The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

変分自己符号化器を用いた表現の多様性のモデル化による表現豊かな音声合成

Expressive Speech Synthesis through modeling the variety of expressions by Variational Autoencoder

阿久澤 圭 *1 Kei Akuzawa　岩澤 有祐 *1 Yusuke Iwasawa　松尾 豊 *1 Yutaka Matsuo

*1東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

Recent advancements in the deep autoregressive generative modeling improve the performance of speech synthesis (SS). However, how to equip the expressiveness into the deep autoregressive based SS-system is an open issue due to the lack of ability to model the global characteristics of speech (such as speaker individualities or speaking styles). In this paper, we propose a model called VAE-Loop, which integrates variational autoencoder (VAE) with VoiceLoop: one of the autoregressive based speech synthesis models. Unlike the traditional SS with autoregressive modeling, the proposed method explicitly model the global characteristic of speech by VAE, enabling control of the expressiveness of the synthesized speech. Experiments on VCTK and Blizzard2012 showed that VAE helps VoiceLoop to generate higher quality speech and control expressions through learning the global characteristics.

1. はじめに

我々は自然な会話の中で様々な感情，話し方，話者の年齢や性別 (以下話者性と呼ぶ) などによって特徴付けられた表現豊かな音声を発する．表現の豊かさは，近年深層学習を基盤とした音声合成モデルが向上させた自然さと明瞭さに加えて，音声合成システムが備えるべき性質の一つとされている (例えば [Eyben 12])．表現豊かな音声合成を実現するための典型的な手法は感情や話し方，話者性のラベルを用いることである (例えば [Luong 17]) が，本研究はそうした教師ラベルなしでの表現豊かな音声合成に取り組む．なぜなら，表現豊かな音声は動画投稿サイトやオーディオブックからの入手が容易であるが，その音声に対応するラベルを得るにはコストがかかるし，ヒューリスティックに定めた感情や話し方のカテゴリを信頼できない場合もある．しかし，ラベルなしでの表現豊かな音声合成は一つのテキストに対して数多くの正解音声が存在するという困難な逆問題に直面する．

近年音声合成全般及びラベルなしでの表現豊かな音声合成において，自己回帰生成モデルが重要な役割を果たしている．例えば [Oord 16] は，既存の統計的または波形接続型の音声合成システムから音声の自然さと明瞭さを向上させた．また，自己回帰モデルを用いた sequence-to-sequence モデル (seq2seq モデル) の VoiceLoop[Taigman 18] は，youtube から入手した多様な韻律を含む音声を用いて音声合成モデルを訓練できることを初めて示した．

しかしラベルなしでの表現豊かな音声合成に自己回帰生成モデルを用いることは 2 つの問題を抱えている．第一に，自己回帰生成モデルはラベルなしでデータの大域的な特徴を学習するのが不得手である (例えば [Gulrajani 17])．話者性や感情などが発話を大域的に特徴づけていることを考えれば，自己回帰生成モデルも前述の逆問題に十分に対抗できない可能性がある．第二に，自己回帰生成モデルは変分自己符号化器 (Variational Autoencoder, VAE)[Kingma 14] といった他の深層生成モデルと違って潜在変数を持たないので，生成する音声の特徴をコントロールすることが難しい．

そこで本研究では前述の自己回帰生成モデルの問題を克服し，自然さ，明瞭さに加えて表現の豊かさを備えた音声合成をラベルなしで実現するために，VoiceLoop を VAE に組み込んだモデル VAE-Loop を提案する．VAE は話者性といった音声データの大域的な特徴の学習に有効であることが近年示されてきた (例えば [Hsu 17]) が，それを音声合成に適用することを提案したのは，我々の知る限り本研究が初めてである．VAE-Loop では，VAE を用いて獲得した潜在変数を自己回帰生成モデルへの入力とすることで，データの大域的特徴を組み入れた生成が行える．また話者性や話し方の潜在表現を学習することで，生成する音声の特徴を柔軟にコントロールできる．

実験ではまず，VoiceLoop が VAE によって獲得した大域的な特徴を利用することで困難な逆問題を緩和し，より高品質な音声を生成できることを示す．加えて，獲得した潜在変数を用いて VAE-Loop が話者性や話し方をコントロールする能力を持つことを示す．

