VIBEVOICE Technical Report

Zhiliang Peng[*], **Jianwei Yu**[*], **Wenhui Wang**[*], **Yaoyao Chang**[*], **Yutao Sun**[*], **Li Dong**[*]
Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, Furu Wei[◇]
Microsoft Research
https://aka.ms/GeneralAI

This report presents VIBEVOICE, a novel model designed to synthesize long-form speech with multiple speakers by employing the next-token diffusion framework [SBW+24]—a unified method for modeling continuous data by autoregressively generating latent vectors via diffusion. To enable this, we introduce a novel continuous speech tokenizer that, when compared to the popular Encodec model, improves data compression by 80 times while maintaining comparable performance. This tokenizer effectively preserves audio fidelity while significantly boosting computational efficiency for processing long sequences. Thus, VIBEVOICE can synthesize long-form speech for up to 90 minutes (in a 64K context window length) with a maximum of 4 speakers, capturing the authentic conversational "vibe" and surpassing open-source and proprietary dialogue models.

Project Page: aka.ms/VibeVoice
Code: github.com/microsoft/VibeVoice
Hugging Face: microsoft/VibeVoice
Demo: aka.ms/VibeVoice-Demo

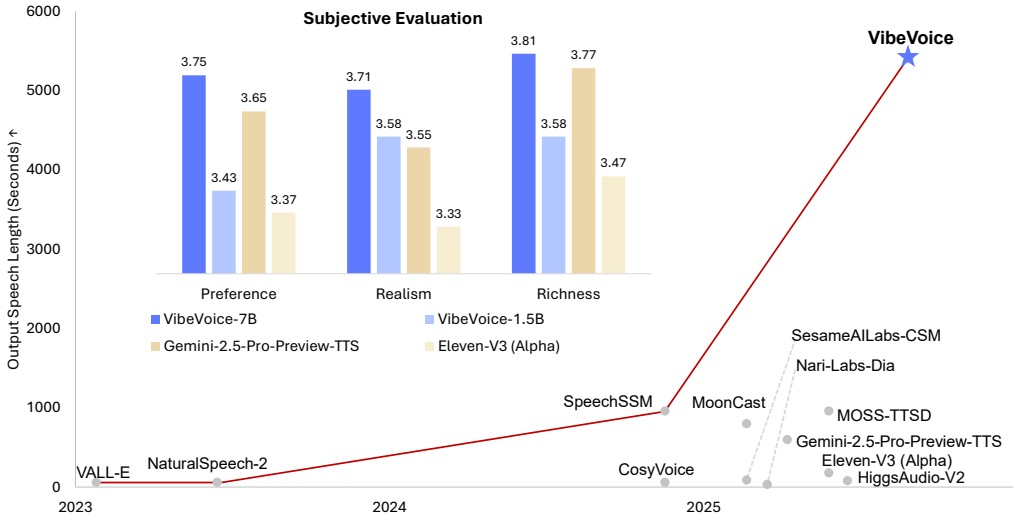

Figure 1: VIBEVOICE is capable of synthesizing 5,000+ seconds of audio while consistently outperforming strong open/closed-source systems in subjective evaluations of preference, realism, and richness.

[*] Core contributors. ◇ Contact person: fuwei@microsoft.com.

