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/ICASSP49660.2025.10890838

MARS6: A Small and Robust Hierarchical-Codec Text-to-Speech Model

Matthew Baas*, Pieter Scholtz*, Arnav Mehta*, Elliott Dyson*, Akshat Prakash*, Herman Kamper*†
*Camb.ai
Email: research@camb.ai

***Abstract*—Codec-based text-to-speech (TTS) models have shown impressive quality with zero-shot voice cloning abilities. However, they often struggle with more expressive references or complex text inputs. We present MARS6, a robust encoder-decoder transformer for rapid, expressive TTS. MARS6 is built on recent improvements in spoken language modelling. Utilizing a hierarchical setup for its decoder, new speech tokens are processed at a rate of only 12 Hz, enabling efficient modelling of long-form text while retaining reconstruction quality. We combine several recent training and inference techniques to reduce repetitive generation and improve output stability and quality. This enables the 70M-parameter MARS6 to achieve similar performance to models many times larger. We show this in objective and subjective evaluations, comparing TTS output quality and reference speaker cloning ability. Project page: https://camb-ai.github.io/mars6-turbo/**

***Index Terms*—text-to-speech, speech synthesis, voice cloning**

I. Introduction

Text-to-speech (TTS) systems have improved many-fold in recent years, showcasing new capabilities in speaker cloning cability and naturalness [1]–[3]. One promising area in TTS is spoken language models (SLMs) [4], where a neural audio codec converts speech into a sequence of discrete tokens. Like text language models, SLMs are trained to predict the next discrete token autoregressively, typically using a transformer-based architecture. But most prior SLM-based TTS systems exhibit a key limitation – they are unstable [5], [6]. When the reference audio or text is complex or out-of-domain, SLMs often perform poorly compared other TTS methodologies.

While there have been several methods proposed to address such limitations, they are typically considered in isolation (e.g. repetition aware sampling [2]), or they drastically increase the runtime (e.g. multiple sampling [2], [7]). To this end, we propose MARS6 – a 70M parameter SLM for robust, rapid and expressive TTS. We combine several recent techniques, and propose some new techniques from outside the TTS domain (e.g. odds ratio preference optimization [8] and a new top-p fallback sampling mechanism). MARS6 consists of an encoder-decoder transformer, and combines a hierarchical speech codec with a hierarchical decoder architecture to process speech tokens at a rate of 12 Hz. Together with the aforementioned inference techniques, this makes MARS6 a highly robust and capable TTS model. It is also a showcase for a 'bag of tricks' that we introduce for SLM-based TTS.

†This author is with E&E Engineering, Stellenbosch University, South Africa. All contributions were made in their capacity as an advisor to Camb.ai Inc.

For our experiments, we construct a difficult in-the-wild TTS evaluation set using the expressive EARS dataset [9]. We compare MARS6 against prior diffusion- and autoregressive-based TTS models using objective and subjective evaluations. MARS6 performs competitively, even against models many times its size. When used with voice cloning based on a snippet of reference audio, MARS6 captures the target speaker identity closely, surpassing prior models in subjective speaker similarity evaluations. Our main contribution is to demonstrate that we can combine several recently proposed techniques with new techniques proposed herein during model design, training, and inference, to stabilize outputs and yield a more robust SLM-based TTS system. Demo, samples, code, and checkpoints: https://camb-ai.github.io/mars6-turbo/.

II. Related Work

Within SLMs, there are broadly three ways to approach speech tokenization. The first is to tokenize speech using acoustic tokens at a fixed sample rate, as done in EnCodec and DAC [10], [11]. The second is to mix acoustic and semantic tokens using two different quantizers [12], e.g. using clustered HuBERT features for semantic and EnCodec for acoustic tokens. The third, which we explore here, is that of hierarchical acoustic codecs, such as SNAC [13]. These codecs quantize speech into acoustic tokens in different codebooks, each with its own sampling rate. This makes lower codebooks more 'coarse', and higher sample-rate codebooks 'fine'. The progenitor SLM TTS model, VALL-E, and its successors [2], [4], [6], uses an autoregressive transformer to predict the most coarse acoustic codebook, and a non-autoregressive model to predict the remaining codebook values.

