離散音声トークン生成によるテキスト音声合成のための音声主観評価値予測に基づくdecoding戦略

山内 一輝[1,a)]　中田 亘[1]　齋藤 佑樹[1]　猿渡 洋[1]

概要：本論文では，離散音声トークン生成に基づくテキスト音声合成モデルおける decoding 戦略について探求する．我々は，テキスト生成における嗜好データに基づく decoding 戦略に着想を得た，離散音声トークン生成に向けた主観評価値予測に基づく新たな decoding 戦略を提案する．主観評価実験により，提案手法が合成音声の degeneration 問題を回避し，自然性を向上させるのに有効であることを示す．提案手法はモジュール性および拡張性が高く，音声合成モデルをアライメントするための有望なアプローチである．

Kazuki Yamauchi[1,a)]　Wataru Nakata[1]　Yuki Saito[1]　Hiroshi Saruwatari[1]

1. はじめに

テキスト音声合成 (Text-to-Speech: TTS) [1] とは，任意のテキストから対応する自然な読み上げ音声を合成する技術であり，音声 AI アシスタントなどに用いられている．音声は人間にとって最も重要なコミュニケーション手段の一つであり，テキストとは異なり発話者の感情や些細なニュアンスなども伝達できる．それゆえ，TTS は人間とコンピュータの自然なコミュニケーションを実現するために重要な技術である．

TTS モデルは音声の中間特徴量を自己回帰的に予測するモデルと，非自己回帰的に予測するモデルに分類される．自己回帰的な TTS モデルは非自己回帰的な TTS モデルに対して，音声の継続長などの韻律の多様性を向上させ，合成音声の表現力を向上させる傾向にある．従来の TTS モデルではメルスペクトログラムが中間特徴量として用いられることが多いが，連続的なメルスペクトログラムを自己回帰的に予測する場合，推論時に予測誤差蓄積の問題が起き，特に長い音声の合成時に合成音声の品質が劣化する問題が起きる．

近年，メルスペクトログラムを中間特徴量として用いる従来の TTS モデルと異なり，ニューラルオーディオコーデック (Neural Audio Codec: NAC) [2], [3] モデルによって抽出された音声の離散的な表現を中間特徴量として利用し，テキストを条件に NAC トークン列を生成することで音声を合成する TTS モデルに関する研究が活発に行われている [4], [5]．NAC は，入力された音声やオーディオ信号を離散的なトークン列に変換し，様々なダウンストリームタスクに利用できるようにする技術である．離散音声トークンを利用することで，自然言語処理分野の特にテキスト生成分野において研究されてきた離散トークン処理手法に類似した音声処理が可能となる [5], [6]．加えて，TTS において NAC のような離散音声トークンを音声の中間表現として用いる利点は，音声の自己回帰的な生成の際に誤差蓄積の問題を軽減することができるという点がある．離散的な音声トークンを自己回帰的に予測する場合，有限の候補から次の音声トークンを選択することになるため，メルスペクトログラムのような連続的なな特徴に比べて，予測誤差蓄積の問題が軽減されることが期待される．このように離散音声トークンを中間表現として用いる自己回帰的な TTS モデルは有望な手法であるが，テキスト生成分野に比べて離散トークンの decoding 戦略について十分探究されていない．

本論文では，離散音声トークン生成に基づく TTS モデルおける decoding 戦略について探求する．我々は，テキスト生成における嗜好データに基づく decoding 戦略に着想を得た，離散音声トークン生成に向けた主観評価値予測に基づく新たな decoding 戦略を提案する．我々はまず主観評価実験により，テキスト生成分野において提案された，離散トークンの自己回帰的な予測の際に起きる繰り返し生成問題を軽減するためのサンプリングに基づく decoding

[1] 東京大学, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] yamauchi-kazuki042@g.ecc.u-tokyo.ac.jp

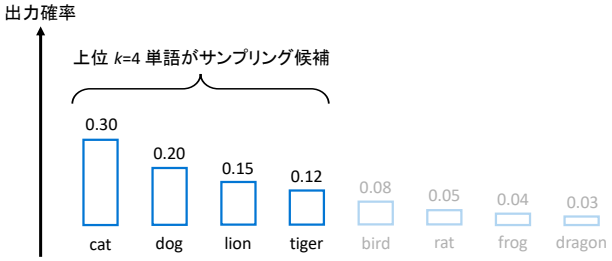

図 1: $k = 4$ とした top-k サンプリングの概念図.

