Decoding Strategy with Perceptual Rating Prediction for Language Model-Based Text-to-Speech Synthesis

Kazuki Yamauchi Wataru Nakata Yuki Saito Hiroshi Saruwatari
The University of Tokyo
yamauchi-kazuki042@g.ecc.u-tokyo.ac.jp

Abstract

Recently, text-to-speech (TTS) synthesis models that use language models (LMs) to autoregressively generate discrete speech tokens, such as neural audio codec, have gained attention. They successfully improve the diversity and expressiveness of synthetic speech while addressing repetitive generation issues by incorporating sampling-based decoding strategies. However, sampling randomness can lead to undesirable output, such as artifacts, and destabilize the quality of synthetic speech. To address this issue, we propose *BOK-PRP*, a novel sampling-based decoding strategy for LM-based TTS. Our strategy incorporates best-of-K (BOK) selection process based on perceptual rating prediction (PRP), filtering out undesirable outputs while maintaining output diversity. Importantly, the perceptual rating predictor is trained with human ratings independently of TTS models, allowing BOK-PRP to be applied to various pre-trained LM-based TTS models without requiring additional TTS training. Results from subjective evaluations demonstrate that BOK-PRP significantly improves the naturalness of synthetic speech.

1 Introduction

Recently, text-to-speech (TTS) synthesis models incorporating language models (LMs) have seen significant advances [4, 8, 14, 22, 25]. These models typically represent speech as discrete tokens, such as neural audio codec [9, 15, 27], and generate these tokens autoregressively. One of the key factors driving the advancement of LM-based TTS is the application of techniques developed for text generation to speech synthesis [2, 3, 28]. For example, sampling-based decoding strategies developed for text generation are also effective in LM-based TTS for addressing repetitive generation issues caused by greedy decoding [6, 7]. Furthermore, LM-based TTS that incorporates sampling-based decoding improves the diversity of prosody, such as duration, enhancing expressiveness.

However, sampling randomness introduces an issue: it can result in undesirable outputs and destabilize generation. To alleviate this, top-k sampling [10] and top-p sampling [11] limit candidate tokens to those with higher probabilities. However, narrowing down candidates reduces output diversity and leads to repetitive generation issues, making them insufficient for fundamentally addressing the issue.

Recently, decoding strategies have been extensively explored to control text generation [18, 20, 23, 26], including efforts to suppress undesirable outputs. For instance, controlled decoding [17] involves sampling multiple candidate tokens and selecting the best tokens based on predicted human preference scores. In contrast, decoding strategies for speech synthesis to prevent sampling randomness from generating undesirable output, such as artifacts, have not been thoroughly explored.

Motivated by this, we propose *BOK-PRP*, a novel sampling-based decoding strategy for LM-based TTS, which incorporates best-of-K (BOK) selection based on perceptual rating prediction (PRP). BOK-PRP involves sampling K blocks (or sequences) of discrete speech tokens and selecting the one with the highest rating given by an external perceptual rating predictor. While conventional strategies

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

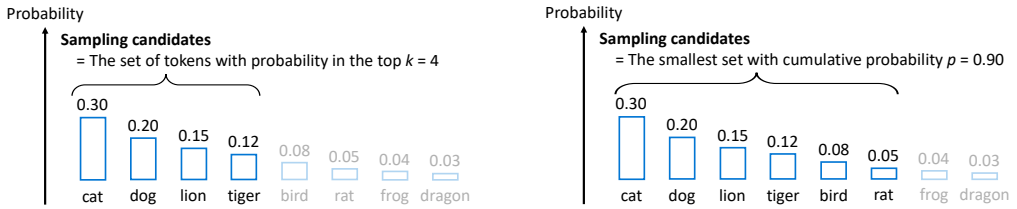
select tokens based solely on the probabilities computed by an LM, BOK-PRP leverages perceptual rating prediction to filter out undesirable outputs while maintaining output diversity. Importantly, the perceptual rating predictor is trained with human ratings independently of TTS models, allowing BOK-PRP to be applied as an inference-time add-on for various pre-trained LM-based TTS models.

