離散音声トークン生成に基づく感情合成音声のための多目的知覚評価値を活用したdecoding戦略

山内 一輝[1]　中田 亘[1]　齋藤 佑樹[1]　猿渡 洋[1]

概要：本稿では，離散音声トークン生成に基づく音声合成のための decoding 戦略について検討する．我々の先行研究では，音声の知覚評価値予測（perceptual rating prediction: PRP）に基づく best-of-K (BOK) を導入したサンプリングベースの decoding 戦略である BOK-PRP が提案され，知覚評価値として音声の自然性に関する主観評価値を用いた BOK-PRP が合成音声の自然性向上に寄与することが示された．本研究では，感情音声合成に焦点を当て，知覚評価値として音声の自然性に加え，感情強度に関する主観評価値を導入する．合成音声の自然性と感情強度の間にトレードオフが存在することに着目し，それらの線型和に基づく多目的知覚評価値を新たに導入する．実験的評価により，多目的知覚評価値を活用した BOK-PRP が，合成音声の自然性と感情強度のトレードオフを効果的に制御および改善することを示す．

キーワード：テキスト音声合成，感情音声合成，言語モデル，decoding 戦略

1. はじめに

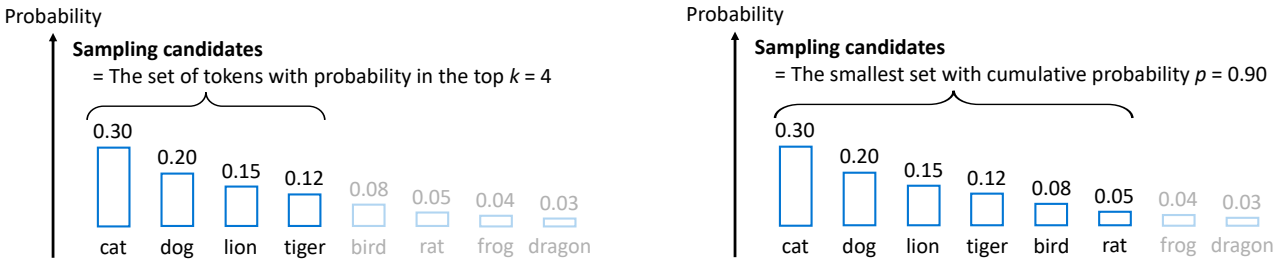
近年，言語モデル (language model: LM) に基づくテキスト音声合成 (text-to-speech: TTS) 手法が大きな注目を集めている [1], [2], [3], [4], [5]．これらのモデルは通常，音声をニューラルオーディオコーデック [6], [7], [8] などの離散トークンとして表現し，これらのトークンを言語モデルを用いて自己回帰的に生成することで音声を合成する．言語モデルに基づく TTS の進歩を支える主要な要因の一つとして，テキスト生成分野で開発された技術を音声合成に応用した点が挙げられる [9], [10], [11]．例えば，言語モデルによるテキスト生成のために開発されたサンプリングベースの decoding 戦略は，言語モデルに基づく TTS においても，greedy decoding が引き起こす繰り返し生成問題の解決に有効であることが確認されている [12], [13]．さらに，サンプリングベースの decoding を言語モデルに基づく TTS に導入することで，韻律の多様性が向上し，合成音声の表現力を高めることができる．特に感情音声合成では，サンプリングによって得られる韻律の多様性が，感情表現の幅を広げ，感情の表現力を向上させる上で重要な役割を果たしている．

しかしながら，サンプリングに伴うランダム性により，アーティファクトなどの望ましくない出力が生成されるなど，合成音声の自然性が不安定化する問題が生じる．これを軽減するために，top-k サンプリング [14] や top-p サンプリング [15] は，サンプリングの候補となるトークンを，言語モデルにより算出された生起確率の高いトークンに限定する．しかしながら，候補を絞ることで出力の多様性が減少し，感情の表現力が低下してしまうため，根本的な解決策には至っていない．

この問題を解決するため，我々の先行研究では，音声の知覚評価値予測（perceptual rating prediction: PRP）に基づく best-of-K（BOK）を導入したサンプリングベースの decoding 戦略である *BOK-PRP* を提案した [16]．先行研究 [16] では音声の知覚評価値として，自然性に関する平均オピニオンスコア（mean opinion score: MOS）に注目し，BOK-PRP により合成音声の主観的自然性が向上することを示した．ただし，先行研究 [16] は，合成音声の韻律表現の多様性が限定的である単純な読み上げ音声合成に焦点を当てており，知覚評価の観点として自然性のみに注目していた．

そこで本稿では，合成音声の韻律表現がより多様となる感情音声合成に焦点を当て，BOK-PRP に自然性に加えて感情の表現力という観点からの知覚評価値を導入する．さらに，我々は言語モデルに基づく感情音声合成モデルによる合成音声には，自然性と感情の表現力の間にトレードオフが存在することに着目した．そこで，自然性と感情強度の知覚評価値の線形結合に基づく多目的知覚評価を新たに

[1] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.