Despite success, VALL-E and its descendants often suffer from stability issues. Several studies have tried to address this [14], [15], e.g. by adding linguistic and phonemic constraints to improve coherence between the output speech and the given input text [16]. But most of these improvements require phoneme alignments during training. The 'bag-of-tricks' we introduce in this paper does not require such resources.

III. MARS6

Fig. 1 shows the MARS6 model, which follows an encoder-decoder architecture. For zero-shot speaker cloning, the encoder takes in reference speaker embeddings together with the target text. The decoder is hierarchical and made of two components: a local and global decoder, similar to the proposal of [17]. The

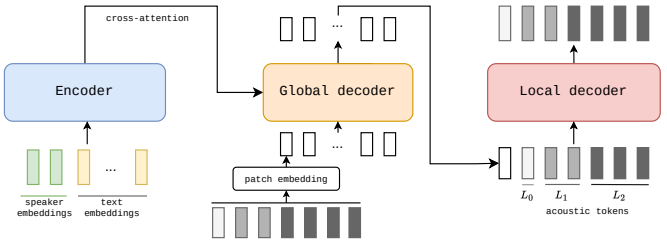

Fig. 1. MARS6 is an encoder-decoder transformer. The encoder converts a speaker embedding and sequence of text embeddings to latent vectors for cross-attention in the global decoder. The hierarchical autoregressive decoder has two parts: The global decoder produces new latent vectors at a low sample rate, where each vector is autoregressively decoded to acoustic tokens using a smaller local decoder model. The entire patch of acoustic tokens then forms the next input vector to the global decoder through a patch embedding.

global decoder takes input acoustic features in patches, and its output is fed into the local decoder to autoregressively predict all acoustic tokens for the next patch. Details are given next.[1]

A. Encoder and input representation

The encoder is a non-causal transformer encoder using Mish activations [18] with sinusoidal positional embeddings, similar to [19]. Its input sequence consists of two parts. First, to clone the target speaker, we compute a speaker embedding using a pretrained speaker verification model and a secondary embedding using CLAP [20]. The former, being trained mostly on non-emotive speech, gives a good base speaker representation. But, for expressive references where the speaker verifier's embeddings are less meaningful, the more broadly trained (but less speaker-specific) CLAP embedding is useful. These two vectors are mapped to the dimension of the transformer using a projection layer, and then joined along the sequence length ('speaker embeddings' in Fig. 1). Second is the sequence of text embeddings corresponding to the desired text being spoken ('text embeddings' in Fig. 1). To reduce the token count and improve speed, the text is tokenized using byte-pair encoding (BPE) [21].

To improve reference coherence and output stability, we adapt a lesson from [22]. We give the encoder a way to learn when an output should be high fidelity (e.g. 48 kHz audio from VCTK [23] downsampled to the 24 kHz codec sampling rate) or lower fidelity (e.g. upsampled 16 kHz audiobook data). To indicate the target quality to the encoder, we prepend the original sample rate to the text, e.g. for 16 kHz, "Mister ..." becomes "[16000] Mister ...".

B. Global decoder

MARS6 operates on hierarchical acoustic tokens from the SNAC acoustic model [13]. SNAC encodes speech into discrete sequences using residual vector quantization with codebooks at different sampling rates, representing different levels in a hierarchy, where earlier codebooks are sampled less frequently. For MARS6 we use the 3-codebook SNAC [13], with codebook sample rates of 12 (L_0), 24 (L_1), and 48 Hz (L_2).

[1]Mars is the Roman god of war. It is also the name of a chocolate bar first produced in 1932. MARS6 was our sixth internal model version.

Like the encoder, this decoder uses Mish activations and sinusoidal positional embeddings. The global decoder takes patches of acoustic tokens from SNAC at 12 Hz, whereby all codebook tokens generated within $\frac{1}{12}$s are flattened and fed through a patch embedding [17] to yield a 12 Hz input vector sequence as shown in Fig. 1. This corresponds to a patch size of seven, since for every $\frac{1}{12}$s, there is one token from the 12 Hz L_0 codebook, two from the 24 Hz L_1 codebook, and four from the 48 Hz L_2 codebook.

C. Local decoder

The global decoder's output must be converted to the full hierarchical codec tokens to vocode the output speech. Each output vector from the global decoder is fed as the first input vector to the local decoder. As shown in Fig. 1, the local decoder then autoregressively predicts each codec token for all codebooks for the current patch in a flattened way, predicting L_0, then two L_1 tokens, then the last four L_2 codebook tokens.