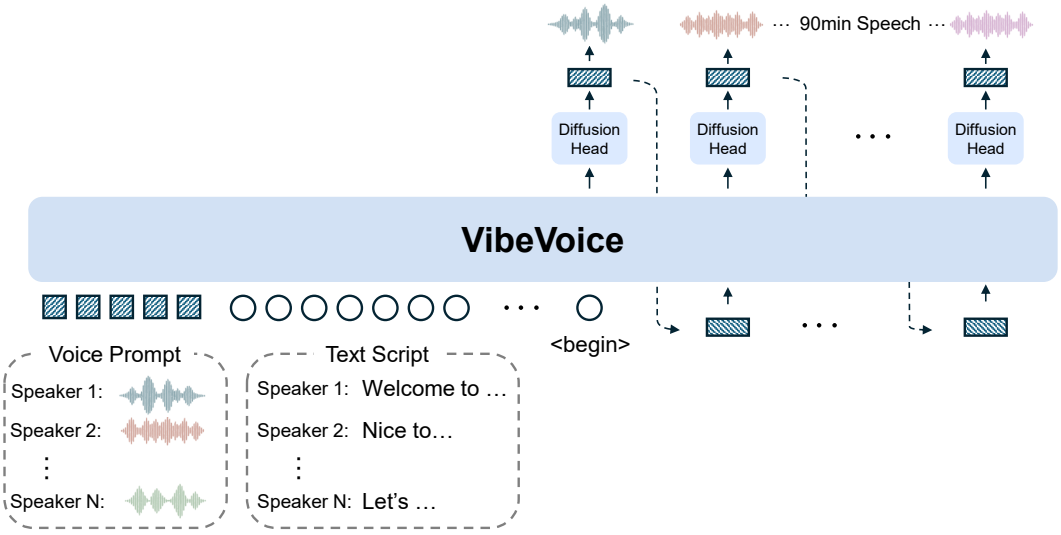

Figure 2: VIBEVOICE employs next token diffusion framework as in LatentLM [SBW+24] to synthesize long-form and multi-speaker audios. Voice prompts and text scripts provide initial input. VIBEVOICE processes hybrid context features, and its hidden states condition a token level Diffusion Head (D), which predicts acoustic VAE for speech segments, subsequently recovered by acoustic decoder (A).

1 Introduction

While recent advancements in Text-to-Speech (TTS) synthesis have achieved remarkable success in generating high-fidelity, natural-sounding speech for single speakers in short utterances [WCW+23, ACC+24a, LVS+23, CNM+24, DWC+24a, JCC+25, YZC+25], a significant frontier remains in the scalable synthesis of long-form, multi-speaker conversational audio, such as podcasts and multi-participant audiobooks. Although traditional systems can technically produce such audio by concatenating individually synthesized utterances, achieving natural turn-taking and content-aware generation are major challenges. Recently, research on multi-speaker long conversational speech generation has begun to emerge [Goo24, PSJ+24, Nar25, Ope25, Ses25]. However, most of these works are either not open-sourced [Goo24, PSJ+24] or still face challenges in terms of generation length and stability [Nar25, Ses25, JYY+25, Ope25].

In this work, we introduce VIBEVOICE, as illustrated in Figure 2, a novel framework developed for the scalable synthesis of long-form and multi-speaker speech. To support long audio generation, we have pioneered the development of a causal speech tokenizer that achieves a 3200X compression rate (i.e., 7.5 Hz frame rate). In our experiments, this highly efficient tokenizer maintains a speech-to-text token ratio of approximately 2:1, meaning two speech tokens are roughly equivalent to one BPE [SHB15] text token.

We utilize a pre-trained Large Language Model (LLM, e.g., Qwen2.5 [YYZ+24]) to interpret complex user inputs, including detailed text sentences and role assignments. We have streamlined the architecture by removing unnecessary prior designs: voice latent features and text scripts are concatenated into a single sequence and fed directly into the LLM. The LLM then processes this context to predict a hidden state, which in turn conditions a lightweight, token-level Diffusion Head [LTL+24]. This diffusion head is responsible for predicting the continuous Variational Autoencoder (VAE) features, which are subsequently recovered into the final audio output by speech tokenizer decoder.

Despite its architectural simplicity, VIBEVOICE yields an exceptionally powerful TTS model. It demonstrates remarkable flexibility in handling multiple speakers and achieves a synthesis length of up to 90 minutes. Scaling the LLM from 1.5B to 7B, the larger model exhibits significant gains in perceptual quality, delivering richer timbre, more natural intonation, and enhanced transfer capabilities, such as in cross-lingual applications.

2 Method

2.1 Speech Tokenizers

We employ two separate tokenizers as input to learn both acoustic and semantic features. In our experiments, generating long-form speech benefits from this separate design.

Acoustic Tokenizer adopts the principles of a Variational Autoencoder (VAE) [KW14], specifically drawing inspiration from the σ-VAE variant proposed in LatentLM [SBW+24] to mitigate potential variance collapse issues of VAEs when used in autoregressive modeling settings. The process involves an encoder network, parameterized by ϕ, which maps the input audio $\boldsymbol{x}$ to the parameters of a latent distribution, primarily the mean μ. Notably, variance σ is a pre-defined distribution ($\mathcal{N}(0, C_\sigma)$) in σ-VAE, rather than a learnable distribution in VAE [KW14]. A latent vector $\boldsymbol{z}$ is then sampled using the reparameterization trick. Following the σ-VAE approach to ensure robust variance for autoregressive modeling, we can formulate this as: $\boldsymbol{z} = \mu + \sigma \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1),\ \sigma \sim \mathcal{N}(0, C_\sigma)$.

