ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889794

Speaking from Coarse to Fine: Improving Neural Codec Language Model via Multi-Scale Speech Coding and Generation

Haohan Guo†, Fenglong Xie‡, Dongchao Yang†, Xixin Wu†, Helen Meng†
†The Chinese University of Hong Kong, Hong Kong SAR, China
‡Xiaohongshu Inc., Shanghai, China
{hguo, dcyang, wuxx, hmmeng}@se.cuhk.edu.hk, fenglongxie@xiaohongshu.com

***Abstract*—The neural codec language model (CLM) has demonstrated remarkable performance in text-to-speech (TTS) synthesis. However, troubled by "recency bias", CLM lacks sufficient attention to coarse-grained information at a higher temporal scale, often producing unnatural or even unintelligible speech. This work proposes CoFi-Speech, a coarse-to-fine CLM-TTS approach, employing multi-scale speech coding and generation to address this issue. We train a multi-scale neural codec, CoFi-Codec, to encode speech into a multi-scale discrete representation, comprising multiple token sequences with different time resolutions. Then, we propose CoFi-LM that can generate this representation in two modes: the single-LM-based chain-of-scale generation and the multiple-LM-based stack-of-scale generation. In experiments, CoFi-Speech significantly outperforms single-scale baseline systems on naturalness and speaker similarity in zero-shot TTS. The analysis of multi-scale coding demonstrates the effectiveness of CoFi-Codec in learning multi-scale discrete speech representations while keeping high-quality speech reconstruction. The coarse-to-fine multi-scale generation, especially for the stack-of-scale approach, is also validated as a crucial approach in pursuing a high-quality neural codec language model for TTS.**

***Index Terms*—Neural Codec, Language Model, TTS, VQ-VAE, Representation Learning**

I. Introduction

The success of large language models (LLMs) in text domain [1], [2], [3] has demonstrated their great capability in discrete sequence generation. It also inspires the birth of a new text-to-speech synthesis (TTS) paradigm based on the neural codec language model (CLM) [4], [5], [6], which treats TTS as a next-token prediction task. This framework usually relies on a neural codec [7], [8], [9] to encode the speech audio into discrete tokens, which can be incorporated with the text sequence and generated by the LM, i.e. an auto-regressive decoder. Finally, we obtain the speech audio from these generated speech tokens via the codec decoder.

However, different from the text, the discrete speech sequence is much longer to keep sufficient capacity to preserve phonetic and acoustic information. This long sequence length not only increases the complexity of TTS modeling but aggregates the "recency bias" of LMs [10], [11], i.e. overly focusing on recent tokens during auto-regressive generation. This issue makes LM focus less on coarse-grained information [12], e.g. phonetics, prosody, and speaking style at higher and different temporal scales, hence causing unstable TTS performance, producing unnatural or even unintelligible speech. Although monotonic attention constraints [13], [14], [15] are proposed to fix stability issues, they still cannot solve "recency bias" fundamentally. Some works [5], [16], [17] turn to directly model shorter speech sequences with a larger frameshift to avoid this issue, but limits the fine-grained expression of LMs in TTS. This dilemma implies the necessity of applying guidance to LMs to pay sufficient attention to both coarse-grained and fine-grained information of speech.

In this work, we propose a novel CLM-based zero-shot TTS approach, CoFi-Speech, that generates speech in a coarse-to-fine manner via a multi-scale speech coding and generation approach. In this framework, the multi-scale speech codec, CoFi-Codec, decomposes speech into multiple discrete sequences with different temporal resolutions and decodes them back with a high reconstruction quality. Then, we propose two LM-based approaches to predict this multi-scale speech representation from coarse to fine: single-LM-based chain-of-scale generation and multiple-LM-based stack-of-scale generation. In experiments, we present subjective and objective evaluations to demonstrate that CoFi-Speech significantly outperforms baseline systems based on single-scale speech sequences on naturalness and similarity, where stack-of-scale generation performs best. Finally, we conduct detailed ablation studies to analyze multi-scale coding and generation, to further validate the effectiveness of "speaking from coarse to fine" in achieving high-quality CLM-based TTS.

II. CoFi-Speech

A. CoFi-Codec

CoFi-Codec aims to decompose speech into multiple discrete sequences with different resolutions to provide a multi-scale speech representation [18], [19], [12]. Fig. 1 (a) presents a three-scale CoFi-Codec model architecture. We first convert speech signals into the Mel spectrogram, and employ multiple encoder blocks to down-sample it into sequences at different temporal scales. Specifically, each encoder block is applied with a 1-D convolutional ResNet block and a

Work performed during the first author's internship at Xiaohongshu. This project is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Fund.