The local decoder is also a causal autoregressive transformer. But unlike the encoder and global decoder, it always operates on a fixed sequence length of seven. So we use fixed, learnt positional embeddings instead of sinusoidal embeddings.

D. Training

The model is trained end-to-end with a standard cross-entropy loss to predict the next acoustic token. Speaker embeddings are computed from the ground truth audio during training, while during inference they are computed from a desired reference speaker. The local decoder is applied in parallel to the global decoder outputs during training and autoregressively during inference. During training, an end-of-sequence token is appended to the acoustic tokens of the utterance, which the local encoder is trained to predict.

IV. Inference and Fine-Tuning Techniques

MARS6 is fast and small because most of its parameters operatore on only a 12 Hz sequence in the global decoder. The shorter sequence can also improve stability. But on its own, this new architecture does not solve the SLM-robustness problem. Below we introduce and incorporate a 'bag of tricks' for inference and fine-tuning to improve stability and performance.

A. Fine-tuning setup

We split model training into two parts: pretraining and fine-tuning. Pretraining involves next-token prediction, as described earlier. We then fine-tune the model using a curated high-quality subset of the training data.

For fine-tuning, we combine odds ratio preference optimization (ORPO) [8] and reverse inference optimization (RIO) [5]. First, we compute the pretraining model predictions on arbitrary text using reference waveforms from a high quality subset of the training data. We then feed these outputs back to MARS6 as references, with the transcript of the original reference, and predict the original reference audio in a cyclic way, as in [5]. We then rank the cyclic outputs based on character error rate and UTMOS [24], and select the worst performing outputs as

'rejected' samples, and the corresponding ground truth audio as 'chosen' samples for ORPO. While not precisely the same as either the original ORPO (where both chosen and rejected samples come from model predictions) or RIO (where both the best and worst-performing cyclic outputs are used), we found this setup to yield the best results in preliminary experiments.

We also found that the model had a tendency to get stuck producing the same acoustic token – this is why prior work incorporate semantic tokens in addition to acoustic tokens [12]. To remedy this, we incorporate a flux loss to penalize repetitive generations [25]. We adapt the flux loss used for the continuous autoregressive TTS [25] to discrete units, defining it as:

$$\mathcal{L}_{\text{flux}} = \frac{\beta}{\epsilon + \text{CrossEntropy}(\hat{\mathbf{y}}_t, y_{t-1})} \tag{1}$$

where β is a scaling coefficient for the loss term, ϵ is a small offset added for numerical stability, $\hat{\mathbf{y}}_t$ is the decoder logit predictions at timestep t, and y_{t-1} is the ground truth codebook index of the *prior timestep*. Intuitively, this penalizes the probability of the token in the prior timestep. We apply this flux loss to L_0 codebook predictions, during both ORPO fine-tuning and pretraining, each with different weightings.

B. Inference algorithms

We combine three inference methods.

1) Repetition aware sampling (RAS): This approach from [2] is used on the local decoder predictions for positions corresponding to the L_0. Using the notation of the original paper, we found $K = 10, t_r = 0.09$ to yield best results.

2) Quality prefixing: As mentioned in Sec. III-A, in training we prepend the original sample rate of the reference to the text to give the model am indication for output quality. In inference, we always set this to "[48000]" to maximize output quality.

3) Top-p backoff sampling: SLM outputs can be made more stable by sampling with a low top-p value. However, sometimes this can cause the model to still get stuck in a loop. We alleviate this by using a backoff approach similar to the temperature backoff used by Whisper [26]. Concretely, we sample with a top-p of 0.2, and check the output length before vocoding. If the predicted audio is unrealistically short, we increment the top-p by 0.2 and sample again.

C. Shallow and deep cloning

MARS6 can clone from a reference in two ways – *shallow clone* and *deep clone*. The prior is where we compute the speaker embeddings from the reference audio and perform inference directly. While simple, the speaker similarity is not optimal. The latter is similar to the approach of VALL-E, where we assume knowledge of the reference transcript, and then assign a prefix to both the encoder and global decoder as the reference transcript and acoustic tokens, respectively. This gives better prosody and speaker transfer from the reference, at the cost of inference time (longer sequence length).