In this paper, we focus on the naturalness mean opinion score (MOS) as the perceptual rating and conduct subjective evaluations to assess whether BOK-PRP effectively improves the naturalness of synthetic speech. Note that it is not obvious whether selecting samples with higher predicted MOS will result in an improvement in the actual MOS. Indeed, our experiments provide an interesting insight: excessively large K results in overfitting to MOS prediction, which in turn degrades the subjective naturalness of synthetic speech. Also, audio samples are available on our demo page [1].

Our contributions are summarized as follows:

- We propose BOK-PRP, a promising decoding strategy for LM-based TTS, which leverages perceptual rating prediction to filter out undesirable outputs while maintaining output diversity.
- We demonstrate that BOK-PRP effectively addresses the instability issues associated with conventional sampling-based strategies and significantly improves the naturalness of synthetic speech.

2 Related works



(a) Overview of top-k sampling ($k = 4$).

(b) Overview of top-p sampling ($p = 0.90$).

Figure 1: Conventional sampling-based decoding strategies

2.1 Conventional decoding strategies

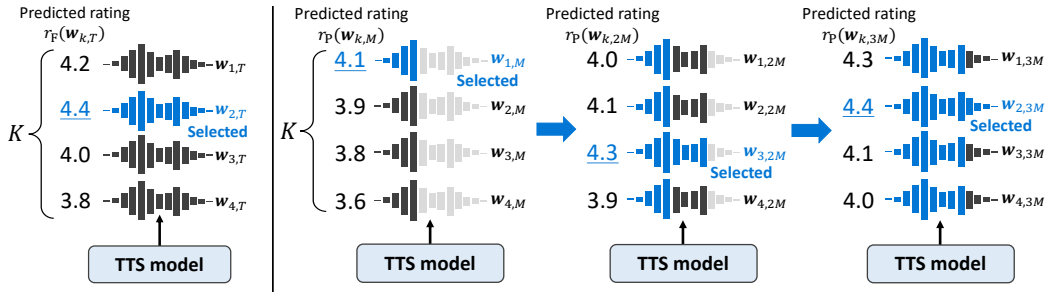
Decoding is the process of selecting output tokens based on the probability distribution computed by LMs during autoregressive generation. Greedy decoding is a simple *deterministic* decoding strategy that selects the token with the highest probability as the next token. However, greedy decoding often leads to repetitive generation, causing the output to get stuck in loops of repeating the same tokens.

Sampling-based decoding strategies, such as top-k sampling [10] and top-p sampling [11], involve *stochastically* selecting tokens based on the distribution of tokens, which introduces diversity and effectively addresses repetitive generation issues. Additionally, top-k and top-p sampling narrow down sampling candidate tokens to suppress undesirable outputs. Specifically, top-k sampling draws tokens from the set of tokens with the 1st to kth highest probabilities (Fig. 1a). Top-p sampling draws tokens from the smallest set of tokens whose cumulative probability exceeds a threshold p (Fig. 1b). However, narrowing down the candidates reduces output diversity and can lead to repetitive generation. Thus, filtering out undesirable outputs while maintaining diversity remains challenging.

2.2 Controlled decoding

Controlled decoding [17] is a decoding strategy that controls text generation through block-wise best-of-K selection process based on preference score prediction. Specifically, during autoregressive generation, blocks of M tokens are sampled across K blocks, and the block with the highest predicted preference score is selected. The preference score predictor (called the prefix scorer in the original paper) is trained to take a partially decoded sequence as input and estimate the preference score for the sequence fully decoded by an LM. This method is highly modular and can be applied as an inference-time add-on to an unseen base model, effectively improving the output's preference score.

[1] `https://kyamauchi1023.github.io/BOK-PRP/`



(a) Sequence-wise BOK-PRP (b) Block-wise BOK-PRP

Figure 2: The decoding process of the two variants of BOK-PRP. (a) Sequence-wise BOK-PRP synthesizes K speech samples and selects the one with the highest predicted rating. (b) Block-wise BOK-PRP is a variant of controlled decoding, adapted for discrete speech token generation.

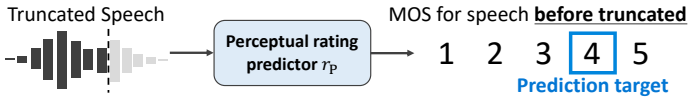


Figure 3: The training process of the perceptual rating predictor for block-wise BOK-PRP.