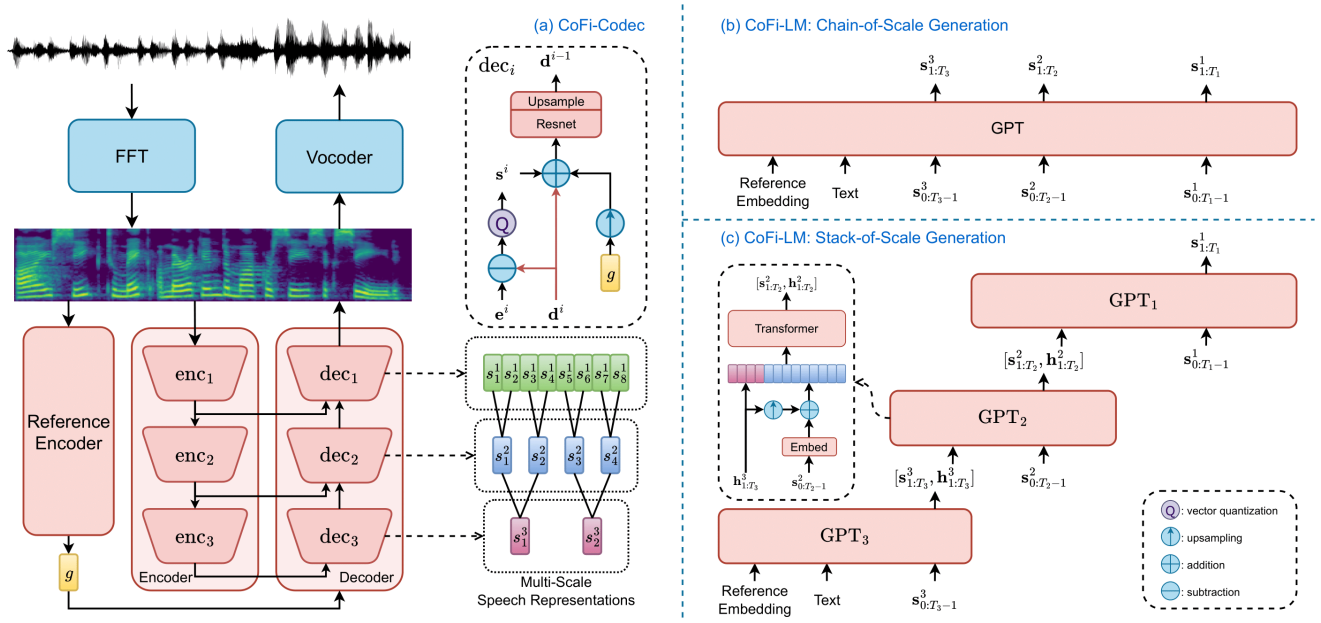

Fig. 1. The system architecture of a three-scale CoFi-Speech, comprising: a CoFi-Codec for speech encoding and decoding, and two types of CoFi-LM for text-to-speech generation. The trainable modules are colored red, and each VQ operation comprises a trainable codebook.

strided convolutional layer for down-sampling. Meanwhile, we follow SoCodec [16] to apply an ECAPA-TDNN-based [20] reference encoder to extract a global embedding g from the Mel spectrogram to capture time-invariant information, e.g., speaker identity, global speaking style, acoustic environment, etc.

In decoding, we apply an equal number of decoder modules to quantize and decode these encoding sequences in a reversed order to reconstruct the Mel spectrogram from coarse to fine. For example, in i-th decoder block, we first extract the residual sequence at this stage by performing $\mathbf{e}^i - \mathbf{d}^i$, i.e. removing high-scale information from the decoding sequence $\mathbf{d}^i$ out of the encoding sequence $\mathbf{e}^i$. We then apply single-stream [21] or multi-stream vector quantization, e.g. RQ [22] and PQ [23], to obtain the quantized speech sequence $\mathbf{s}^i$. It is added with $\mathbf{d}^i$ and global embedding g, subsequently processed by a ResNet block and a transposed convolutional layer for up-sampling to produce the next decoding sequence $\mathbf{d}^{i-1}$. Notably, the highest-scale decoder block has no higher-scale decoding sequence as the input. Finally, the Mel spectrogram is reconstructed from the decoder, and converted to the waveform via a pre-trained neural vocoder.