V. Experimental Setup

A. Evaluation data and baselines

Many evaluation benchmarks do not capture the diversity of in-the-wild speech. We therefore construct a new evaluation set on the emotive EARS dataset [9]. It includes emotional speech, different reading styles, free-form conversational speech, and non-verbal sounds recorded in an anechoic environment from 107 English speakers. We select 43 speakers for the test set and 64 for the validation set. Ignoring the non-verbal, free-form and 'slow' utterances, we select half of the samples (audio and transcript) for each style, and pair each sample with another of the same speaker and style to serve as the voice cloning reference. MARS6 and the baseline models have, to the best of our knowledge, not seen any part of EARS.

We compare the 70M-parameter MARS6 against three strong baseline models, all much larger: XTTSv2 [1] (460M parameters), StyleTTS2 [3] (148M parameters), and MetaVoice-1B [27] (1.2B parameters). We use the best available checkpoints and the best inference settings from each paper.

B. MARS6 implementation

1) Model: We use standard 8-layer, 512-dimensional transformers for the encoder and global decoder, and a 4-layer local decoder. For the two speaker embeddings, we use WavLM-SV [28] and the pretrained MS-CLAP [20]. We train the BPE tokenizer to a vocabulary size of 512.

2) Training: We train MARS6 for 2M steps using AdamW [29] with a linearly decaying learning rate starting at $5 \cdot 10^{-4}$ (after a 10k step linear warmup) and ending at $2.5 \cdot 10^{-5}$. We use an AdamW β of $(0.9, 0.995)$, weight decay of $2 \cdot 10^{-2}$, and batch size of 96.

3) Data: We train MARS6 on the following publically available datasets: LibriHeavy [30], GLOBE [31], VCTK [23], AniSpeech [32], and CCv2 [33]. We limit the number of utterances from each speaker to be at most 80k. Together this results in a training dataset of roughly 46k hours.

C. Evaluation metrics

1) Objective evaluation: We measure intelligibility using the word/character error rate (W/CER) between the predicted outputs on our EARS test set and the ground truth audio. We obtain transcripts of the generated audio using the Whisper-base speech recognition model [26]. We objectively measure speaker cloning ability using the equal-error rate (EER) for a pretrained speaker verification system [34]. The verification system produces a similarity score between pairs of utterances. We compute these similarities on *(ground truth reference, generated)* pairs and *(ground truth reference, other ground truth)* pairs from the same speaker. The former pairs are assigned a label of 0, and latter a label of 1. Thsese can then be used to compute an EER as in [35]. A higher EER indicates that it is harder to distinguish generated speech from ground truth examples of the reference speaker, up to a theoretical maximum of 50%. We also report an approximated mean naturalness metric using the pretrained UTMOS model [24] predicting naturalness scores on a scale of 1-5.

TABLE I
RESULTS MEASURING THE INTELLIGIBILITY (W/CER), NATURALNESS (UTMOS, MOS) AND SPEAKER SIMILARITY (EER, SIM) ON OUR EARS TEST SET. FOR MOS AND SIM, 95% CONFIDENCE INTERVALS ARE SHOWN.

Model	WER ↓	CER ↓	EER ↑	UTMOS ↑	MOS ↑	SIM ↑
Testset topline	5.74	2.50	-	3.50	3.34 ± 0.11	3.46 ± 0.08
XTTSv2 [36]	1.74	0.83	29.4	3.81	3.58 ± 0.08	2.24 ± 0.11
MetaVoice-1B [27]	30.70	27.41	**31.2**	3.13	2.84 ± 0.11	2.47 ± 0.11
StyleTTS2 [3]	**1.34**	**0.36**	23.1	**4.40**	$\mathbf{4.08 \pm 0.07}$	2.80 ± 0.12
MARS6 (deep)	7.42	5.17	30.7	3.79	3.34 ± 0.10	$\mathbf{3.07 \pm 0.11}$
MARS6 (shallow)	3.96	2.38	23.1	3.65	3.44 ± 0.08	2.24 ± 0.11
w/o RIO ORPO [8]	14.54	12.92	22.7	3.60	—	—
w/o RAS [2]	7.31	5.73	24.0	3.76	—	—
w/o quality prefixing	7.06	4.95	26.1	3.56	—	—