3 Methods

Fig. 2 illustrates the decoding process of BOK-PRP. BOK-PRP involves the best-of-K selection process, where the sample with the highest predicted perceptual rating, specifically predicted naturalness MOS in this paper, is selected from the K samples drawn from an LM. We employ UTMOS [21], a widely used naturalness MOS predictor, as the rating predictor. We explore two variants: (a) sequence-wise BOK-PRP and (b) block-wise BOK-PRP. For text generation, the block-wise manner can generate highly-rated tokens more efficiently than the sequence-wise manner [17].

In the following, let $\boldsymbol{x}$ be an input text and let p denote a pre-trained LM that computes the probabilities of discrete speech tokens autoregressively. A set of K samples, $Y_{K,T} = \{\boldsymbol{y}_{1,T}, \ldots, \boldsymbol{y}_{K,T}\}$, is drawn from $p(\cdot|\boldsymbol{x})$, where each sample $\boldsymbol{y}_{k,T} = [y_{k,1}, \ldots, y_{k,T}]$ represents a token sequence of length T. Also, let $\boldsymbol{w}_{k,T} = \mathrm{Dec}(\boldsymbol{y}_{k,T})$ be a corresponding speech waveform re-synthesized by the decoder Dec. BOK-PRP can be applied to any LM-based TTS models that fit the described formulation.

3.1 Sequence-wise BOK-PRP

In this strategy, we simply draw fully decoded samples $Y_{K,T}$ from $p(\cdot|\boldsymbol{x})$, and select the optimal sequence $\hat{\boldsymbol{y}}_T$ with the highest predicted perceptual rating, denoted as:

$$\hat{\boldsymbol{y}}_T = \underset{\boldsymbol{y}_{k,T} \in Y_{K,T}}{\operatorname{argmax}} \; r_{\mathrm{F}}\left(\mathrm{Dec}\left(\boldsymbol{y}_{k,T}\right)\right), \tag{1}$$

where r_{F} denotes the perceptual rating predictor for *fully* synthesized speech. In this paper, we simply employ the pre-trained UTMOS model as is for r_{F}.

3.2 Block-wise BOK-PRP

In this strategy, we iteratively decode tokens for each block of M tokens. Given partially decoded tokens $\hat{\boldsymbol{y}}_{(n-1)M}$ in the nth iteration, we sample continuing token candidates $Y_{K,nM}$ from $p(\cdot|\boldsymbol{x}, \hat{\boldsymbol{y}}_{(n-1)M})$, and select the one $\hat{\boldsymbol{y}}_{nM}$ with the highest predicted perceptual rating, denoted as:

$$\hat{\boldsymbol{y}}_{nM} = \underset{\boldsymbol{y}_{k,nM} \in Y_{K,nM}}{\operatorname{argmax}} \; r_{\mathrm{P}}\left(\mathrm{Dec}\left(\boldsymbol{y}_{k,nM}\right)\right), \tag{2}$$

where r_{P} denotes the perceptual rating predictor for *partially* synthesized speech. This iteration is repeated until tokens are fully decoded. Unlike r_{F}, since r_{P} takes as input a waveform re-synthesized from partially decoded tokens, it is inappropriate to directly use UTMOS for r_{P}. Therefore, we train

the UTMOS model by introducing a truncation process, as illustrated in Fig. 3. Specifically, during training, the input speech was truncated at a random time, and the model was trained to predict MOS for speech before the truncation. The truncation time was selected in 0.5-second increments.

4 Experiments

We conducted MOS tests to evaluate BOK-PRP. Additionally, we investigated the impact of the number of samples K on the subjective naturalness of synthetic speech through an ablation study.

4.1 Experimental conditions

LM-based TTS model Fig. 4 illustrates an overview of the LM-based TTS model used in our experiments. We employed Descript Audio Codec (DAC) [15] for discrete speech tokenization and Transformer encoder-decoder [24] to generate DAC tokens. The model architecture and training settings followed those of DiscreteTTS-v2.2 in UTDUSS [19], which achieved top performance in TTS track of Interspeech2024 speech processing using discrete speech unit challenge [5].