本研究の貢献は，自己回帰生成モデルをラベルなしでの表現豊かな音声合成に適用する際の問題点を克服するために，VAE が獲得した大域的な特徴を利用することを提案したことである．この手法により、高コストなラベルありデータが得られないより現実的な問題設定においても，高品質で多様な表現を持つ音声の生成が可能になる．

2. 関連研究

深層学習を基盤にした seq2seq モデルによる音声合成モデルは，テキストから音響特徴量を直接予測する単純な構造を持ち，また自然で明瞭な音声を生成することが示されている．(例えば [Wang 17b][Taigman 18]) 特に VoiceLoop[Taigman 18] は多様な韻律を含む youtube から入手した音声を用いて訓練できることが示されているので，ラベルなしでの表現豊かな音声合成に有効であると考えて提案モデルの一部に組み入れた．

また [Taigman 18] や [Wang 17a] は，話し方を教師なしで学習しコントロールできる seq2seq モデルを提案した．これらの手法と違い，本研究が用いる VAE は背後に潜在表現の明確な定式化があり，話し方の潜在表現を確率分布として陽に学習することができる．その結果，簡単で柔軟な話し方のコント

連絡先: 阿久澤圭，東京大学大学院工学系研究科技術経営戦略学専攻，〒 113-8656 東京都文京区本郷 7-3-1，akuzawa-kei@weblab.t.u-tokyo.ac.jp

ロールが可能になる．

seq2seq モデルに限定しないラベルなしでの表現豊かな音声合成の研究には [Eyben 12] などがある．こうした研究では，クラスタリングなどの教師なし学習の手法で獲得した感情の表現を訓練に用いることが多い．これらの手法では表現の学習と音声合成モデルの訓練を別々に行う必要があるが，提案モデルは一括で訓練を行うことができる．

VAE を音声の生成に利用した研究には [Hsu 17] などがある．これらの研究は VAE によって音声データが持つ話者性などの特徴を潜在変数として獲得できることを示した．これらの研究と異なり，本研究では音声合成を目的としてテキストによる条件付けを行う．また音声以外の領域では，本研究と同様に VAE と自己回帰生成モデルを組み合わせる研究が多くある (例えば [Gulrajani 17])．

3. モデル

本節では，まず提案モデルの基盤となる VAE と VoiceLoop について解説し，その後提案モデルについて述べていく．

3.1 Variational Autoencoder

VAE は Kingma らによって提案された深層生成モデルである [Kingma 14]．VAE では観測変数 $\boldsymbol{x}$ と潜在変数 $\boldsymbol{z}$ のデータ生成過程を，確率密度分布のモデルパラメータ $\boldsymbol{\theta}$ を用いて $\boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ のように定める．すると，真の事後分布 $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ を近似した分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$($\boldsymbol{\phi}$ はモデルパラメータ) を用いて，周辺尤度の下界 $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ が以下のように求まる．

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) \\ &\geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) \\ &= -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]\end{aligned}$$

そして訓練では，下界 $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$ が最大になるようにパラメータ $\boldsymbol{\theta}, \boldsymbol{\phi}$ について最適化を行う．本稿では $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ をエンコーダー，$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ をデコーダーと呼ぶ．

3.2 VoiceLoop

VoiceLoop は，Taigman らによって提案された seq2seq モデルである [Taigman 18]．$\boldsymbol{x}_1, ..., \boldsymbol{x}_T$ を予測したい音響特徴量の可変長の系列とする．VoiceLoop は条件付き独立性

$$p_{\boldsymbol{\xi}}(\boldsymbol{x}|\boldsymbol{c}) = \prod_{t=1}^{T} p_{\boldsymbol{\xi}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{c}) \quad (1)$$

$$p_{\boldsymbol{\xi}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{c}) \sim \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_{\boldsymbol{\xi}}(\boldsymbol{x}_{1:t-1}), \boldsymbol{I}) \quad (2)$$

を利用して，各タイムステップ $t \in T$ ごとに $p_{\boldsymbol{\xi}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{c})$ を逐次的に推定する．ここで (2) 式は，モデルパラメータ $\boldsymbol{\xi}$ とタイムステップ t 以前の音響特徴量 $\boldsymbol{x}_{1:t-1}$ によって決まる変数 $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ を平均，単位行列を分散とするガウス分布に確率変数 $\boldsymbol{x}_t$ が従うことを仮定している．