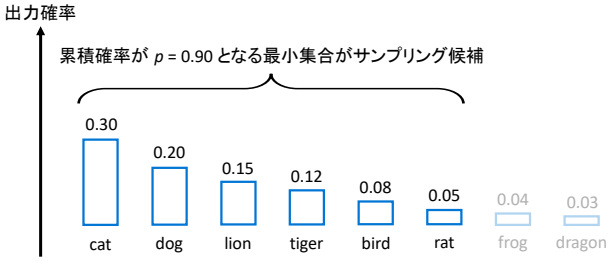

図 2: $p = 0.90$ とした top-p サンプリングの概念図.

戦略が，音声合成に対しても有効であることを示す．また，我々の提案手法が従来のサンプリングに基づく decoding 戦略を上回り，合成音声の自然性の向上に有効であることを示す．

2. 関連研究

2.1 言語モデルとサンプリング戦略

テキスト生成の分野において，言語モデルを用いて自己回帰的に単語を生成するテキスト生成手法が重要視されている．このような手法では，言語モデルは既に生成された単語列を条件に，次に生成され得る単語の尤度を出力する．言語モデルの出力に基づき次の単語を選択 (decoding) する戦略として最も単純な方法は，尤度が最も高い単語を選択するという手法 (greedy search) である．しかし，greedy search により単語を decoding すると，同じ単語が繰り返し生成されてしまうという問題がしばしば生じる．すなわち，過去に同じ単語の繰り返し列が生成されている場合，次の時刻においてもそれと同じ単語の尤度が高くなることで，同じ単語の繰り返しが生成されてしまう．

2.2 top-k/top-p サンプリング

自己回帰的に単語（離散トークン）を予測する際に起こる繰り返し問題を軽減するための decoding 戦略として，確率的なサンプリングに基づく top-k サンプリング [7] や top-p サンプリング [8] が提案されている．これらの手法はテキストを decoding する際にサンプリングを導入しており，出力トークンを多様化し，同じトークン列の繰り返し問題を軽減している．さらに，これらの手法では出力テキストの多様性を保持しながら，サンプリングされるトークンの候補を絞ることで不適切なトークンの生成を防ぐ工夫をしている．具体的には，top-k サンプリングでは，言語モデルの出力である各トークンの尤度に基づいて，サンプリングの候補を尤度の高い上位 k トークンに絞る（図 1）．また，top-p サンプリングでは，各トークンの尤度から計算された出力確率に基づき，累積確率が p を範囲で最小の集合を作成し，それらをサンプリングの候補とする（図 2）．このような工夫により top-k/top-p サンプリングは，サンプリングの候補の限定を行わない単純なサンプリング手法に比べて不適切なトークンが生成される問題を軽減しているが，あくまで生成結果は確率的であり，生成結果の適切さは保証されないため，完全な解決には至っていない．

2.3 Controlled decoding

言語モデルの出力テキストがより人間にとって好ましいものとなるように，言語モデルの出力を制御可能な decoding 戦略である controlled decoding [9] という手法が提案された．この手法では，まず，途中まで decoding されたテキストに対し，その人間にとって好ましさを予測するモデルである prefix scorer を人間の嗜好データに基づいて学習する．推論時は，まず M 単語からなるブロック（フレーズ）を K パターンサンプリングし，そのうち prefix scorer により予測された人間にとっての好ましさのスコア (prefix score) が最も高いものを選択する．さらに，その続きとなる M 単語を K パターンサンプリングし，最も prefix score の高いものを選択するという手順を，テキストの生成が終了するまで繰り返す．テキストの生成にサンプリングを用いる従来の手法は，生成結果の多様化させる一方で，サンプリングによって望ましくない結果が生成されてしまうという問題が発生する．一方で，controlled decoding は複数パターンのサンプリング結果から最も望ましいものを選択するため，そのような問題を軽減し出力結果の人間にとっての好ましさを安定化させることができる．また，同じ単語の繰り返し列をはじめとして，尤度が高いフレーズと人間にとって好ましいフレーズは異なる場合があるため，controlled decoding は言語モデルの出力結果を望ましい方向に制御するための有望な手法である．

