド音声合成におけるスタイル制御と転送のための潜在的な表現の学習。In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Sig- nal Processing (ICASSP)*, pp.6945-6949.IEEE, 2019.

Ziegler, Z. and Rush, A. Latent normalizing flows for discrete sequences（ジークレット・シーケンスに対する潜在的正規化フロー）.In *International Conference on Machine Learning*, pp.7673-7682.PMLR, 2019.

の補足資料です。
エンド・ツー・エンド音声合成のための逆説的学習を用いた条件付き変分オートエンコーダ

A. モノトニックアライメントサーチ

図4はMASの擬似コードである。データの厳密な対数尤度ではなくELBOを最大化するアライメントを探索するが、2.2.1節で述べたGlow-TTSのMAS実装を利用することが可能である。

```
def monotonic_alignment_search(value)。
    """与えられた対数尤度行列に対して最も可能性の高いアライメントを返します。Args:
        値: 対数尤度行列.その (i, j)-番目のエントリは、j番目の潜在変数の対数尤度を
        含みます。
        は, 与えられたi番目の事前平均と事前分散に対応する。
        ... 数学::
            value_{i,j} = log N(f(z)_{j}; \mu_{i}, \sigma_{i})
        (dtype=float, shape=[text_length, latent_variable_length])
    リターンです。
        path: 最も可能性の高いアライメント。
        (dtype=float, shape=[text_length,latent_variable_length])。
    """
    t_x, t_y = value. shape # [text_length, letent_variable_length].
    パス = zeros([t_x, t_y])

    # これまでの最尤アライメントに対する対数尤度を保存するキャッシュ。
    Q = -INFINITY * ones([t_x, t_y])

    for y in range(t_y):
        for x in range(max(0, t_x + y - t_y), min(t_x, y + 1)):
            if y == 0: # 基本的な場合。y が 0 の場合, 可能な x の値は 0 だけである.
                Q[x, 0] = value[x, 0] です。
            を追加しました。
                if x == 0:
                    v_prev = -INFINITY
                を追加しました。
                    v_prev = Q[x-1, y-1]
                v_cur = Q[x, y-1]である。
                Q[x, y] = value[x, y] + max(v_prev, v_cur)

    # 最後の観測からバックトラックを行う。
    インデックス = t_x - 1
    for y in range(t_y - 1, -1, -1):
        path[index, y] = 1
        index != 0 かつ (index == y または Q[index, y-1] < Q[index-1, y-1]) の場
            合： index = index - 1

    リターンパス
```

*図4.*モノトニックアライメントサーチのシュードコード。

B. モデル構成

本節では、Glow-TTSやHiFi-GANの構成を踏襲しつつ、VITSに新たに追加した部分を中心に説明する。エンコーダやWaveNet残差ブロックはGlow-TTSと同じものを用い、デコーダや多周期判別器はGlow-TTSの生成器や多周期判別器と同じである。