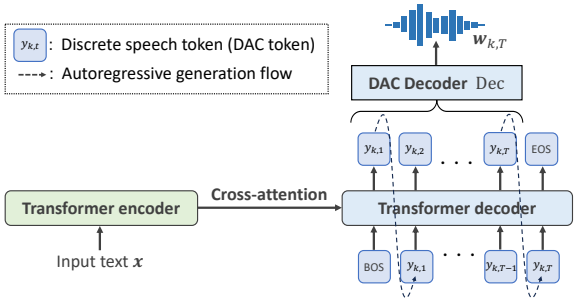


Figure 4: The LM-based TTS model we used.

Dataset We used LJSpeech [12], which contains speech from a single English speaker, to train the TTS model. All speech samples were resampled to 16 kHz, and the dataset was divided into training (12,600 utterances), validation (250 utterances), and evaluation (250 utterances) subsets.

Compared methods In the main experiment, we compared the following decoding strategies:

- **Greedy decoding**: A deterministic strategy that selects the token with the highest probability.
- **Naive sampling**: A sampling-based strategy that draws a token from the full set of available tokens, equivalent to top-k sampling with k set to the maximum and top-p sampling with p set to 1.
- **Top-k top-p sampling**: A sampling-based strategy that first uses top-k to narrow down the sampling candidates and then applies top-p to filter the candidates further.
- **Sequence-wise BOK-PRP**: The proposed strategy with sequence-wise best-of-K selection process.
- **Block-wise BOK-PRP**: The proposed strategy with block-wise best-of-K selection process.

For top-k top-p sampling, the values of k, p, and temperature parameter were set to 190, 0.50, and 0.40, respectively. These parameters were tuned with Optuna [1], maximizing UTMOS on the validation set. This temperature parameter value was consistently applied to other sampling processes as well. For BOK-PRP, the value of K and M were set to 8 and 16, respectively. Note that a block of 16 tokens corresponds to approximately 0.5 seconds of speech. The model architecture and training settings of the perceptual rating predictor for block-wise BOK-PRP followed the official implementation of UTMOS[2], except for the truncation process described in Section 3.2.

Evaluation metrics We conducted MOS tests via crowdsourcing to assess the subjective naturalness of synthetic speech. Participants rated the naturalness—how human-like and natural it sounded—of randomly selected speech samples on a 5-point scale, ranging from 1 (very unnatural) to 5 (very natural). Participants were limited to native English speakers. A total of 200 participants took part in the evaluations, with each participant rating 24 samples. The significance of the differences in MOS was assessed using a Student's t-test at a 5% significance level. Additionally, we calculated the average UTMOS score for all speech synthesized from the transcriptions in the evaluation set.

4.2 Experimental results

Main results: BOK-PRP significantly improves subjective naturalness. The evaluation results are shown in Table 1. The naturalness MOS of speech synthesized using block-wise BOK-PRP was

[2] https://github.com/sarulab-speech/UTMOS22

Table 1: Naturalness MOS with 95% confidence intervals and average UTMOS for the compared methods. **Bold** indicates the highest MOS. Ground truth means natural speech samples from LJSpeech.

Method	Naturalness MOS ↑	UTMOS ↑
Greedy decoding	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-wise BOK-PRP (proposed)	3.71 ± 0.07	4.46
Block-wise BOK-PRP (proposed)	$\mathbf{3.73 \pm 0.07}$	4.43
Ground truth	3.92 ± 0.07	4.43

Table 2: Ablation study on the number of samples K. Naturalness MOS with 95% confidence intervals and average UTMOS. **Bold** indicates the highest MOS.

K	Naturalness MOS ↑	UTMOS ↑
2	3.72 ± 0.08	4.40
4	3.74 ± 0.08	4.43
8	$\mathbf{3.83 \pm 0.07}$	4.43
16	3.79 ± 0.07	4.45
32	3.65 ± 0.08	4.46

significantly higher compared to that using conventional decoding strategies, such as top-k top-p sampling. This suggests that incorporating our proposed best-of-K selection process with perceptual rating prediction effectively enhances the naturalness of synthetic speech.

We also found that there was no significant difference between block-wise and sequence-wise BOK-PRP. In this paper, we conducted experiments on the simple single-speaker TTS task, and the training data, which contained only a few seconds of speech with limited prosodic diversity, resulted in minimal efficiency gains from the block-wise manner. On the other hand, for advanced tasks that require generating diverse expressions, such as spontaneous style TTS [16] and emotional TTS [13], token samples become diverse, making block-wise manner potentially more efficient than sequence-wise manner. Further experiments on these tasks will be conducted in future work.