The architecture is a mirror-symmetric encoder-decoder structure. The encoder employs a hierarchical design with 7 stages of modified Transformer blocks [VSP+17] (using 1D depth-wise causal convolutions instead of self-attention module) for efficient streaming processing. Six downsampling layers achieve a cumulative 3200X downsampling rate from a 24kHz input, yielding 7.5 tokens/frames per second. Each encoder/decoder component has approximately 340M parameters. The training objective follows the DAC [KSL+23], including its discriminator and loss designs.

Semantic Tokenizer mirrors the hierarchical architecture of the Acoustic Tokenizer's encoder, but without VAE components, as its objective is deterministic content-centric feature extraction. The main difference is the training objective, which uses Automatic Speech Recognition (ASR) as the proxy task. During training, its output is decoded by several Transformer decoder layers to predict text transcripts, aligning the semantic encoder's representations with textual semantics. This decoder is discarded after pre-training.

2.2 VIBEVOICE

VIBEVOICE employs a Large Language Model (LLM) as its core sequence model, integrated with specialized audio encoding and diffusion-based decoding modules to achieve scalable, high-fidelity multi-speaker speech synthesis. The overall inference architecture is depicted in Figure 2.

Input Representation: The model input X is formed by concatenating the voice font features and the text script embeddings, specified by users, interleaved with role identifiers ($\texttt{Speaker}_k$): $X = [\texttt{Speaker}_1 : \boldsymbol{z}_1, \texttt{Speaker}_2 : \boldsymbol{z}_2, ..., \texttt{Speaker}_N : \boldsymbol{z}_N] + [\texttt{Speaker}_1 : T_1, \texttt{Speaker}_2 : T_2, ..., \texttt{Speaker}_N : T_N]$, where $\boldsymbol{z}_N$ is projected by Eq **??** and T_N is each role's text scripts. For the generated speech segment $\boldsymbol{s}$, it will be encoded by acoustic tokenizer and semantic tokenizer to form the hybrid speech representation for the auto-regressive modeling.

Token-Level Diffusion: To synthesize speech in a streaming way, VIBEVOICE employs a lightweight diffusion head [LTL+24] conditioned on the LLM's hidden state of each token, $\boldsymbol{h}_i$. During training, this diffusion head is optimized to reverse a forward noising process by predicting the noise [HJA20] added to the clean acoustic VAE features $z_{a,i}$. During inference, this diffusion head iteratively refines a randomly sampled Gaussian noise vector to predict the target acoustic VAE feature, $\boldsymbol{z}_{a,i}$. This denoising process is enhanced using Classifier-Free Guidance (CFG), which interpolates between a conditional prediction (guided by $\boldsymbol{h}_i$) and an unconditional prediction. An efficient sampler, such as DPM-Solver++ [LZB+22, LZB+25], is utilized to accelerate this iterative process, ultimately yielding a clean acoustic feature estimate.

We instantiated VIBEVOICE's core LLM using the 1.5B and 7B parameter versions of Qwen2.5 [YYZ+24]. The diffusion head [LTL+24] comprises 4 layers. During VIBEVOICE training, the pre-trained acoustic and semantic tokenizers remained frozen, with only the LLM and diffusion head parameters being learnable. We employed a curriculum learning strategy for the LLM input sequence length, progressively increasing from 4,096 to 65,536 tokens. The guidance scale is 1.3 and the iterative denoising step is 10 for VIBEVOICE.

Model	Subjective				Objective		
	Realism	Richness	Preference	Average	WER (Whisper)	WER (Nemo)	SIM-O
Nari Labs Dia [Nar25]	-	-	-	-	11.96	10.79	0.541
Mooncast [JYY+25]	-	-	-	-	2.81	3.29	0.562
SesameAILabs-CSM [Ses25]	2.89 ±1.15	3.03 ±1.11	2.75 ±1.08	2.89 ±1.12	2.66	3.05	0.685
Higgs Audio V2 [Bos25]	2.95 ±1.13	3.19 ±1.06	2.83 ±1.16	2.99 ±1.13	5.94	5.97	0.543
Elevenlabs v3 alpha [Ele]	3.34 ±1.11	3.48 ±1.05	3.38 ±1.12	3.40 ±1.09	2.39	2.47	0.623
Gemini 2.5 pro preview tts [Goo]	3.55 ±1.20	3.78 ±1.11	3.65 ±1.15	3.66 ±1.16	1.73	2.43	-
VIBEVOICE-1.5B	3.59 ±0.95	3.59 ±1.01	3.44 ±0.92	3.54 ±0.96	**1.11**	**1.82**	0.548
VIBEVOICE-7B	**3.71** ±0.98	**3.81** ±0.87	**3.75** ±0.94	**3.76** ±0.93	1.29	1.95	**0.692**

Table 1: Human subjective and objective evaluation results. For all subjective metrics and SIM-O, higher scores are better. For WER, lower scores are better. Best results are in **bold**.