(a) Top-k サンプリングの概要 ($k = 4$).　(b) Top-p サンプリングの概要 ($p = 0.90$).

図 1: 従来のサンプリングベースの decoding 戦略.

提案する．線形結合の重み係数を手動で調整することで，推論時に自然性と感情強度のどちらをより重視するかを制御することが可能となる．実験的評価では，多目的知覚評価に基づく BOK-PRP が，合成音声の自然性と感情強度のトレードオフを効果的に制御および改善するために有効であるかを評価するための主観評価を実施する．

2. 関連研究

2.1 従来の decoding 戦略

Decoding とは，言語モデルによって算出されたトークンの確率分布に基づいて出力トークンを選択するプロセスである．Greedy decoding は，次のトークンとして，最も生起確率の高いトークンを選択する，最も単純な決定論的 decoding 戦略である．しかし，greedy decoding は出力が同じトークンを繰り返すループに陥ってしまう問題を引き起こす可能性がある．

これに対し，top-k サンプリング [14] や top-p サンプリング [15] などのサンプリングベースの decoding 戦略は，トークンの確率分布に基づいて，確率的にトークンを選択する．これらの戦略は多様性をもたらし，繰り返し生成の問題を効果的に解決する．さらに，top-k および top-p サンプリングではサンプリング候補のトークンを絞り込み，不適切な出力を抑制する．具体的には，top-k サンプリングでは，確率が 1 位から k 位までのトークンの集合からトークンを選択する (図 1a)．一方，top-p サンプリングでは，累積確率が閾値 p を超える最小のトークン集合からトークンを選択する (図 1b)．しかし，候補を絞り込むことで出力の多様性が減少し，サンプリングベースの戦略がもつ表現力向上の利点が十分に発揮されなくなってしまう．そのため，不適切な出力をフィルタリングしながら出力の多様性を維持することは依然として課題である．

2.2 BOK-PRP: 知覚評価値予測に基づく best-of-K

BOK-PRP という言語モデルに基づく TTS のための，効果的なサンプリングベースの decoding 戦略が提案されている [16]．この戦略は，音声の知覚評価値予測に基づく best-of-K を導入している．具体的には，BOK-PRP は，言語モデルに基づく TTS モデルにより K 通りの合成音声をサンプリングし，外部の知覚評価値予測器による評価が最も高いサンプルを出力として選択する．従来の戦略が言語モデルによって算出された確率分布に基づいてトークンを選択するのに対し，BOK-PRP は知覚評価値予測を活用することで，望ましくない出力を除外しつつ，出力の多様性を維持することが可能となる．

数学的定式化　以下では，$\boldsymbol{x}$ を入力テキスト，p を事前学習された言語モデル（自己回帰的に離散音声トークンの確率分布を計算するモデル）とする．K 個のサンプル集合 $Y_K = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$ を $p(\cdot|\boldsymbol{x})$ から生成する．ここで，各サンプル $\boldsymbol{y}_k = [y_{k,1}, \ldots, y_{k,T}] (k = 1, \ldots, K)$ は長さ T の離散音声トークン列を表す．また，$\boldsymbol{w}_k = \mathrm{Dec}(\boldsymbol{y}_k)$ をデコーダ $\mathrm{Dec}(\cdot)$ によって合成された音声波形とする．BOK-PRP は，上記の定式化に適合する任意の言語モデルに基づく TTS モデルに適用可能である．

BOK-PRP のアルゴリズム　まず，$p(\cdot|\boldsymbol{x})$ から K 通りの音声サンプル Y_K を生成する．次に，最も予測知覚評価値の高い最適なトークン列 $\hat{\boldsymbol{y}}$ を以下のように選択する：

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}_k \in Y_K}{\mathrm{argmax}}\, r\left(\mathrm{Dec}\left(\boldsymbol{y}_k\right)\right). \tag{1}$$