$p_{\boldsymbol{\xi}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{c})$ の推定は以下の手順に従う．VoiceLoop はバッファと呼ばれる潜在変数 $S \in \mathbb{R}^{d \times k}$ を持つ．S はそれぞれが d 次元の変数である $S[1]...S[k]$ から成り立っている．それぞれのタイムステップ t で，バッファは次のように変化する．

$$\begin{aligned}S_t[i+1] &= S_{t-1}[i] \quad if \;\; 1 \leq i < k \\ S_t[1] &= u\end{aligned}$$

ここで u は，テキスト情報 $\boldsymbol{c}_t$，音響特徴量 $\boldsymbol{x}_{t-1}$，話者の埋め込みベクトル s，およびバッファ自身を用いて，以下のように決まる変数である．

$$\begin{aligned}C_t &= [\boldsymbol{c}_t + tanh(F_u(s)), \boldsymbol{x}_{t-1}] \\ u &= N_u([S_{t-1}, C_t])\end{aligned} \quad (3)$$

そして，VoiceLoop はバッファ S_t と話者の埋め込みベクトル s を用いて音響特徴量の予測値 $\hat{\boldsymbol{x}}_t$ を以下のように出力する．

$$\hat{\boldsymbol{x}}_t = N_o(S_t + F_o(s)) \quad (4)$$

ただし，F_u, N_u, F_o, N_o はそれぞれ全結合層，または全結合層と活性化関数を多層重ねたものである．また，$\hat{\boldsymbol{x}}_t$ は (2) 式の $\boldsymbol{\mu}_{\boldsymbol{\xi}}(\boldsymbol{x}_{1:t-1})$ と同値である．

3.3 提案モデル: VAE-Loop

VAE-Loop は，VAE を補助特徴量 $\boldsymbol{c}$ で条件づけたモデル $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c})$ を出発点とする．本研究においては $\boldsymbol{x}$ が音響特徴量，$\boldsymbol{c}$ が音素に対応する．そして，その際の尤度の下界は以下のように書ける．

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{c}) = &-D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})]\end{aligned} \quad (5)$$

ここで，本研究では簡単化のために $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ と仮定した．そして，事前分布 $p(\boldsymbol{z})$，近似事後分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ はそれぞれ以下の正規分布に従うと仮定する．

$$p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}), \;\; q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}\boldsymbol{I})$$

ここで，$\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\sigma}_{\boldsymbol{\phi}}$ はパラメータ $\boldsymbol{\phi}$ とエンコーダーの入力 $\boldsymbol{x}$ によって決まる変数である．

VAE-Loop では，VoiceLoop によるテキスト $\boldsymbol{c}$ から音響特徴量 $\boldsymbol{x}$ の生成過程を，VAE におけるデコーダー $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$ と見なす．VoiceLoop をデコーダーとして用いるために，(1),(2),(3) 式を以下のように置き換える．

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{z}, \boldsymbol{c}) \quad (6)$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{z}, \boldsymbol{c}) \sim \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:t-1}, \boldsymbol{z}), \boldsymbol{I}) \quad (7)$$

$$u = N_u([S_{t-1}, C_t, \boldsymbol{z}]) \quad (8)$$

(8) 式では，VoiceLoop による音声の生成過程に潜在変数 $\boldsymbol{z}$ を組み入れている．$\boldsymbol{z}$ は S_{t-1}, C_t と違いタイムステップ t に依存しない大域的な特徴であり，自己回帰生成モデルの欠点を補っている．

$\boldsymbol{z}$ は訓練時にエンコーダー $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ からサンプリングされ，音声の大域的な特徴を学習する．エンコーダーは通常の VAE と同様に多層のニューラルネットワーク (Deep Neural Networks, DNN) でモデル化する．一方推論時には，$\boldsymbol{z}$ は単純な事前分布 $p(\boldsymbol{z})$ を介してサンプリングされる．VAE-Loop の概念図は図 1 で確認できる．

提案モデルの目的関数は，(5), (6) 式を用いて以下のように書ける．

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{c}) = &-D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + \\ &\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\sum_{t=1}^{T} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1}, \boldsymbol{z}, \boldsymbol{c})]\end{aligned} \quad (9)$$