In training, the model tends to overfit low-scale sequences with richer information, leading to high-scale representation collapse, i.e. nothing is preserved in the sequence. Hence, we propose scale-wise nested dropout (SWND), which masks quantized sequences $\mathbf{s}_{1:b}$ given an index b randomly sampled from $[0, N_s - 1]$, where $b = 0$ indicates that no sequence is masked. This approach forces high-scale sequence to preserve effective information for reconstruction, avoiding representation collapse. Moreover, we apply GAN training [17], [16] on CoFi-Codec to improve the generation quality of the Mel spectrogram. The loss function is written as follows:

$$\mathcal{L}_c = \lambda_{vq} * \mathcal{L}_{vq} + \lambda_{reg} * \mathcal{L}_{reg} + \lambda_{adv} * \mathcal{L}_{adv} \quad (1)$$

where $\mathcal{L}_{vq}$ is the vector-quantization loss, i.e. the averaged L2 loss between embeddings before and after VQ, $\mathcal{L}_{reg}$ is the L2-based regression loss on Mel spectrogram, and $\mathcal{L}_{adv}$ is the adversarial loss following [16].

B. CoFi-LM

To predict the multi-scale discrete representation in a coarse-to-fine manner, we propose CoFi-LM in two styles: chain-of-scale (CoS) generation and stack-of-scale (SoS) generation.

Chain-of-scale generation is a simple extension of CLM to the multi-scale speech representation, which employs one LM to predict speech sequences in descending order of the temporal scale, as shown in Fig. 1 (b). In training, the text and speech sequences are embedded into the same space and processed by a decoder-only Transformer to estimate the probabilities of the next speech tokens. Notably, for multi-stream speech sequence, we follow the "delay pattern"[24], [25], [26] to achieve the multi-stream prediction. This approach explicitly constrains the LM to model multi-scale information to alleviate "recency bias". However, concatenating multiple sequences leads to a longer chain, causing higher computing costs and modeling complexity. To mitigate this issue, we propose stack-of-scale generation.

Stack-of-scale generation employs a stack of LMs to generate speech sequences at different scales in cascade, as shown in Fig. 1 (c). It first generates the highest-scale sequence from the text and the reference embedding. The hidden states $\mathbf{h}^3_{1:T_3}$

in the last transformer layer are employed to generate the next-scale sequence. It is placed at the head of the input sequence as the prompt, and also upsampled to be added to the lower-scale speech sequence to emphasize the alignment between these two sequences. In this way, we recursively generate all speech sequences from coarse to fine. This stage-wise approach enables us to better model multi-scale information while avoiding introducing long context to transformers.

Finally, we apply in-context learning (ICL) [27], [28] to CoFi-LMs for zero-shot TTS. Specifically, we extract speaker embedding and speech tokens from the reference audio as speech prompts. The transcript of the reference audio is also concatenated with the input text as the text prompt.

III. Experimental Protocol

A. Training Configuration

We conduct experiments with WenetSpeech4TTS (Basic) [29], a Chinese dataset with 7k hours of speech data. All audio files are normalized to the sample rate of 16kHz and converted to 80-dim Mel spectrograms with a frameshift of 10ms. We create a byte-pair encoding (BPE) based text dictionary with 8192 tokens to encode the text.

CoFi-Codec is applied with 512-dim ResNet modules, each consisting of four residual units with two 1-D convolutional layers. enc_1 and dec_1 are applied with four more residual units to better encode and reconstruct Mel spectrograms. We train a three-scale CoFi-Codec with ordered product quantization (OPQ) [16] to compress the Mel spectrogram into two single-stream sequences with frameshifts of 120ms and 40ms and one four-stream sequence with a frameshift of 20ms. Each stream in each sequence has a codebook with 16,384 codewords. In training, we set $\lambda_{vq} = 1, \lambda_{reg} = 1, \lambda_{adv} = 0.1$ and use EMA [21] to update codebooks with the decay rate of 0.99. The discriminator in Mega-TTS [30] is employed for adversarial training. The sampling probabilities for SWND are $p_0 = 0.8, p_1 = 0.1, p_2 = 0.1$. Finally, a pre-trained BigVGAN-based [31] neural vocoder is employed to convert the Mel spectrogram to the waveform.