ここで，$r(\cdot)$ は知覚評価値予測器を表す．最後に，最適な出力音声 $\hat{\boldsymbol{w}} = \mathrm{Dec}\left(\hat{\boldsymbol{y}}\right)$ を合成する．

3. 提案手法

本研究では，best-of-K の基準とする知覚評価値予測のために，自然性 MOS 予測器 $r_{\mathrm{nat}}(\cdot)$ および感情強度 MOS 予測器 $r_{\mathrm{emo}}(\cdot)$ の 2 つの知覚評価値予測器を用いる．さらに，$r_{\mathrm{nat}}(\cdot)$ と $r_{\mathrm{emo}}(\cdot)$ を組み合わせることで，新たな多目的知覚評価値予測器 $r(\cdot)$ を構成する．

3.1 自然性 MOS 予測

自然性 MOS 予測器 $r_{\mathrm{nat}}(\cdot)$ として，広く用いられている UTMOS [17] を利用する．UTMOS は音声波形を入力として，自然性 MOS の予測値を出力する．この出力は 1

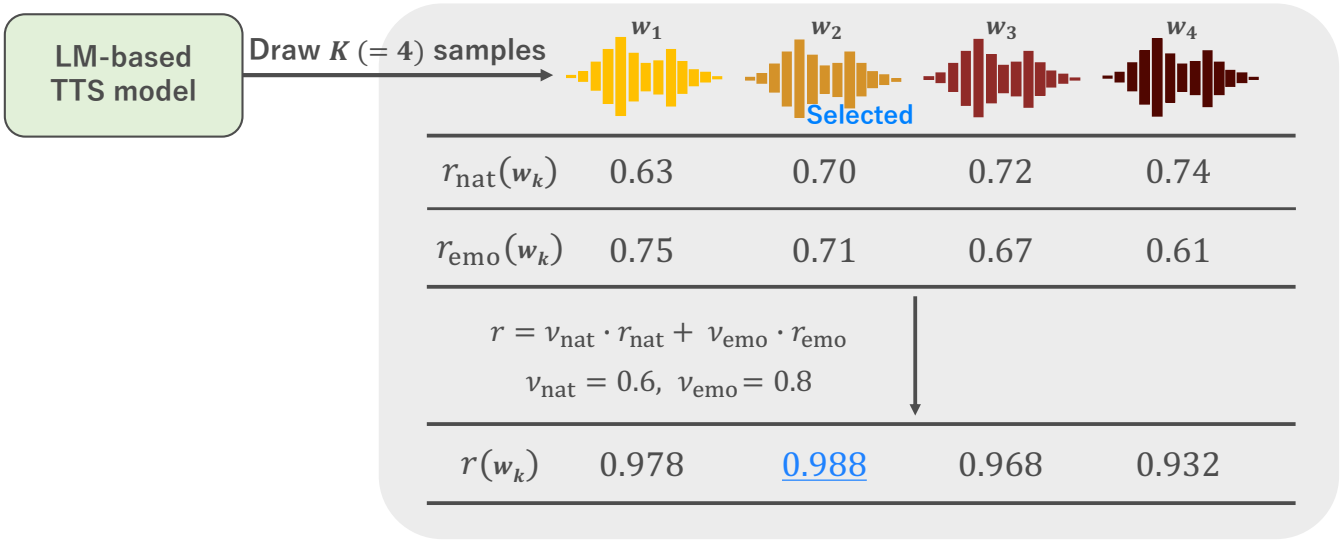


図 2: 提案する多目的知覚評価値予測に基づく BOK-PRP の decoding プロセスの概要．まず，K 通りの合成音声 $\boldsymbol{w}_k (k = 1, \ldots, K)$ がサンプリングされ，それぞれの自然性と感情強度の評価値 r_{nat}，r_{emo} が計算される．次に，それらと重み係数 ν_{nat}，ν_{emo} に基づいて最終的な評価値 r が算出され，最も r の高いサンプルが最終的な出力として選択される．

（非常に不自然）から 5（非常に自然）までの連続実数値である．本研究では，出力値を 0 から 1 までの実数値にスケールするために，$r_{\text{nat}}(\cdot)$ を以下のように定義する：

$$r_{\text{nat}}(\boldsymbol{w}_k) = \frac{\text{UTMOS}(\boldsymbol{w}_k) - 1}{4}. \tag{2}$$