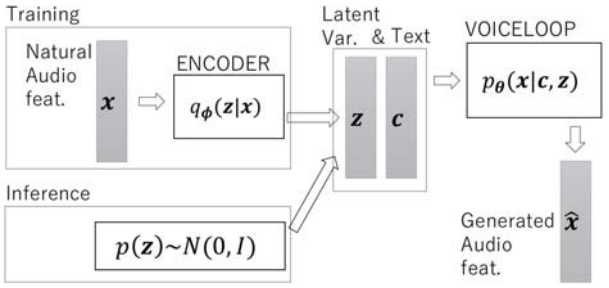

図 1: 提案モデル VAE-Loop の概念図．訓練時はエンコーダーから，推論時は事前分布から潜在変数のサンプリングを行う．

ここで (9) 式の第一項は正則化項，第二項は再構成誤差である，第二項は VoiceLoop が推定した $\hat{\boldsymbol{x}}_t$ と真の $\boldsymbol{x}_t$ の間で平均二乗誤差をとることで推定できる．つまり，$\boldsymbol{z}$ を用いる点を除いて通常の VoiceLoop の目的関数と同様である．

4. 実験

4.1 データセット

本研究では，複数話者のデータセットと，様々な感情や話し方を含むデータセットの二つを訓練に用いた．複数話者のデータセットとして，109 人の英語話者の音声を集めた VCTK Corpus[Veaux 17](以下 VCTK と呼ぶ) を用いた．[Taigman 18] とデータの前処理などの条件を揃えるために，VCTK に含まれるすべてのデータではなく，VoiceLoop のソースコードページ *1 から入手できる男性 4 人女性 17 人計 21 人の北米の話者による合計約 5 時間の音声データを用いた．それぞれの音声の長さは 5 秒以下である．

二つ目のデータセットとして，Blizzard Challenge2012[King 12] で使用された，4 編のオーディオブックからなるデータセット (以下 Blizzard2012 と呼ぶ) を用いた [Braunschweiler 10]．オーディオブックは多様な感情や話し方を含むことから，表現豊かな音声合成の研究によく利用される．Blizzard2012 は VCTK と違って全て同一の男性話者によって朗読されている．VCTK と条件を揃えるため，また勾配爆発を避けるため長さが 5 秒以下の音声のみを用いたところ，全体で約 10 時間の音声となった．

二つのデータセットに共通の前処理として冒頭の無音期間の削除を行った．また，全体の 9 割を訓練データとして，残りの 1 割を検証データと 50 サンプルのテストデータに分けた．実験に用いる特徴量は [Taigman 18] と同様に，モデルへの入力を音素，モデルが出力する音響特徴量を 60 次元のメルケプストラム，基本周波数 (F0)，非周期性成分，有声・無声フラグとした．

4.2 実験設定

実験では VoiceLoop と提案モデル VAE-Loop について比較を行った．提案モデルのエンコーダーには，時間方向への畳み込みを基盤とした DNN を用いた．具体的には，ストライド幅 2 の畳み込み，ドロップアウト，バッチ正規化，ReLU を 5 層繰り返した後に，可変長の音響特徴量を扱うための時間方向の MaxPooling，全結合層を重ねた構造を用いた．一方，提案モデルのデコーダーおよびベースラインに用いた VoiceLoop のハイパーパラメータは，[Taigman 18] での著者らによる実装と同様とした．

訓練時には最適化アルゴリズムを Adam，エポック数を 150 とした．学習率は [1e-3, 1e-4, 5e-5, 1e-5] の中から検証誤差が最小になるものを選んだところ，VCTK では 1e-4，Blizzard2012 では 5e-5 となった．

*1 https://github.com/facebookresearch/loop/

表 1: MOS (Mean±CI)

Method	VCTK	Blizzard2012
Ground Truth(1)	4.07 ± 0.23	3.94 ± 0.30
VoiceLoop(w/o labels)	2.51 ± 0.34	2.23 ± 0.24
VoiceLoop(w/ labels)	3.24 ± 0.27	N/A
VAE-Loop($\sigma = 1$)	**3.25 ± 0.29**	2.47 ± 0.32
VAE-Loop($\sigma = 0.7$)	N/A	**2.89 ± 0.32**
VAE-Loop($\sigma = 0$)	N/A	**3.03 ± 0.32**