LMs are applied with 12-layer decoder-only Transformers with a feature dimension of 1024. For multi-stream speech sequences, we apply a "delay pattern". We employ the sampling strategy with a top-p of 0.8, a top-k of 50, and a repetition penalty of 2.0. CoFi-Codec and LMs are trained using AdamW [32] for 100k iterations with a batch size of 1.6k seconds. The learning rate exponentially decays from 3×10^{-4} to 1×10^{-4}.

B. Evaluation Metrics

We create an 860-utterance test set with high diversity in speaking style and audio quality for system evaluation. In zero-shot TTS, each utterance is paired with another out-of-training-set audio file as the reference audio. We calculate character error rate (CER, %), and speaker similarity (SIM, $\times 10^{-2}$) to measure intelligibility, and speaker similarity.[1] In IV-A, we use a subset with 100 test cases for subjective evaluation, where each case includes includes audio samples generated by different models for the same input text and reference audio. There are 10 native speakers invited to the test to rate each audio with a score ranging from 1 to 5 in terms of naturalness (NMOS) and speaker similarity (SMOS), respectively. In codec evaluation, We adopt Mel-cepstrum distortion (MCD, dB) and CER to measure reconstruction quality.

[1]The ASR tool for transcribing is available at https://github.com/modelscope/FunASR. The tool for extracting speaker embedding is available at https://huggingface.co/Wespeaker/wespeaker-cnceleb-resnet34

IV. Results

A. TTS System Comparison

As shown in Table I, we compare CoFi-Speech with baseline systems: VALL-E [33] and SoCodec-TTS, trained with the same dataset as ours, and the SotA industrial TTS system, CosyVoice trained with over 100k hours of data. The baseline system VALL-E auto-regressively generates an eight-stream speech sequence with a short frameshift of 20ms along both time and stream axes. This model troubled by "recency bias" presents serious stability issues on this challenging test set, performing the worst synthesis quality. SoCodec-TTS compresses speech into a short speech sequence with a frameshift of only 120ms to achieve stable LM inference. It presents high naturalness but low speaker similarity due to the lack of fine-grained information. In contrast, CoFi-Speech leverages both coarse-grained and fine-grained information, presenting better naturalness and similarity, especially for the SoS-based approach with the highest NMOS and SMOS, even outperforming CosyVoice trained with much more data. This significant improvement underscores the efficacy of CoFi-Speech in CLM-TTS.[2]

TABLE I
THE SUBJECTIVE TESTS OF CLM-BASED TTS SYSTEMS ON NATURALNESS AND SPEAKER SIMILARITY WITH 95% CONFIDENCE INTERVAL (CI).

Systems		NMOS ↑	SMOS ↑
CosyVoice [28]		4.28 ± 0.09	3.70 ± 0.12
VALL-E [29]		3.46 ± 0.10	2.92 ± 0.12
SoCodec-TTS [16]		4.30 ± 0.08	3.51 ± 0.12
CoFi-Speech	-CoS	4.12 ± 0.09	3.75 ± 0.12
	-SoS	**4.42 ± 0.08**	**3.90 ± 0.11**

B. Multi-Scale Speech Coding

We conduct objective evaluations to investigate if CoFi-Codec can decompose speech into sequences at multiple temporal scales while keeping high reconstruction quality. As shown in Fig. 2, we evaluate speech reconstructed from top-b scales. For example, for a three-scale CoFi-Codec, "$b = 2$" means that we only reconstruct speech using $\mathbf{s}^3$ and $\mathbf{s}^2$, and mask $\mathbf{s}^1$ with all-zero vectors. The result demonstrates that CoFi-Codec learns a good multi-scale representation with the help of SWND. The high-scale sequence preserves more effective information, presenting lower reconstruction loss

[2]Samples are available at https://hhguo.github.io/DemoCoFiSpeech

over that of codec without SWND. Moreover, compared with the single-scale CoFi-Codec and Encodec [7] with a four-stream sequence on a frameshift of 20ms, CoFi-Codec presents the lowest MCD and CER. This result validates CoFi-Codec as the expected codec, providing multi-scale discrete speech representation and keeping high reconstruction quality.

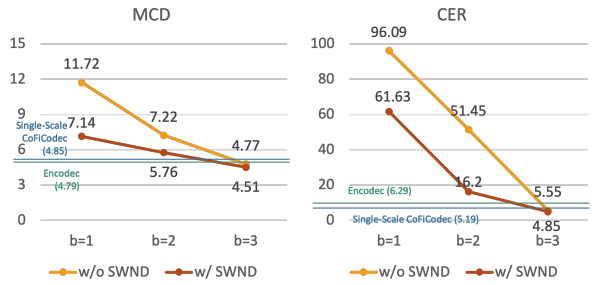

Fig. 2. The objective evaluation of multi-scale coding with or without scale-wise nested dropout (SWND).