$r_{\text{nat}}(\cdot)$ の値が高いほど，予測される自然性が高いことを意味する．

3.2 感情強度 MOS 予測

感情強度 MOS 予測器 $r_{\text{emo}}(\cdot)$ として，事前学習済みの次元感情認識モデル [18][*1]を利用する．このモデルは，音声から覚醒度（arousal），支配性（dominance），および感情価（valence）を 0（低い）から 1（高い）の範囲で予測する．本研究では，怒り（angry），喜び（happy），悲しみ（sad）の 3 つの感情の音声合成に焦点を当て，$r_{\text{emo}}(\cdot)$ を以下のように定義する：

$$r_{\text{emo}}(\boldsymbol{w}_k) = \begin{cases} \dfrac{\text{arousal}(\boldsymbol{w}_k) + (1 - \text{valence}(\boldsymbol{w}_k))}{2} \\ \qquad \text{if target emotion is “angry”}, \\ \dfrac{\text{arousal}(\boldsymbol{w}_k) + \text{valence}(\boldsymbol{w}_k)}{2} \\ \qquad \text{if target emotion is “happy”}, \\ \dfrac{(1 - \text{arousal}(\boldsymbol{w}_k)) + (1 - \text{valence}(\boldsymbol{w}_k))}{2} \\ \qquad \text{if target emotion is “sad”}. \end{cases} \tag{3}$$

$r_{\text{emo}}(\cdot)$ の値が高いほど，ターゲット感情に対する予測感情強度が高いことを意味する．

[*1] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Multi-Attributes

3.3 多目的知覚評価

本研究では，言語モデルに基づく感情音声合成モデルによる合成音声には，自然性と感情強度の間にはトレードオフが存在することに注目する．つまり，合成音声の感情強度が高いほど，自然性が低下する傾向にある．そこで，以下のように定義される多目的知覚評価値予測器 $r(\cdot)$ を導入する：

$$r(\boldsymbol{w}_k) = \nu_{\text{nat}} \cdot r_{\text{nat}}(\boldsymbol{w}_k) + \nu_{\text{emo}} \cdot r_{\text{emo}}(\boldsymbol{w}_k). \tag{4}$$

ここで，重み係数 ν_{nat} および ν_{emo} は，$\nu_{\text{nat}}^2 + \nu_{\text{emo}}^2 = 1$ を満たす実数である．この設定は，テキスト生成の分野で提案された，ユーザー依存の preference 制御を可能とするための強化学習手法である Directional Preference Alignment with multi-objective rewards [19] に着想を得ている．ν_{nat} と ν_{emo} の値を手動で調整することで，best-of-K において，自然性と感情強度のどちらかを重視するかを制御することが可能となる．図 2 は，提案する多目的知覚評価値に基づく BOK-PRP の decoding プロセスを示している．

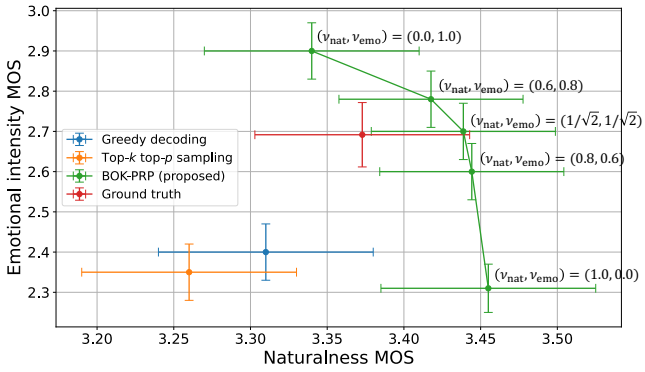
4. 実験的評価

提案する多目的知覚評価に基づく BOK-PRP が，合成音声の主観的な自然性と感情強度のトレードオフを効果的に改善するかを検証するために，主観評価を実施する．

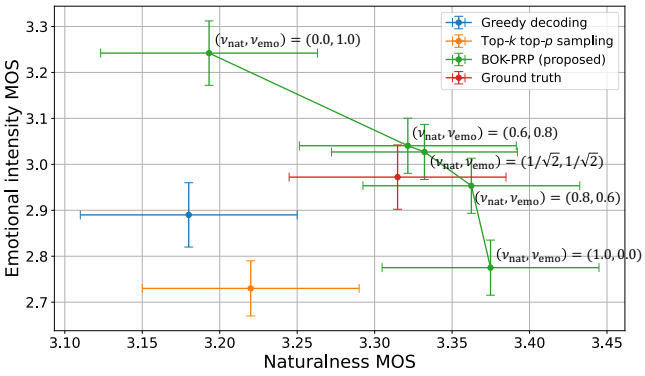
4.1 実験条件

言語モデルに基づく感情的 TTS モデル 言語モデルに基づく感情的 TTS モデルとして，事前学習済みの Cosyvoice-300M-Instruct モデル [2][*2]を使用する．このモデルは，感