HiFi-GANは、デコーダの入力次元が異なることと、副判別器を追加していることを除いて、それぞれ、HiFi-GANと同じである。

B.1. 事前エンコーダと事後エンコーダ

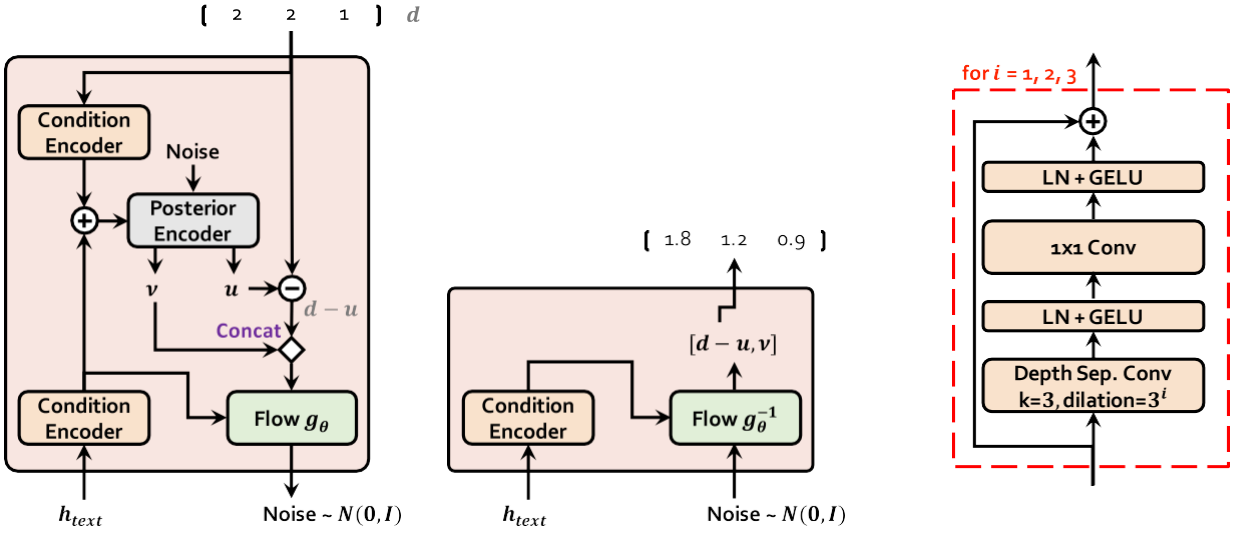
先行エンコーダの正規化フローは、4つのアフィン結合層のスタックで、各結合層は4つのWaveNet残差ブロックから構成されています。アフィン結合層は体積保存変換であると制限しているため、結合層はスケールパラメータを生成しない。

16個のWaveNet残差ブロックからなる事後エンコーダは、リニアスケールの対数振幅スペクトログラムを取り込み、192チャンネルの潜在変数を生成する。

B.2. デコーダとディスクリミネータ

このデコーダの入力は事前または事後エンコーダから生成された潜在変数であるため、デコーダの入力チャネルサイズは192である。デコーダの最後の畳み込み層では、バイアスパラメータを削除する。これは、混合精度学習時に不安定な勾配スケールを引き起こすためである。

識別器には、周期[2,3,5,7,11]の5つのサブ識別器からなるマルチ周期識別器と3つのサブ識別器からなるマルチスケール識別器を用いている。学習効率を上げるため、マルチスケール識別器のうち、生波形で動作する最初の部分識別器のみを残し、平均化された波形で動作する2つの部分識別器を破棄する。この判別器は、周期[1, 2, 3, 5, 7, 11]のマルチ周期判別器と見なすことができる。



(a) 学習手順
希釈された深さ方向に分離可能なコンボ

(b) 推論手順

(c)
きけんじょうブロック

*図 5.*確率的持続時間予測器の(a)学習手順と(b)推論手順を示すブロックダイアグラム。確率的持続時間予測器の主な構成要素は、（c）拡張され、深さ方向に分離可能な畳み込み残差ブロックである。

B.3. ストキャスティックス・デュレーション・プレディクター

図5aおよび図5bは、それぞれ確率的持続時間予測器の学習と推論の手順を示している。確率的持続時間予測器の主な構成要素は、図5cに示すように、拡張および深さ方向に分離可能な畳み込み（DDSConv）残差ブロックである。DDSConvブロックの各畳み込み層は、層正規化層とGELU活性化関数に続いている。我々は、大きな受容野サイズを維持しながらパラメータ効率を向上させるために、拡張および深さ方向に分離可能な畳み込み層を使用することを選択した。

継続時間予測器の事後エンコーダと正規化フローモジュールは、フローベースのニューラルネットワークであ

り

のようなアーキテクチャを持つ。違いは、後置エンコーダがガウス雑音列を2つのランダムな変数 v と u を用いて近似的な事後分布 $q_\varphi(u, v|d, c_{text})$ を表現し、正規化フローモジュールは $d - u$ と v をガウス雑音列に変換して、セクション 2.2.2 で述べたように増大・脱量子化データログ $p_\theta(d - u, v|c_{text})$ の対数尤度を表現しています。