3 Results

3.1 VIBEVOICE Podcast

We conducted both objective and subjective evaluations to benchmark the performance of the proposed VIBEVOICE against recent state-of-the-art conversational speech generation systems [Nar25, JYY+25, Ses25, Bos25, Ele, Goo].

To manage the labor-intensive and time-consuming nature of subjective evaluation, we designed a compact test set. This set consists of 8 long conversational transcripts with a total duration of about 1 hour. We used speech prompts to ensure consistent timbre across the different models. Since Gemini 2.5 Pro preview TTS does not support speech-prompt control, we used its default male and female voices for comparison instead.

For our objective evaluation, we measure Word Error Rate (WER) and speaker similarity. WER is obtained by transcribing the generated speech using Whisper-large-V3 [RKX+23] and Nemo ASR [XJM+23]. Speaker similarity (SIM-O) is computed by extracting speaker embeddings with WavLM-large [CWC+22a].

For subjective evaluation, we recruited 24 human annotators to provide Mean Opinion Scores (MOS) across three dimensions: **Realism** (how natural and human-like the speech sounds, including prosody, emotion, and the smoothness of speaker turns), **Richness** (the expressiveness of the speech in terms of tone and emotion, including variation and adaptation to context), and **Preference** (overall listener enjoyment and subjective preference, reflecting naturalness, pleasantness, and engagement). The evaluation covered six models with all eight test samples, meaning that each annotator listened to approximately six hours of audio in total.

We can observe that: **The proposed VIBEVOICE models outperform all other top-tier models on long conversational speech generation across both objective and subjective metrics.** Compared with the VIBEVOICE-1.5B model, the VIBEVOICE-7B model achieves significantly better performance on all objective metrics and SIM-O, while maintaining a comparable WER.

3.2 VIBEVOICE Short Utterance

We evaluate VIBEVOICE on the SEED test sets [ACC+24b], a widely used benchmark composed of short utterances. For evaluation, approximately 1,000 English samples and 2,000 Chinese samples are drawn from the CommonVoice dataset, denoted as *test-en* and *test-zh*, respectively. We compute word error rate (WER) using Whisper-large-v3 for *test-en* and Paraformer [GZMY22] for *test-zh*. For speaker similarity (SIM), we adopt a WavLM-large [CWC+22b] model fine-tuned on the speaker verification task.

Table 2 presents the results on the SEED test sets. Although our model is primarily trained on long-form speech, it demonstrates strong generalization on short-utterance benchmarks. In addition, by employing a lower frame rate, our model substantially reduces the number of decoding steps required to synthesize one second of speech.

Model	Frame Rate	test-zh		test-en	
		CER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑
MaskGCT [WZL+24]	50	2.27	0.774	2.62	0.714
Seed-TTS [ACC+24b]	-	1.12	0.796	2.25	0.762
FireRedTTS [GLS+24]	25	1.51	0.635	3.82	0.460
CosyVoice 2 [DWC+24b]	25	1.45	0.748	2.57	0.652
Spark TTS [WJM+25]	50	1.20	0.672	1.98	0.584
VIBEVOICE-1.5B	7.5	1.16	0.744	3.04	0.689

Table 2: Results on the SEED test sets.

Tokenizer	N_q	Token Rate	test-clean			test-other		
			PESQ	STOI	UTMOS	PESQ	STOI	UTMOS
Ground-Truth	-	-	-	-	4.056	-	-	3.483
Encodec [DCSA22]	8	600	2.72	**0.939**	3.04	2.682	**0.924**	2.657
DAC [KSL+23]	4	400	2.738	0.928	3.433	2.595	0.908	2.945
Encodec [DCSA22]	4	300	2.052	0.901	2.307	2.052	0.884	2.088
SpeechTokenizer [ZZL+23]	4	300	1.931	0.878	3.563	1.737	0.837	3.018
DAC [KSL+23]	1	100	1.246	0.771	1.494	1.245	0.751	1.499
WavTokenizer [JJW+25]	1	75	2.373	0.914	4.049	2.261	0.891	3.431
WavTokenizer [JJW+25]	1	40	1.703	0.862	3.602	1.662	0.834	3.055
Ours (Acoustic)	1	7.5	**3.068**	0.828	**4.181**	**2.848**	0.823	**3.724**

Table 3: Objective evaluation of reconstruction quality on the LibriTTS test-clean and test-other datasets. N_q denotes the number of quantizers (VAE for us). Token Rate indicates the number of tokens/frames generated per second of audio. Higher PESQ, STOI, and UTMOS scores indicate better performance. Best results are in **bold**.