4.3 MOS による評価

本節では大域的な特徴のラベルが得られないときも VAE-Loop が逆問題を緩和して高品質な音声を生成する能力を持つと示すために，crowdMOS toolkit[Ribeiro 11] と Amazon Mechanical Turk を用いた Mean Opinion Score(MOS) による評価を行った．MOS とは，評価者がアルゴリズムによって生成された音声を聞いて「人間の音声のように自然に聞こえるか」に基づいて 5 段階評価でつける点数である．評価者はそれぞれのモデルごとに 50 個の音声サンプルを聞いて評価した．評価者はアメリカ在住者に限定し，二つのデータセットに対してそれぞれ 15 人以上の評価者が参加した．表 1 は，MOS とその 95%信頼区間 (Confidence Interval, CI) を表している．ただし，表 1(1) は Vocoder によって再構成した後の音声サンプルを，MOS の上限として用いている．

表 1 より，VCTK において提案モデルの評価がベースラインの評価を上回ることが観測された．さらに，VAE-Loop は話者ラベルを用いていないにも関わらず，話者ラベルを用いる場合の VoiceLoop と同レベルの評価を得ていることが観測された．また我々の非公式な試聴では，VAE-Loop がベースラインに比べて明瞭な音声の生成に失敗する（例えば，吐息やある特定の音素を数秒間生成し続ける）ことが少ないと観測された．したがってこの MOS の差は，VAE-Loop が話者性という音声の大域的な特徴を学習することで困難な逆問題を軽減したことを示唆していると考えられる．

Blizzard2012 では，テストサンプルの生成に使う $p(\boldsymbol{z})$ が高い分散を持つとき，VAE-Loop も VoiceLoop と同様に明瞭な音声の生成に失敗する現象が度々確認された．そこで，$p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \sigma^2\boldsymbol{I})$ として，様々な σ を用いて $\boldsymbol{z}$ のサンプリングを行った．ただし，$\sigma = 0$ とは常に $\boldsymbol{z} = 0$ をサンプリングすることを指す．その結果，$p(\boldsymbol{z})$ の分散を抑えると提案モデルの MOS が向上し，ベースラインを上回ることが観測された．ここで，σ を小さくすると毎回似たような $\boldsymbol{z}$ がサンプリングされるので，σ を変えることは安定した音声の生成と音声の多様性の間でトレードオフを取ることだと捉えることができる．

4.4 潜在変数を用いた音声の特徴のコントロール

本節では，提案モデルが潜在変数を用いて生成する音声の特徴をコントロールできることを示す．図 2 は，VCTK を用いて訓練した VAE-Loop が生成した音声の F0 の軌跡を表している．左右の図では，別々の発話文を同一の潜在変数 $\boldsymbol{z}$ を用いて生成している．図 2 で $\boldsymbol{z}$ の値に応じて F0 の高さが決まることから，提案モデルが女性らしい高い声，男性らしい低い声といった特徴を潜在変数を用いてコントロールできることが確認できた．また，二つの $\boldsymbol{z}$ の中間の値を用いて音声を生成すると，F0 の値も平均化されることが確認できた．図 3 は，Blizzard2012 を訓練に用いたときの F0 の軌跡を表している．VCTK と同様に，潜在変数の値が F0 の軌跡を特徴付けてい

ることが確認できた．音声の例は次の URL で試聴できる．: https://akuzeee.github.io/VAELoopDemo/

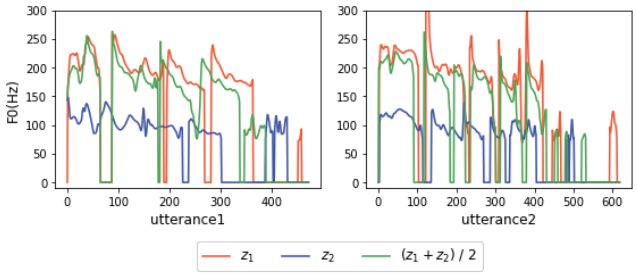

図 2: VCTK で訓練した VAE-Loop を用いた，二つの発話文に対する F0 の軌跡．z_1 は高い声 (我々の試聴によれば女性の声)，z_2 は低い声 (男性の声) を生成する．また二つの潜在変数の平均をとることで，F0 の値も平均化される．

