潜在表現からの部分的推論による音声対話システム向け低遅延テキスト音声合成*

☆白井成彦, 森大毅 (宇都宮大学)

1 はじめに

リアルタイム性を重視する音声対話システムにおいて応答速度は極めて重要な要素であり、音声合成プロセスの高速化はその要である。高品質な音声合成モデルである VITS [1] は、推論時間の 96%以上を HiFi-GAN ベースのデコーダが占めており、これがリアルタイム動作への障壁となっている。本研究は、VITS のデコーダに入力される潜在表現の部分的デコードにより、音声対話システムにおける音声合成指令から合成音声再生開始までのレイテンシを短縮することを目的とする。

2 部分的デコードと結合手法

2.1 潜在表現からの部分的デコード

合成音声再生開始までのレイテンシを低減する手法として、入力を文節などに区切って音声合成することが考えられる。しかし、音声の韻律は文節間の統語構造や文の長さなど個々の文節を超えた範囲の単語の影響を受けるため、各文節を独立に合成すると自然性は大きく損なわれてしまう。

そこで、図 1 のように、VITS の Text Encoder が推定したパラメータに基づいてサンプリングされた潜在表現 z の列をセグメントごとに分割し、それぞれを独立に Decoder へ入力して音声を生成する「部分的デコード」を行う。Decoder が波形またはスペクトログラムを生成し終わるまでの時間は入力の潜在表現列長におおよそ比例するため、最初のセグメントに対応する合成音声の再生を早く開始することができる。

提案手法のポイントは、潜在表現列を得る部分までは VITS と全く同じという点にある。これにより、文節ごとに独立に合成するのと異なり、統語構造や文の長さなどを反映した本来の VITS の韻律生成能力をそのまま活かすことができる。

2.2 スペクトログラムのクロスフェードによるセグメント結合

本研究では、波形を直接生成するオリジナルの VITS に比べ出力の点数が少ないため高速な MS-iSTFT-VITS [2] を用いる。この場合、Decoder の出力は複素スペクトログラムとなる。部分的デコードの結果得られた複素スペクトログラムを結合し、全体を逆短時間フーリエ変換すれば合成音声が得られる。しかし、デコードされたセグメント同士を接続すると、位相の不連続性によるクリックノイズが生じるため合成音声の品質が低下する。

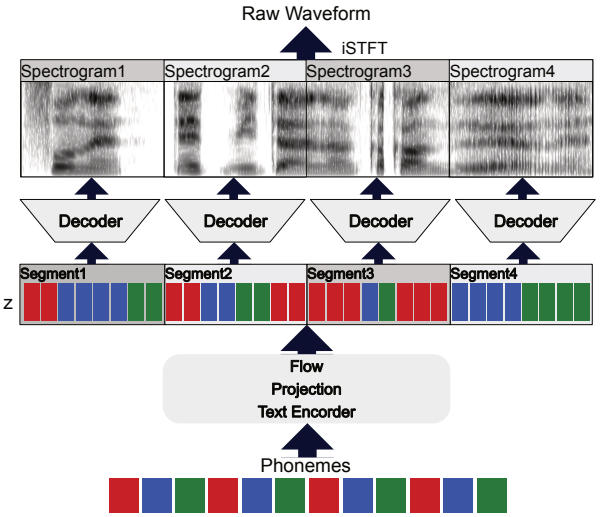

Fig. 1 Inference flow applying partial decoding from latent representations

そこで、隣接するセグメントの振幅スペクトル同士の 1 次元たたみこみにより類似性が最大となる重ね位置を探索し、クロスフェードで滑らかに結合する。

3 聴覚実験

セグメント結合手法が合成音声の自然性に与える影響を検証するため、以下の 4 つの条件で対話音声を合成し、聴覚実験を行った。

条件 1: Text-level Split テキストを入力時点で文節ごとに分割し、それぞれ独立して音声合成を行い、生成された波形同士をそのまま結合する。

条件 2: Naive Concatenation 部分的デコードの結果得られたスペクトログラムを単純に結合する。

条件 3: Adjusted Concatenation 隣接するセグメントのスペクトログラムの重ね位置を調整し

*Low-latency text-to-speech synthesis for spoken dialogue systems using partial inference from latent representations. by SHIRAI, Naruhiko, MORI, Hiroki , (Utsunomiya University)