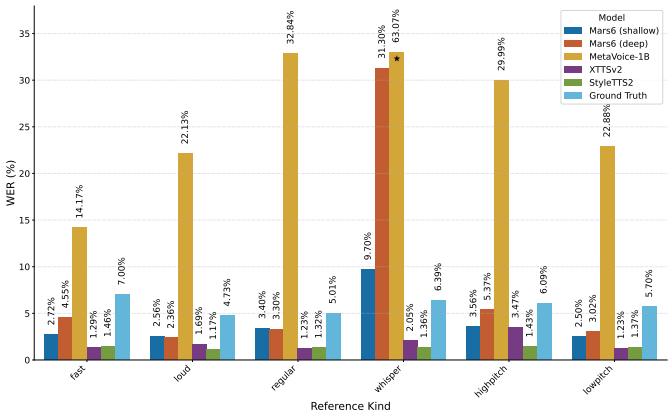

Fig. 2. Comparison of word error rates for different speaker reference styles.

2) Subjective evaluation: We perform two subjective evaluations using Amazon Mechanical Turk. In the first, we collect a mean opinion score (MOS) on a 1-5 scale. In the second, we collect a speaker similarity score (SIM) on a 1-4 scale following the protocol of the Voice Conversion Challenge 2020 [35]. From the EARS test set, we select 36 utterances from each baseline, the ground truth, and MARS6 (both using shallow and deep clone). We include trapping and calibration samples to filter out anomalous listeners, resulting in 1326 ratings from 2340 unique listeners. For SIM, each evaluated utterance (from the baselines, MARS6, or actual ground truth audio) is paired with another random utterance from the same speaker and speaking style. We present the listener these samples side-by-side and ask them to rate how similar the speaker sounds on a 1-4 scale similar to [35]. After filtering anomalous listeners, we have 1980 SIM ratings from 40 unique listeners.

VI. RESULTS

A. Intelligibility and reference similarity

The results on the EARS test set are given in Table I. Results are mixed: for intelligibility, StyleTTS is a clear winner. In terms of speaker similarity, MARS6 using deep clone has the best SIM score, but in terms of EER, MetaVoice-1B is best. For naturalness (MOS and UTMOS), StyleTTS2 again is the best. But these results are perhaps a bit misleading, as can be seen by both XTTS, StyleTTS, and MARS6 having better W/CER and UTMOS values than the ground truth test utterances.

While this requires further investigation, the audio samples on the demo give some insight. Because the EARS is emotive, spontaneous, and diverse, it is less intelligible than pure read speech. Models like StyleTTS2 and XTTSv2 appear to produce audio that is 'de-emphasized' compared to that of the reference, particularly for highly emotive references. Meanwhile, SLM-based models like MetaVoice and MARS6 appear to clone the prosody of the reference more strongly at the cost of intelligibility, indicated by the higher speaker similarity metrics (especially for deep clone). This effect is clearly heard when a whispered reference is used, where StyleTTS2 and XTTSv2 produce clean sounding outputs that are not whispered, while MARS6 correctly produces a whispered output, even if it is slightly less intelligible (higher W/CER). So, for highly expressive speech, lower W/CER numbers do not always correspond to outputs that are faithful to the reference utterance.

We ablate the RAS, quality prefixing (Sec. IV-B) and RIO ORPO fine-tuning (Sec. IV-A) in the last three rows of Table I by measuring the model's shallow clone performance i.t.o objective metrics. Removing any of the individual techniques degrades intelligibility. Speaker similarity is also worse when removing RIO ORPO. This shows that each technique is important for MARS6.

B. Effect of reference style and cloning method

To demonstrate this effect a bit more, as well as profile the cases where MARS6 is making intelligibility errors, we make use of the style labels in EARS. Using these labels we plot the WER metric grouped by the style of the reference utterance in Fig. 2. The trends for most styles appear constant, except for one reference style – whispers. Most of the W/CER in Table I from both MetaVoice and MARS6 are attributed to whispered outputs! This, together with the audio samples, provides evidence for our earlier hypothesis. MARS6 is able to produce coherent whisper outputs, however, Whisper-base cannot accurately transcribe whispers. This also causes the poorly-cloned outputs of XTTSv2 and StyleTTS2 to be rated much higher in terms of intelligibility.