C. Multi-Scale Speech Generation

We also conduct objective evaluations to investigate if multi-scale speech generation can improve TTS quality. As shown in Fig. 3, we compare LMs based on three codecs: a single-scale codec with a frameshift of 20ms; a two-scale CoFi-Codec with frameshifts of 120ms and 20ms; the three-scale CoFi-Codec introduced in Sec. III-A. All codecs share the same model and training configurations. The results demonstrate that multi-scale speech generation can significantly improve TTS quality, presenting improved intelligibility and speaker similarity in CoS- and SoS-based LMs as more scales are used. Moreover, compared with CoS-LM, SoS-LM performs better with a lower CER and a higher SIM. To eliminate the impact of different numbers of parameters between CoS-LM and SoS-LM, we train a two-scale CoS-large with 24 transformer layers. It doubles the inference cost, presenting a better CER and SIM than the smaller one, but still has a gap to the two-scale and three-scale SoS-LMs. This result substantially validates the SoS as a better multi-scale generation approach.

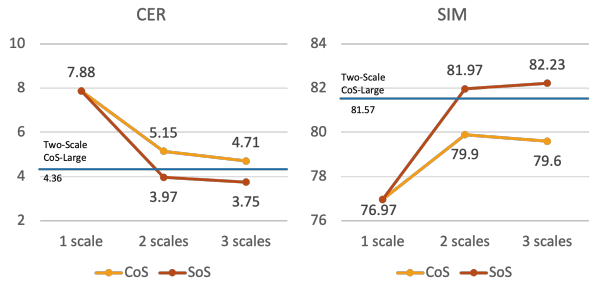

Fig. 3. The objective evaluation of multi-scale generation.

D. Recency Bias: Attention Visualization

Finally, we attempt to investigate the "recency bias" issue in CLM-TTS by visualizing attention maps of different LMs when synthesizing the same utterance. We aggregate attention maps across all heads and layers in one model by adding them together and apply the map with a scaling factor of 0.25 and a max clipping of 1.0 for clearer visualization. As shown in Fig. 4, single-scale LM shows significant recency bias, where each speech frame pays much more attention to the nearest speech tokens and text tokens only relevant to the current pronunciation. However, in CoFi-Speech-CoS, the high-scale sequence generation presents wider attention to the text and speech context, effectively modeling coarse-grained information. It then guides the next-scale sequence generation to render fine-grained details. Similarly, CoFi-Speech-SoS employs multiple LMs to achieve coarse-to-fine generation. The first LM absorbs richer contextual information to generate the high-scale sequence and guides the next LMs to generate lower-scale sequences recursively. This validates the effectiveness of "speaking from coarse to fine" in addressing "recency bias" via explicit multi-scale modeling.

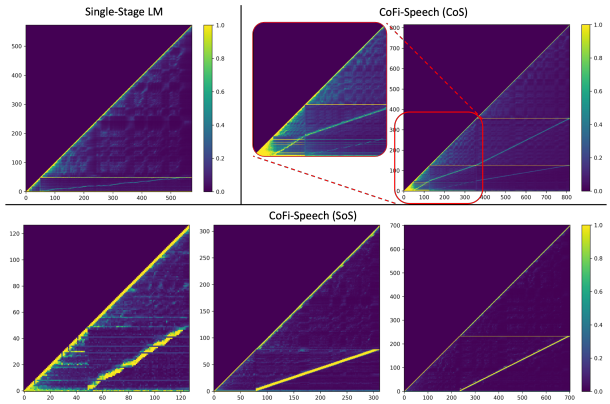

Fig. 4. The aggregated attention maps of different LMs. For CoFi-Speech (SoS), maps displayed from left to right refer to GPT_3, GPT_2, and GPT_1.