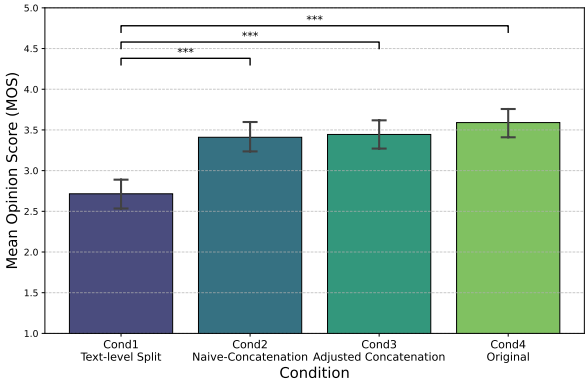

Fig. 2　Naturalness evaluation results (MOS) under each condition

クロスフェードにより滑らかに結合する。

条件 4: Original 分割処理を一切行わず、テキスト全体を一括で音声合成を行う。 VITS 本来の品質（Topline）として使用する。

日本語話し言葉コーパス（CSJ）[3] で事前学習し、宇都宮大学パラ言語情報研究向け音声対話データベース（UUDB）[4] でファインチューニングした MS-iSTFT-VITS [2] により生成した 16 文（音素数平均 28.0、標準偏差 20.0）× 4 条件の計 64 刺激を評価対象とした。音声の研究室に所属する 9 名を被験者として 5 段階評定尺度（MOS）による主観評価を実施した。

サンプリング周波数は 16 kHz とし、STFT 及び iSTFT のパラメータは、FFT サイズ 1024、ホップ長 256、ウィンドウサイズ 1024、窓関数には Hann 窓を用いた。音声合成には Intel 社製 i7-14700KF、5.500GHz を用いた。

また各条件において音声が出力可能になるまでに発生するレイテンシを計測した。

4　実験結果

聴覚実験の結果を図 2 に示す。エラーバーは 95% 信頼区間を表し、Tukey HSD による多重比較の結果もあわせて示す。

条件 1（2.72）が最も低いスコアを示した。条件 3（3.44）は、条件 2（3.41）よりも平均スコアが高くなった。条件 4（3.59）が最も高いスコアを示した。

Tukey HSD 法による多重比較の結果、条件 1 と条件 2、条件 1 と条件 3、条件 1 と条件 4 の間で平均値の差が有意 ($p < 0.001$) であり、それ以外の組では有意ではなかった。

以上の結果から、部分的デコードを行う手法（条件 2, 3）の導入による合成音声の自然性への悪影響は認められなかった。また、文節ごとに独立して音声合成を行う手法（条件 1）は自然性の低下が顕著であることがわかった。

Table 1　Comparison of latency across different conditions.

Condition	**Latency [ms]**
Cond 1	35.68 ± 9.64
Cond 2	50.07 ± 18.82
Cond 3	49.59 ± 19.36
Cond 4	90.71 ± 51.65

各条件における音声が再生可能になるまでのレイテンシの平均と標準偏差を表 1 に示す。

提案法である部分的デコード（条件 2, 3）はオリジナルの MS-iSTFT-VITS（条件 4）と比較してレイテンシが減少していることがわかる。これは部分的デコードにより、音声再生開始に必要な部分のみがデコードされたことでレイテンシが低減されたと考えられる。

5　おわりに

本稿では、MS-iSTFT-VITS を用いたストリーミング音声合成において、部分的デコードによるテキスト音声合成の低遅延化手法を提案した。また、セグメント同士の接続の際に生じる位相の不連続性を解消するため、振幅スペクトル同士を類似性が最大となるよう探索し、クロスフェードで滑らかに結合した。

聴取実験の結果、提案手法は従来のテキスト全体の一括生成に迫る品質を保ちながらレイテンシを低下させることができた。

一方で、提案手法では入力されたテキスト全体で潜在表現を生成する必要がある。よって音声対話システムにおいて発話文を LLM によって生成する場合、発話文が最後まで生成されないと提案手法では音声合成を開始できない。入力の逐次化は今後の課題と言える。

参考文献

[1] Kim *et al.*, Proc. ICML, Vol. 139, 5530–5540, 2021.
[2] Kawamura *et al.*, Lightweight and High-Fidelity End-to-End Text-to-Speech with Multi-Band Generation and Inverse Short-Time Fourier Transform, Proc. ICASSP 2023, 2023.
[3] Maekawa, Proc. SSPR, Vol. 139, 7–12, 2003.
[4] Mori *et al.*, Speech Communication, Vol. 53, 36–50, 2011.