Additionally, we found that while UTMOS was correlated with MOS, the comparison results based on UTMOS did not always align with those based on MOS. This suggests that selecting tokens based on higher predicted perceptual rating (i.e., UTMOS) does not always lead to an improvement in the actual subjective rating (i.e., naturalness MOS) of the synthetic speech.

Ablation study on K: Excessively large K degrades naturalness. We evaluated speech synthesized by block-wise BOK-PRP while varying K across values of 2, 4, 8, 16, and 32. The evaluation results are shown in Table 2. The naturalness MOS of speech synthesized with K set to 8 is significantly higher compared to when K is set to 2 or 32. This suggests that while somewhat large K is necessary to stabilize the naturalness of synthetic speech, excessively large K results in overfitting to perceptual rating prediction, which in turn degrades the subjective naturalness.

5 Conclusions and future directions

We propose BOK-PRP, a novel sampling-based decoding strategy for LM-based TTS, which introduces best-of-K (BOK) selection process based on perceptual rating prediction (PRP). We demonstrated through subjective evaluations that BOK-PRP outperforms conventional sampling-based strategies, such as top-k top-p sampling, significantly improving the naturalness of synthetic speech.

In future work, we will investigate the effectiveness of BOK-PRP for LM-based spontaneous style TTS [16] and LM-based emotional TTS [13]. Furthermore, although our primary focus in this paper was on enhancing the naturalness MOS of synthetic speech, BOK-PRP can be extended to perceptual rating predictions from other perspectives, such as prosodic naturalness and emotional suitability. In future work, we will explore whether BOK-PRP can effectively improve human ratings of synthetic speech from various perspectives beyond naturalness. We also hope that our work advances future research on decoding strategies for LM-based audio generation beyond just TTS.

Acknowledgements

This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011 (experimental evaluation) and JST, ACT-X, JPMJAX23CB (algorithm development). Also, this work was partially funded by the Royal Society of New Zealand Catalyst Seeding programme: New Zealand–Japan Joint Research Projects (JSP-UOA1901-JR), the Kajima Foundation's Support Program for International Joint Research Activities (2024-kyodoshin-05) and the Acoustics and Vibration Research Centre at the University of Auckland. We would also like to express our gratitude to Osamu Take from the University of Tokyo for his valuable discussions.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2623–2631, Aug. 2019.

[2] 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*, 31:2523–2533, 2022.

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

[4] 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*, pages 4978–4982, 2024.

[5] X. Chang, J. Shi, J. Tian, Y. Wu, Y. Tang, Y. Wu, S. Watanabe, Y. Adi, X. Chen, and Q. Jin. The interspeech 2024 challenge on speech processing using discrete units. In *Proc. INTERSPEECH*, pages 2559–2563, 2024.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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)*, Apr. 2020.

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

[13] S. Ji, J. Zuo, M. Fang, Z. Jiang, F. Chen, X. Duan, B. Huai, and Z. Zhao. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305, 2024.

[14] 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*, 11:1703–1718, 2023.

[15] 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.

[16] 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*, pages 1785–1789, 2024.

[17] 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 *Proc. International Conference on Machine Learning (ICML)*, pages 36486–36503, Jul. 2024.

[18] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[19] W. Nakata, K. Yamauchi, D. Yang, H. Hyodo, and Y. Saito. UTDUSS: UTokyo-SaruLab System for Interspeech2024 Speech Processing Using Discrete Speech Unit Challenge. *arXiv preprint arXiv:2403.13720*, 2024.

[20] L. Qin, S. Welleck, D. Khashabi, and Y. Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9538–9551, 2022.

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

[22] 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*, pages 4433–4437, 2024.

[23] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3008–3021, 2020.

[24] 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. Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[25] 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*, pages 682–686, 2024.

[26] K. Yang and D. Klein. FUDGE: Controlled text generation with future discriminators. In *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3511–3535, June 2021.

[27] 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*, 30:495–507, 2021.

[28] 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)*, pages 15757–15773, Dec. 2023.