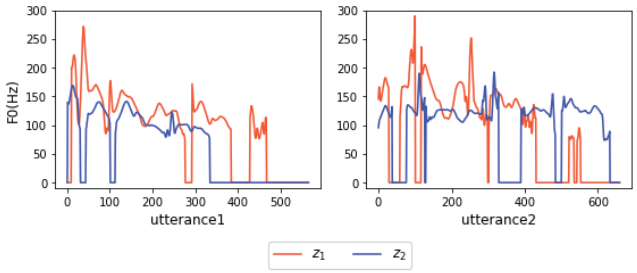

図 3: Blizzard2012 で訓練した VAE-Loop を用いた，二つの発話文に対する F0 の軌跡．z_1 は高低差の大きい声 (我々の試聴によれば芝居掛かった声)，z_2 は高低差の少ない声 (落ち着いた声) を生成する．

5. 議論と結論

本研究では，既存の自己回帰生成モデルが抱える問題を克服し，自然さ，明瞭さに加えて表現の豊かさを持つ音声合成をラベルなしで実現するために，VAE を用いて多様な表現をモデル化することを提案した．実験では，ラベルなしでの表現豊かな音声合成という困難な逆問題において，VoiceLoop が VAE によって獲得した大域的特徴を利用することでより高品質な音声を生成すること，そして音声の話者性や話し方を操作できることを示した．

本研究の限界に，Truncated Back Propagation Through Time を含む十分なハイパーパラメータ探索を実験コストの問題から行えなかったことがある．しかし 4.2 節で述べた実験設定に限れば二つのデータセットで提案モデルの有効性を確かめることができた．また VAE が困難な逆問題を緩和し高品質な音声の生成に役立つことは示せたが，「表現の豊かさ」を直接評価することは課題として残された．

将来の研究の方向性としては，少量のラベルありデータを用いた半教師あり学習への拡張が考えられる．また，VAE のエンコーダーの構造を改良するなどして，より系列データに特化した潜在表現を獲得することも重要だと考える．

参考文献

[Braunschweiler 10] Braunschweiler, N., Gales, M. J. F., and Buchholz, S.: Lightly supervised recognition for automatic alignment of large coherent speech recordings, in *Proc. interspeech 2010*, pp. 2222–2225 (2010)

[Eyben 12] Eyben, F., Buchholz, S., and Braunschweiler, N.: Unsupervised clustering of emotion and voice styles for expressive TTS, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4009–4012 (2012)

[Gulrajani 17] Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A.: PixelVAE: A Latent Variable Model for Natural Images, in *Proc. 5th International Conference on Learning Representations* (2017)

[Hsu 17] Hsu, W.-N., Zhang, Y., and Glass, J.: Learning Latent Representations for Speech Generation and Transformation, in *Proc. Interspeech 2017*, pp. 1273–1277 (2017)

[King 12] King, S. and Karaiskos, V.: The Blizzard Challenge 2012, in *Proc. Blizzard Challenge workshop* (2012)

[Kingma 14] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes., in *Proc. 2nd International Conference on Learning Representations* (2014)

[Luong 17] Luong, H. T., Takaki, S., Henter, G. E., and Yamagishi, J.: Adapting and controlling DNN-based speech synthesis using input codes, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4905–4909 (2017)

[Oord 16] Oord, van den A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *CoRR*, Vol. abs/1609.03499, (2016)

[Ribeiro 11] Ribeiro, F., Florncio, D., Zhang, C., and Seltzer, M.: CROWDMOS: An approach for crowdsourcing mean opinion score studies, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2416–2419 (2011)

[Taigman 18] Taigman, Y., Wolf, L., Polyak, A., and Nachmani, E.: VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop, in *Proc. 6th International Conference on Learning Representations* (2018)

[Veaux 17] Veaux, C., Yamagishi, J., and MacDonald, K.: CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (2017)

[Wang 17a] Wang, Y., Skerry-Ryan, R. J., Xiao, Y., Stanton, D., Shor, J., Battenberg, E., Clark, R., and Saurous, R. A.: Uncovering Latent Style Factors for Expressive Speech Synthesis, *CoRR*, Vol. abs/1711.00520, (2017)

[Wang 17b] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, in *Proc. Interspeech 2017*, pp. 4006–4010 (2017)