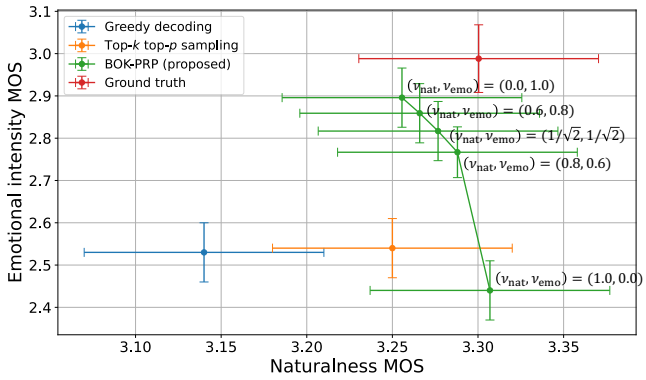
[*2] https://github.com/FunAudioLLM/CosyVoice



(a) ターゲット感情が “怒り” の場合.



(b) ターゲット感情が “喜び” の場合.



(c) ターゲット感情が “悲しみ” の場合.

図 3: 主観評価実験の結果．グラフの横軸は自然性 MOS，縦軸は感情強度 MOS を表す．上下に伸びるバーは 95% 信頼区間を示す．Ground truth は，ESD に含まれる自然音声のサンプルを指す．

情や発話スタイルを指定することで表現豊かな合成音声をサンプリングできる．なお，公式実装では decoding 戦略として top-k top-p サンプリングが使用されている．

データセット　評価には，Emotional Speech Dataset (ESD) [20] の英語サブセットを使用する．このデータセットには，10 人の話者が 5 つの感情（怒り，喜び，悲しみ，驚き，中立）を表現した音声が含まれており，各話者・感情あたり 350 発話（各話者約 1,750 発話，1.2 時間）が含まれている．公式実装 [20] が提供するテストセットには，各話者・感情あたり 30 発話（合計 1,500 発話）が含まれる．このテストセットに含まれるテキストから各手法により音声を合成し，自然音声のサンプルと比較して評価を行う．

比較手法　本実験では，以下の decoding 戦略を比較する：

- **Greedy decoding**：最も生起確率の高いトークンを選択する決定論的戦略
- **Top-k top-p サンプリング**：まず top-k に基づきサンプリング候補を絞り込み，次に top-p に基づきさらに候補を絞り込むサンプリングベースの戦略
- **BOK-PRP (proposed)**：提案する多目的評価値に基づく best-of-K を導入したサンプリングベースの戦略

Top-k top-p サンプリングでは，k を 500，p を 0.90，温度パラメータを 1.0 に設定する．BOK-PRP では，K の値を 10 に設定する．また，BOK-PRP には多目的知覚評価値予測器 r を利用し，重み係数の組み合わせとしては以下の 5 通りを比較する：

$$(\nu_{\text{nat}}, \nu_{\text{emo}}) = (1.0, 0.0), (0.8, 0.6), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \quad (5)$$

$$(0.6, 0.8), (0.0, 1.0). \quad (6)$$

評価指標　クラウドソーシングを通じて MOS テストを実施し，合成音声の主観的な自然性と感情強度を評価した．クラウドソーシングのプラットフォームとしては Prolific [*3] を利用した．参加者はランダムに選ばれた合成音声のサンプルの自然性を 1（非常に不自然）から 5（非常に自然），事前に指定された感情の強度を 1（非常に弱い）から 5（非常に強い）の 5 段階で評価した．評価には合計 250 人が参加し，各参加者は 24 サンプルを評価した．ただし，参加

[*3] https://www.prolific.com/

者は英語を母語とする話者に限定した．MOS の差の有意性は，有意水準を 5% としたスチューデントの t 検定を用いて評価した．

4.2 実験結果

図 3 に主観評価の結果を示す．まず，すべての感情において，BOK-PRP で感情強度を重視するよう重み係数を調整することで，感情強度 MOS を有意に向上させることが可能であることが確認された．特に，感情強度を重視した場合，greedy decoding や top-k top-p サンプリングといった従来の decoding 戦略と比較して，BOK-PRP により感情強度 MOS が有意に向上している．一方，感情強度を重視するように重みを調整すると，自然性 MOS が低下する傾向がある．特に，図 3b に示されるように，$(\nu_{\text{nat}}, \nu_{\text{emo}}) = (0.0, 1.0)$ の場合，$(\nu_{\text{nat}}, \nu_{\text{emo}}) = (1.0, 0.0)$ の場合と比較して自然性 MOS が有意に低い．逆に，自然性を重視する場合，greedy decoding や top-k top-p サンプリングと比較して自然性 MOS が有意に向上している．これらの結果は，多目的知覚評価に基づく BOK-PRP が以下を実現することを示している：