3. 提案手法

我々は，controlled decoding に着想を得た，自己回帰的に離散音声トークンを生成する際の主観評価値予測に基づく新たな decoding 戦略を提案する．提案手法はまず，途中まで合成された音声に対する自然性に関する主観評価値を予測する prefix scorer を学習し，それを用いて音声を数秒単位で逐次的に decoding する．提案手法の概要を図 3 に示す．

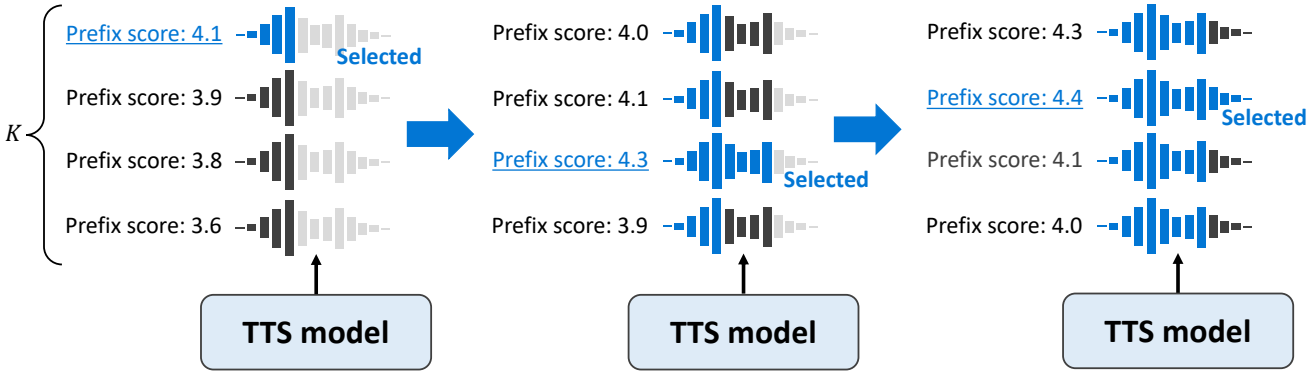

図 3: 提案する音声主観値予測に基づく controlled decoding の概要．M 個の離散音声トークンからなるブロック単位で逐次的に音声を decoding する．

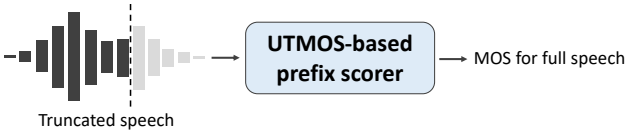

図 4: UTMOS-based prefix scorer の学習方法の概要．

3.1 UTMOS-based prefix scorer

音声の自然性に関する主観評価値を予測するモデルとして UTMOS [10] がある．我々は，UTMOS をベースとして，途中まで合成された音声に対する自然性に関する主観評価値を予測する prefix scorer を構築する．prefix scorer は途中まで decoding された音声を入力として受け取り，それが完全に decoding されたときの自然性に関する主観評価値を予測する必要がある．そこで，我々は UTMOS の学習時に，入力音声をランダムな時刻以降切り捨てるという操作を行うことで，UTMOS ベースの prefix scorer を学習した（図 4）．なお，ここで選択する時刻は 0.5 秒単位とした．

3.2 音声主観値予測に基づく controlled decoding

提案法では，TTS の推論時に，テキスト生成における controlled decoding と同様にブロック単位で離散音声トークンを decoding する．具体的には，まず M 個の離散音声トークンからなるブロックを K パターンサンプリングする．次に，K パターンのブロックをそれぞれ DAC モデルのデコーダに入力し音声波形を生成し，UTMOS ベースの prefix scorer に入力して主観評価予測値 (prefix score) を得る．そして，最も prefix score の高いブロックを選択し，それを条件としてその続きとなるブロックを decoding していくという手順を，離散音声トークンの生成が終了するまで繰り返す．