VII. CONCLUSION

In this work we looked to improve the robustness of discrete neural codec-based TTS models. To this end, we proposed MARS6, which combines several existing and new techniques for speech language model design, training, and inference. To evaluate robustness, we proposed a new test set built on the EARS dataset, consisting of harder and more diverse speech utterances than in other benchmarks. We compared MARS6 against several prior state-of-the-art TTS baselines, and found that MARS6 achieves competitive results with models many multiples larger, particularly in terms of target speaker similarity. Taken together, we show how many recent language and speech language modelling techniques can be effectively combined to achieve a compact, robust, and expressive TTS model.

REFERENCES

[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 *Interspeech*, 2024.

[2] 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.

[3] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *NeurIPS*, 2019.

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

[5] Y. Hu, C. Chen, S. Wang, E. S. Chng, and C. Zhang, "Robust zero-shot text-to-speech synthesis with reverse inference optimization," *arXiv preprint arXiv:2407.02243*, 2024.

[6] 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.

[7] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao *et al.*, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[8] J. Hong, N. Lee, and J. Thorne, "ORPO: Monolithic preference optimization without reference model," *arXiv preprint arXiv:2403.07691*, 2024.

[9] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Interspeech*, 2024.

[10] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[11] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," in *NeurIPS*, 2024.

[12] A. Baade, P. Peng, and D. Harwath, "Neural codec language models for disentangled and textless voice conversion," in *Interspeech 2024*, 2024, pp. 182–186.

[13] H. Siuzdak, "SNAC: Multi-scale neural audio codec," Feb. 2024. [Online]. Available: https://github.com/hubertsiuzdak/snac

[14] Y. Song, Z. Chen, X. Wang, Z. Ma, and X. Chen, "ELLA-V: Stable neural codec language modeling with alignment-guided sequence reordering," *arXiv preprint arXiv:2401.07333*, 2024.

[15] T. Dang, D. Aponte, D. Tran, and K. Koishida, "LiveSpeech: Low-latency zero-shot text-to-speech via autoregressive modeling of audio discrete codes," *arXiv preprint arXiv:2406.02897*, 2024.

[16] C. Wang, C. Zeng, B. Zhang, Z. Ma, Y. Zhu, Z. Cai, J. Zhao, Z. Jiang, and Y. Chen, "HAM-TTS: Hierarchical acoustic modeling for token-based zero-shot text-to-speech with model and data scaling," *arXiv preprint arXiv:2403.05989*, 2024.

[17] L. Yu, D. Simig, C. Flaherty, A. Aghajanyan, L. Zettlemoyer, and M. Lewis, "MEGABYTE: Predicting million-byte sequences with multiscale transformers," in *NeurIPS*, 2023.

[18] D. Misra, "Mish: A self regularized non-monotonic activation function," in *BMVC*, 2020.

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

[20] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP*, 2024.

[21] P. Gage, "A new algorithm for data compression," *The C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[22] Z. Allen-Zhu and Y. Li, "Physics of language models: Part 3.3, knowledge capacity scaling laws," *arXiv preprint arXiv:2404.05405*, 2024.

[23] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[24] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," in *Interspeech*, 2022.

[25] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu *et al.*, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[27] MetaVoice, "MetaVoice-1B," https://github.com/metavoiceio/metavoice-src, 2024, accessed: 2024-09-12.

[28] Microsoft, "WavLM-Base-Plus-SV Model," https://huggingface.co/microsoft/wavlm-base-plus-sv, 2024, accessed: 2024-09-12.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[30] W. Kang, X. Yang, Z. Yao, F. Kuang, Y. Yang, L. Guo, L. Lin, and D. Povey, "Libriheavy: a 50,000 hours asr corpus with punctuation casing and context," in *ICASSP*, 2024.

[31] W. Wang, Y. Song, and S. Jha, "GLOBE: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech," *arXiv preprint arXiv:2406.14875*, 2024.

[32] ShoukanLabs, "AniSpeech Dataset," https://huggingface.co/datasets/ShoukanLabs/AniSpeech, 2024, accessed: 2024-09-12.

[33] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The casual conversations v2 dataset," in *CVPR Workshops*, 2023.

[34] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[35] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.

[36] Coqui, "XTTS-v2: A multilingual text-to-speech model," 2024. [Online]. Available: https://huggingface.co/coqui/XTTS-v2