(1) 合成音声の主観的な自然性と感情強度のトレードオフを，重み係数を調整することで制御可能である．

(2) 従来の decoding 戦略と比較して，トレードオフを有意に改善することができる．

5. おわりに

本研究では, 言語モデルに基づく TTS のための decoding 戦略である BOK-PRP を感情音声合成に適応した．我々は，言語モデルに基づく感情音声合成モデルによる合成音声の自然性と感情強度の間にトレードオフが存在することに着目し，これらの予測値の線形結合に基づく多目的知覚評価値を新たに導入した．主観評価実験により，多目的知覚評価値に基づく BOK-PRP が従来のサンプリングベースの戦略を上回り，合成音声の自然性と感情強度のトレードオフを効果的に制御および改善可能であることが示された．

また，今後の研究において，言語モデルに基づく自発音声合成 [21] および言語モデルに基づく音声対話生成 [22] に対する BOK-PRP の有効性を検証する予定である．さらに，本稿では主に合成音声の自然性および感情強度に関する MOS の向上に焦点を当てたが，BOK-PRP は韻律的自然性や対話の流暢さといった他の観点からの知覚評価値予測にも拡張可能である．今後の研究では，BOK-PRP が自然性や感情強度を超えた多様な観点から合成音声に対する人間の評価を効果的に向上させることができるかを探究する予定である．

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] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, “Xtts: a massively multilingual zero-shot text-to-speech model,” in *Proc. INTERSPEECH*, 2024, pp. 4978–4982.

[2] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, “Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” *arXiv preprint arXiv:2407.05407*, 2024.

[3] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, “Speak, read and prompt: High-fidelity text-to-speech with minimal supervision,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.

[4] Y. Song, Z. Chen, X. Wang, Z. Ma, G. Yang, and X. Chen, “Tacolm: Gated attention equipped codec language model are efficient zero-shot text to speech synthesizers,” in *Proc. INTERSPEECH*, 2024, pp. 4433–4437.

[5] J. Xue, Y. Deng, Y. Han, Y. Gao, and Y. Li, “Improving audio codec-based zero-shot text-to-speech synthesis with multi-modal context and large language model,” in *Proc. INTERSPEECH*, 2024, pp. 682–686.

[6] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *arXiv preprint arXiv:2210.13438*, 2022.

[7] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2023.

[8] 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 (TASLP)*, vol. 30, pp. 495–507, 2021.

[9] 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 (TASLP)*, vol. 31, pp. 2523–2533, 2022.

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

[11] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, “SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities,” in *Proc. Findings of Empirical Methods in Natural Language Processing (EMNLP Findings)*, Dec. 2023, pp. 15 757–15 773.

[12] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, “Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2406.05370*, 2024.

[13] W. Chengyi, C. Sanyuan, W. Yu, Z. Ziqiang, Z. Long, L. Shujie, C. Zhuo, L. Yanqing, W. Huaming, L. Jinyu, H. Lei, Z. Sheng, and W. Furu, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[14] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 889–898.

[15] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[16] K. Yamauchi, W. Nakata, Y. Saito, and H. Saruwatari, "Decoding strategy with perceptual rating prediction for language model-based text-to-speech synthesis," in *Proc. Audio Imagination: NeurIPS 2024 Workshop on AI-Driven Speech, Music, and Sound Generation*, 2024.

[17] 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.

[18] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L. Paola Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *Odyssey 2024: The Speaker and Language Recognition Workshop)*, 2024.

[19] H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang, "Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 8642–8655.

[20] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[21] W. Li, P. Yang, Y. Zhong, Y. Zhou, Z. Wang, Z. Wu, X. Wu, and H. Meng, "Spontaneous style text-to-speech synthesis with controllable spontaneous behaviors based on language models," in *Proc. INTERSPEECH*, 2024, pp. 1785–1789.

[22] K. Mitsui, Y. Hono, and K. Sawada, "Towards human-like spoken dialogue generation between ai agents from written dialogue," *arXiv preprint arXiv:2310.01088*, 2023.