4. 実験的評価

4.1 実験条件

TTS モデルの設定: 本研究で使用した TTS モデルの概

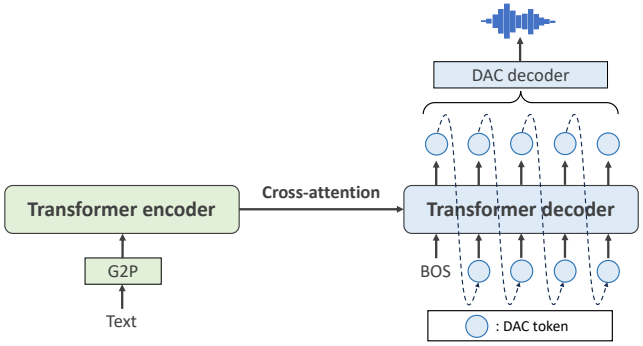

図 5: 使用した TTS モデルの概要．

要を図 5 に示す．このモデルは，NAC トークンを音声の中間表現として用いる音声処理システムである UTDUSS [11] に含まれる TTS モデルに基づいている．我々は離散音声トークン抽出のために Descript Audio Codec(DAC) [3] を学習した．テキストから DAC トークン列を生成するための音響モデルとしては，Transformer TTS [12] に触発され，入力テキストから DAC トークンを自己回帰的に予測する Transformer エンコーダデコーダモデル [13] を学習した．このモデルは，エンコーダ入力として音素列を取り，デコーダがクロスアテンションメカニズムを通じて DAC トークン列を自己回帰的に生成する．また，DAC トークンを波形に変換するためのボコーダには離散単位抽出に使用した DAC モデルのデコーダをそのまま使用した．

DAC モデルの設定について，コードブックサイズ（すなわちコードブック内のコードベクトルの数）を 512 に，残差ベクトル量子化 (Residual Vector Quantization: RVQ) [14] のコードブック数を 1 に設定した．DAC エンコーダのダウンサンプリングレートは $2, 3, 4, 4, 5$ および $5, 4, 4, 3, 2$ に設定し，ダウンサンプリングレートを 480 とした．敵対的学習の multi-period discriminator については，周期を $2, 3, 5, 7, 11, 13, 17$ に設定し，FFT ウィンドウサイズを $2048, 1024, 512, 256$ に設定した．損失の重み付けについては，ハイパーパラメータ探索に基づいて RVQ のコミットメント損失の重みを 2.0，コードブック損失の重みを 8.0 に

表 1: サンプリング戦略に関する評価結果（95% 信頼区間付き MOS および UTMOS）．太字は最も MOS が高かったことを示す．

Method	MOS (↑)	UTMOS (↑)
greedy search	3.35 ± 0.09	4.27
naive sampling	3.57 ± 0.08	4.31
top-k top-p sampling	3.62 ± 0.08	4.36
sequence-level best-of-K	3.71 ± 0.07	4.46
block-wise best-of-K	**3.73 ± 0.07**	4.43
ground-truth	3.92 ± 0.07	4.43

設定し，メルスペクトログラム予測損失の重み (15.0) に近い値に調整した．DAC モデルはバッチサイズ 128，学習率 0.0001 で，8 つの NVIDIA A100 GPU 上で，230k イテレーション，約 30 時間で学習された．その他のハイパーパラメータについては、[*1]公式モデルの設定に従った．

音響モデルのネットワークアーキテクチャおよび訓練設定については，主に Transformer TTS の公式実装[*2]に従った．この音響モデルは Adam optimizer [15] を用いて，バッチサイズ 32，学習率 0.001 で，単一の NVIDIA A100 GPU 上で，20 万回のイテレーション，約 3 時間で学習された．

データセット: DAC モデルおよび音響モデルの学習には，LJSpeech [16] を使用し，学習用（12,600 発話），検証用（250 発話），評価用（250 発話）サブセットに分割して用いた．また，すべての音声は 16 kHz にリサンプリングして用いた．