V. Conclusions

This work introduces CoFi-Speech, a novel approach to enhance CLM-based TTS through coarse-to-fine generation. The framework comprises two key components: 1) CoFi-Codec, a multi-scale speech codec that encodes speech into the discrete, multi-scale representation and decodes them back, and 2) CoFi-LM, which predicts this representation using two strategies: single-LM-based chain-of-scale generation and multiple-LM-based stack-of-scale generation. The zero-shot TTS experiment demonstrates that CoFi-Speech, especially for the stack-of-scale approach, significantly outperforms baseline systems using single-scale speech sequences in terms of naturalness and speaker similarity. The ablation study further validates that: 1) CoFi-Codec is capable of learning multi-scale discrete representations while preserving high-quality speech reconstruction, and 2) CoFi-LM effectively improves TTS performance via explicit multi-scale modeling. Finally, the attention visualization substantiates the efficacy of our approach in addressing "recency bias", providing compelling evidence for its effectiveness in improving CLM-based TTS.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Proc. NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[2] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[4] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[5] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[6] M. Lajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[7] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

[8] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "HiFi-Codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[9] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Proc. NeurIPS*, vol. 36, 2024.

[10] A. Peysakhovich and A. Lerer, "Attention sorting combats recency bias in long context language models," *arXiv preprint arXiv:2310.01427*, 2023.

[11] Z. Wang, H. Zhang, X. Li, K.-H. Huang, C. Han, S. Ji, S. M. Kakade, H. Peng, and H. Ji, "Eliminating position bias of language models: A mechanistic approach," *arXiv preprint arXiv:2407.01100*, 2024.

[12] H. Guo, F. Xie, X. Wu, F. K. Soong, and H. Meng, "Msmc-tts: Multi-stage multi-codebook vq-vae based neural tts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[13] B. Han, L. Zhou, S. Liu, S. Chen, L. Meng, Y. Qian, Y. Liu, S. Zhao, J. Li, and F. Wei, "VALL-E R: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment," *arXiv preprint arXiv:2406.07855*, 2024.

[14] C. Du, Y. Guo, H. Wang, Y. Yang, Z. Niu, S. Wang, H. Zhang, X. Chen, and K. Yu, "VALL-T: Decoder-only generative transducer for robust and decoding-controllable text-to-speech," *arXiv preprint arXiv:2401.14321*, 2024.

[15] H. Wang, C. Du, Y. Guo, S. Wang, X. Chen, and K. Yu, "Attention-constrained inference for robust decoder-only text-to-speech," *arXiv preprint arXiv:2404.19723*, 2024.

[16] H. Guo, F. Xie, K. Xie, D. Yang, D. Guo, X. Wu, and H. Meng, "SoCodec: A semantic-ordered multi-stream speech codec for efficient language model based text-to-speech synthesis," *arXiv preprint arXiv:2409.00933*, 2024.

[17] H. Li, L. Xue, H. Guo, X. Zhu, Y. Lv, L. Xie, Y. Chen, H. Yin, and Z. Li, "Single-Codec: Single-codebook speech codec towards high-performance speech generation," in *Proc. Interspeech*, 2024, pp. 3390–3394.

[18] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Proc. NeurIPS*, 2019, pp. 14 866–14 876.

[19] H. Guo, F. Xie, F. K. Soong, X. Wu, and H. Meng, "A multi-stage multi-codebook vq-vae approach to high-performance neural tts," in *Proc. Interspeech*, 2022.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*. ISCA, 2020, pp. 3830–3834.

[21] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. NeurIPS*, 2017.

[22] Y. Chen, T. Guan, and C. Wang, "Approximate nearest neighbor search by residual vector quantization," *Sensors*, vol. 10, no. 12, pp. 11 259–11 273, 2010.

[23] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, 2010.

[24] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. D'efossez, "Simple and controllable music generation," *Proc. NeurIPS*, vol. 36, 2024.

[25] T. Dang, D. Aponte, D. Tran, and K. Koishida, "LiveSpeech: Low-latency zero-shot text-to-speech via autoregressive modeling of audio discrete codes," in *Proc. Interspeech*, 2024, pp. 3395–3399.

[26] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint arXiv:2402.01912*, 2024.

[27] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao *et al.*, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[28] 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.

[29] L. Ma, D. Guo, K. Song, Y. Jiang, S. Wang, L. Xue, W. Xu, H. Zhao, B. Zhang, and L. Xie, "WenetSpeech4TTS: A 12,800-hour Mandarin TTS corpus for large speech generation model benchmark," in *Proc. Interspeech*, 2024, pp. 1840–1844.

[30] Z. Jiang, J. Liu, Y. Ren, J. He, Z. Ye, S. Ji, Q. Yang, C. Zhang, P. Wei, C. Wang, X. Yin, Z. MA, and Z. Zhao, "Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis," in *Proc. ICLR*, 2024.

[31] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. ICLR*, 2023.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2018.

[33] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.