3.3 Tokenizer Reconstruction

The fidelity of audio reconstructed from acoustic tokens is a critical indicator of the tokenizer's efficacy in preserving essential acoustic information, particularly under high compression rates. To quantify this, we measured PESQ [RBHH01], STOI [THHJ10] and UTMOS [SXN+22] on both the LibriTTS test-clean and test-other datasets [ZDC+19]. Table 3 shows that our acoustic tokenizer, uniquely operating at an ultra-low 7.5 Hz, achieves leading PESQ and UTMOS scores on both test-clean (PESQ: 3.068, UTMOS: 4.181) and test-other (PESQ: 2.848, UTMOS: 3.724) subsets. This demonstrates its capacity for high-fidelity, perceptually excellent audio reconstruction despite aggressive compression, which is a key factor for VIBEVOICE's scalability with long-form audio.

4 Conclusion, Limitations, and Risks

We introduced VIBEVOICE, a novel framework for long-form and multi-speaker speech generation. By integrating efficient hybrid speech representations from specialized ultra-low frame rate (7.5 Hz) acoustic and semantic tokenizers with an end-to-end LLM-based next-token diffusion framework, VIBEVOICE achieves state-of-the-art performance. It scalably synthesizes high-quality audio for up to 90 minutes with up to 4 speakers, demonstrably surpassing existing baselines in both subjective perceptual quality—including preference, realism, and richness—and objective metrics like WER, thereby significantly advancing the capabilities of conversational TTS.

English and Chinese only: Transcripts in languages other than English or Chinese may result in unexpected audio outputs.

Non-Speech Audio: The model focuses solely on speech synthesis and does not handle background noise, music, or other sound effects.

Overlapping Speech: The current model does not explicitly model or generate overlapping speech segments in conversations.

Potential for Deepfakes and Disinformation: High-quality synthetic speech can be misused to create convincing fake audio content for impersonation, fraud, or spreading disinformation. Users

must ensure transcripts are reliable, check content accuracy, and avoid using generated content in misleading ways.

We do not recommend using VIBEVOICE in commercial or real-world applications without further testing and development. This model is intended for research and development purposes only. Please use responsibly.

References

[ACC+24a] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[ACC+24b] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[Bos25] Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio, 2025.

[CNM+24] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[CWC+22a] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[CWC+22b] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[DCSA22] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[DWC+24a] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[DWC+24b] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[Ele] Elevenlabs. Elevenlabs v3 alpha. https://elevenlabs.io/docs/models#eleven-v3-alpha.

[GLS+24] Haohan Guo, Kun Liu, Feiyu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kaituo Xu. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *CoRR*, abs/2409.03283, 2024.

[Goo] Google. Gemini 2.5 Pro Preview TTS. https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro-preview-tts.

[Goo24] Google. NotebookLM. https://notebooklm.google/, 2024.

[GZMY22] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech*, pages 2063–2067. ISCA, 2022.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[JCC+25] Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

[JJW+25] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

[JYY+25] Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, Zhengtao Wang, Xu Tan, Xinyu Zhou, Tao Qin, and Xiangyang Li. Mooncast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*, 2025.

[KSL+23] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 27980–27993, 2023.

[KW14] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*, 2014.

[LTL+24] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[LVS+23] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[LZB+22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[LZB+25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025.

[Nar25] Nari Labs. Nari Labs Dia. https://github.com/nari-labs/dia, 2025.

[Ope25] OpenMOSS Team. MOSS-TTSD. https://github.com/OpenMOSS/MOSS-TTSD, 2025.

[PSJ+24] Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan. Long-form speech generation with spoken language models. *arXiv preprint arXiv:2412.18603*, 2024.

[RBHH01] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

[RKX+23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[SBW+24] Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024.

[Ses25] SesameAILabs. SesameAILabs CSM Model. https://github.com/SesameAILabs/csm, 2025.

[SHB15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[SXN+22] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

[THHJ10] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017.

[WCW+23] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023.

[WJM+25] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

[WZL+24] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *CoRR*, abs/2409.00750, 2024.

[XJM+23] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning*, pages 38462–38484. PMLR, 2023.

[YYZ+24] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[YZC+25] Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi DAI, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025.

[ZDC+19] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

[ZZL+23] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023.