すべての入力条件は条件エンコーダで処理され、それぞれ2つの1x1畳み込み層とDDSConv残差ブロックから構成される。事後エンコーダと正規化フローモジュールは4つのカップリング層からなるニューラル・スプラインフローを持つ。各結合層はまずDDSConvブロックを通して入力と入力条件を処理し、10個の有理二次関数を構築するために使われる29チャンネルのパラメータを生成する。すべてのカップリング層と条件エンコーダの隠れ次元を192に設定した。図 6a と 6b は確率的持続時間予測器に使用される条件エンコーダとカップリング層のアーキテクチャーを示す。

(a) 確率的持続時間予測装置における条件エンコーダ　　(b) 確率的持続時間予測装置におけるカップリング層

*図 6.*確率的持続時間予測器に用いられる(a)条件エンコーダと(b)結合層のアーキテクチャ。

C. サイドバイサイド評価

50項目に対して500件の評価を行い、VITSとグランドトゥルースの7点比較平均意見スコア(CMOS)評価を実施した。その結果、表5に示すように、LJ Speechデータセットと VCTKデータセットにおいて、それぞれ-0.106と-0.270のCMOSを達成した。この結果は、我々のモデルがGlow-TTSやHiFi-GANといった一般に公開されている最高のTTSシステムよりも優れており、MOS評価においてグランドトゥルースと同等のスコアを達成したにもかかわらず、評価者が我々のモデルよりもグランドトゥルースを好む傾向がわずかに残っていることを示している。

*表 5.*VITSのCMOSをグランドトゥルースと比較して評価した。

データセット	CMOS
LJ スピーチ	-0.106
VCTK	-0.262

D. 音声変換

多人数音声の場合、話者識別情報をテキストエンコーダーに与えないため、テキストエンコーダーから推定される潜在変数が話者非依存表現を学習する。話者非依存表現を用いることで、ある話者の音声記録を別の話者の音声に変換することができる。与えられた話者ID sとその話者の発話に対して、対応する発話の音声から線形スペクトログラムx_{lin} を得ることができる。x_{lin} は、事後エンコーダと事前エンコーダの正規化フローによって、話者非依存な表現eに変換できる。

$$z \sim q_\varphi(z|x_{lin}, s) \tag{12}$$

$$e = f_\theta(z|s) \tag{13}$$

そして、正規化フローf^{-1}とデコーダGの逆変換により、表現eから対象話者ID $\hat{s}$の音声$\hat{y}$を合成することができる。

$$\hat{y}=G(f_{\theta}^{-1}(e|\hat{s})|\hat{s}) \tag{14}$$

話者非依存表現を学習し、それを音声変換に用いることは、Glow-TTSで提案された音声変換手法の延長線上にあると考えることができる。本手法では、Glow-TTSのようなメルスペクトログラムではなく、生の波形を提供する。音声変換の結果は図7に示す通りである。これは、ピッチレベルの異なるピッチトラックが同じような傾向を示している。

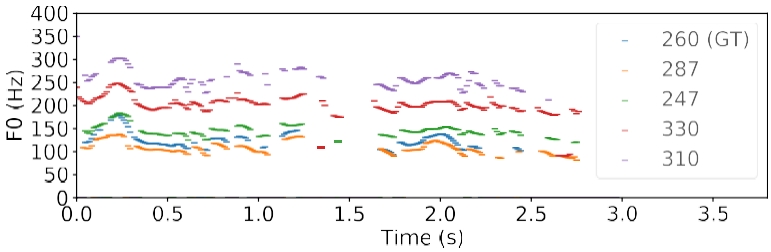


*図7.*異なる話者IDを持つグランドトゥルースサンプルと対応する音声変換サンプルのピッチトラック。