比較手法: 本実験では，以下の 5 つの手法を比較した．

- **greedy search**: 最も尤度の高いトークンを選択していく decoding 戦略
- **naive sampling**: サンプリングの候補の限定を行わない単純なサンプリングに基づく decoding 戦略
- **top-k top-p sampling**: top-k によりサンプリング候補を絞った後に，top-p によりさらにサンプリング候補を絞るサンプリング手法に基づく decoding 戦略
- **sequence-level best-of-K**: top-k top-p sampling を用いて K パターンの音声をサンプリングし，最も UTMOS の高い音声を選択する decoding 戦略
- **block-wise best-of-K**: UTMOS-based prefix scorer を用いた音声主観値予測に基づく controlled decoding

top-k top-p sampling における k および p の値はそれぞれ $190, 0.50$ とした．また，sequence-level best-of-K および block-wise best-of-K におけるサンプル数 K の値は 8 とし，block-wise best-of-K におけるブロックサイズ M の値は 16 とした．なお，16 個の離散音声トークン列は約 0.5 秒の音声に対応する．block-wise best-of-K において離散音声トークンをサンプリング際には top-k top-p sampling を用いた．

4.2 評価基準

クラウドソーシングを用いて，音声の自然性に関する MOS テストを実施した．各手法の合成音声をランダムに提示し，音声の自然性（人間らしく自然な音声に聴こえるか）を 5 段階で評価させた．MOS 評価の受聴者はアメリカまたはイギリス在住のネイティブ英語話者に限定し，受聴者数は 200 人，1 人の評価回数は 24 とした．また，先述の 5 つの手法による合成音声の他に参考として LJSpeech コーパスの自然音声も評価させた．

また，MOS テストによる主観評価だけでなく，参考として UTMOS の計測も行った．具体的には，分割した LJSpeech の評価用セットに含まれる音声の書き起こし文 250 文から各手法により音声を合成し，それらに対する UTMOS の平均値を計算した．

4.3 評価結果

評価結果を表 1 に示す．この結果における有意差の有無は，両側 t 検定によって p 値が 0.05 以下かどうかにより評価した．まず，top-k top-p sampling および naive sampling によって合成した音声の自然性に関する MOS は，greedy search に対して有意に高いことが示された．この結果から，テキスト生成分野において提案された，離散トークンの自己回帰的な予測の際に起きる繰り返し生成問題を軽減するためのサンプリングに基づく decoding 戦略が，音声合成に対しても有効であることが示された．

また，提案手法である block-wise best-of-K によって合成した音声の自然性に関する MOS は，top-k top-p sampling など従来のサンプリングに基づく decoding 戦略に対して有意に高いことが示された．この結果は，我々の提案手法が従来のサンプリングに基づく decoding 戦略を上回り，合成音声の自然性の向上に有効であることを示唆する．

一方で，block-wise best-of-K と sequence-level best-of-K の間には有意差が存在しなかった．この結果は，複数パターンのサンプリング結果から最も望ましいものを選択す

[*1] https://github.com/descriptinc/descript-audio-codec/blob/main/conf/final/24khz.yml
[*2] https://github.com/soobinseo/Transformer-TTS

表 2: サンプル数 K に関する ablation study．実験により得られた 95% 信頼区間付き MOS および UTMOS．太字は最も MOS が高かったことを示す．

サンプル数 K	MOS (↑)	UTMOS (↑)
2	3.72 ± 0.08	4.40
4	3.74 ± 0.08	4.43
8	**3.83 ± 0.07**	4.43
16	3.79 ± 0.07	4.45
32	3.65 ± 0.08	4.46

表 3: ブロックサイズ M に関する ablation study．実験により得られた 95% 信頼区間付き MOS および UTMOS．太字は最も MOS が高かったことを示す．

ブロックサイズ M	MOS (↑)	UTMOS (↑)
16	**3.79 ± 0.07**	4.43
35	3.75 ± 0.07	4.43
70	3.78 ± 0.07	4.44
140	3.77 ± 0.07	4.43

るという decoding 戦略が，不適切なトークンが生成されてしまう問題を軽減し出力音声の自然性を安定させるのに有効である一方で，ブロック単位の逐次的な decoding が必ずしも合成音声の自然性向上に有効とは言えないことを示唆する．

また，計測した UTMOS は MOS とある程度相関があったが，必ずしも傾向は一致しなかった．これは，合成音声に対する UTMOS などの主観評価値予測モデルによる評価結果を上げるように過剰適合することが，必ずしも主観評価値の向上に有効とはならないことを示唆する．

4.4 Ablation study

提案手法である block-wise best-of-K において，サンプル数 K ブロックサイズ M が与える影響について調査した．

まず，M を 16 とし，K の値を $2, 4, 8, 16, 32$ と変化させたときの合成音声の評価結果を比較した．評価結果を表 2 に示す．この結果から，K を 8 とし合成した音声の自然性に関する MOS は，K を 2 や 32 とした場合に対して有意に高いことが示された．この結果は，合成音声の自然性を安定させるためにはある程度のサンプル数が必要である一方で，サンプル数を大きくしすぎると深層学習モデルによる主観評価値予測に過剰適合してしまい，かえって人間による主観評価値が低くなってしまうことを示唆する．実際，サンプル数 K を大きくするほど，UTMOS は高くなっていることがわかる．

また，K を 8 とし，M の値を $16, 32, 70, 140$ と変化させたときの合成音声の評価結果を比較した．評価結果を表 3 に示す．この結果から，ブロックサイズ M を変化させても主観評価値に有意差は生まれないことが示された．この結果は，block-wise best-of-K と sequence-level best-of-K の間には有意差が存在しなかったという結果とも整合している．block-wise best-of-K は長時間の音声のストリーミング方式による合成などの応用が可能であるため，その点では sequence-level best-of-K に対する優位性はあるが，合成音声の自然性向上に対して有効な逐次的 decoding 戦略を提案することは今後の課題である．

5. おわりに

本論文では，離散音声トークン生成に基づく TTS モデルおける decoding 戦略について探求した．我々は，テキスト生成で提案された controlled decoding に着想を得た，離散音声トークン生成に向けた主観評価値予測に基づく新たな decoding 戦略を提案した．我々はまず主観評価実験により，テキスト生成分野において提案された，離散トークンの自己回帰的な予測の際に起きる繰り返し生成問題を軽減するためのサンプリングに基づく decoding 戦略が，音声合成に対しても有効であることを示した．また，我々の提案する新たな decoding 戦略が，従来のサンプリングに基づく decoding 戦略を上回り，合成音声の自然性の向上に有効であることを示した．一方で，音声を逐次的に decoding することが，合成音声の自然性向上に対して有効であることは示されなかった．合成音声の自然性向上に対して有効な逐次的 decoding 戦略を提案することは今後の課題である．また，自然性以外の観点からの音声主観評価値予測に基づく controlled decoding の提案も今後の展望である．

謝辞： 本研究は, JST, Moonshot R&D 助成金番号 JPMJPS2011（実験的評価）と JST，ACT-X，JPMJAX23CB（アルゴリズム開発）の支援を受けたものである．

参考文献

[1] Y. Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis units,” in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[2] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” in *Proc. NIPS*, 2023.

[4] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, “Soundstorm: Efficient parallel audio generation,” *arXiv*, vol. abs/2305.09636, 2023.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, “AudioLM: A language modeling approach to audio generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2022.

[6] C. Wang, S. Chen, Y. Wu, Z.-H. Zhang, L. Zhou, S. Liu,

Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv*, vol. abs/2301.02111, 2023.

[7] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proc. ACL*, Melbourne, Australia, Jul. 2018, pp. 889–898.

[8] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proc. ICLR*, Virtual Conference, Apr. 2020.

[9] S. Mudgal, J. Lee, H. Ganapathy, Y. Li, T. Wang, Y. Huang, Z. Chen, H.-T. Cheng, M. Collins, T. Strohman, J. Chen, A. Beutel, and A. Beirami, "Controlled decoding from language models," in *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*, 2023.

[10] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. INTERSPEECH*, 2022, pp. 4521–4525.

[11] N. Wataru, Y. Kazuki, Y. Dong, H. Hiroaki, and S. Yuki, "UTDUSS: UTokyo-SaruLab System for Interspeech2024 Speech Processing Using Discrete Speech Unit Challenge," *arXiv preprint arXiv:2403.13720*, 2024.

[12] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI*, Hawaii, U.S.A., Jul. 2019, pp. 6706–6713.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[14] A. Vasuki and P. Vanathi, "A review of vector quantization techniques," *IEEE Potentials*, vol. 25, no. 4, pp. 39–47, Jul. 2006.

[15] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[16] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.