Interspeech 2024
1-5 September 2024, Kos, Greece

TacoLM: GaTed Attention Equipped Codec Language Model are Efficient Zero-Shot Text to Speech Synthesizers

Yakun Song[1], *Zhuo Chen*[2], *Xiaofei Wang*[3], *Ziyang Ma*[1], *Guanrou Yang*[1], *Xie Chen*[1,†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, X-LANCE Lab,
Shanghai Jiao Tong University, China
[2]ByteDance [3]Microsoft, One Microsoft Way, Redmond, USA

{ereboas, chenxie95}@sjtu.edu.cn

Abstract

Neural codec language model (LM) has demonstrated strong capability in zero-shot text-to-speech (TTS) synthesis. However, the codec LM often suffers from limitations in inference speed and stability, due to its auto-regressive nature and implicit alignment between text and audio. In this work, to handle these challenges, we introduce a new variant of neural codec LM, namely TacoLM. Specifically, TacoLM introduces a gated attention mechanism to improve the training and inference efficiency and reduce the model size. Meanwhile, an additional gated cross-attention layer is included for each decoder layer, which improves the efficiency and content accuracy of the synthesized speech. In the evaluation of the Librispeech corpus, the proposed TacoLM achieves a better word error rate, speaker similarity, and mean opinion score, with 90% fewer parameters and 5.2 times speed up, compared with VALL-E. Demo and code is available at https://ereboas.github.io/TacoLM/.

Index Terms: text-to-speech synthesis, language modeling, gated attention, zero-shot learning

1. Introduction

With the development of deep neural networks, text-to-speech (TTS) technology has made significant progress [1, 2, 3, 4]. Among them, zero-shot TTS only requires a short audio sample as prompts to synthesize high-quality speech for any unseen speaker. Zero-shot TTS does not need to be fine-tuned with speech data with respect to new speakers, which is of more practical value but still challenging. Early studies use continuous audio signals as input, relying on an explicit speaker encoder to embed a speaker's timbre, prosody, and speaking style [5, 6]. Some studies further rely on specifically designed speech disentanglement approaches to extract speaker-agnostic information [7, 8]. However, when the speaker embedding is inaccurate, the model's ability to generalize to zero-shot scenarios decreases dramatically.

Recent developments in diffusion and language modeling bring breakthroughs to the zero-shot TTS. The former [9, 10, 11] leverages the diffusion model [12] and its variants [13, 14] to estimate the target speech that shares the same distribution as the enrollment, while the latter [15] usually employs language models on a pre-trained neural codec to extract discrete audio representations and reconstruct high-quality waveforms. Both systems achieve impressive performance in the field of zero-shot speech synthesis. As the neural codec language model doesn't require a separated duration prediction model, which potentially enables a more direct end-to-end optimization, we focus on the improvement of this algorithm family in this work. In the domain of neural codec language models, VALL-E [16] is a prototypical and highly effective two-stage approach. Specifically, VALL-E takes the phoneme and acoustic tokens as prompts, and leverages an autoregressive (AR) and a non-autoregressive (NAR) language model, to generate coarse and fine-grained acoustic tokens of the unenrolled speaker, respectively. This powerful approach obviates the need for encoding the speaker into embeddings, while allowing for direct extraction of the environment and acoustic information from the audio samples.

Despite its achievements, the VALL-E still suffers from many drawbacks. One drawback that affects the practical application experience is the slow speed, arising from the use of an AR model to generate coarse-grained acoustic tokens. While the multi-head attention mechanism in the Transformer can model the relationship between pairs of tokens well, it often faces large computational and memory costs. The speed deficiency is especially prominent in the inference process, since the AR model needs to generate outputs in a token-by-token manner. Another limitation of VALL-E is the occasional mismatch between the synthesized speech and text prompts, which can manifest as repetitions, transpositions, or omissions. This is because the model does not align text and speech well. How to realize zero-shot TTS in both an efficient and accurate manner remains an open challenge.

In this study, we address the zero-shot TTS task with our proposed TacoLM (Ga**T**ed **A**ttention Equipped **Co**dec **L**anguage **M**odel), which is a two-stage (AR + NAR) framework, similar to VALL-E. TacoLM incorporates a MEGA module, as detailed in [17], which is based on a single-head gated attention mechanism with an exponential moving average. As a result, TacoLM is computationally efficient and requires minimal memory and storage. In addition, TacoLM employs a gated cross-attention mechanism to exchange information between text and audio, aiming at enhancing the accuracy of the synthesized speech. To evaluate TacoLM in the zero-shot scenario, we conducted experiments on the LibriSpeech [18] dataset. The experimental results show that TacoLM is superior to the advanced baseline VALL-E in terms of both objective (WER and speaker similarity) and subjective metrics (CMOS and SMOS). Ablation studies further corroborate the individual contributions of TacoLM's components to its overall effectiveness. Our contributions can be summarized as follows:

- We propose TacoLM, a two-stage zero-shot text-to-speech framework, which first discretizes the audio and text, and then relies on language models to demonstrate strong zero-shot speech synthesis capabilities. The training of TacoLM is also open-source to facilitate research in this direction.

† Corresponding author

10.21437/Interspeech.2024-1531

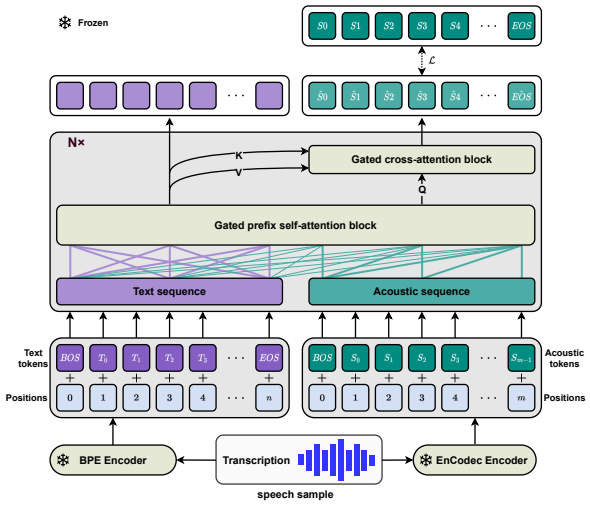

Figure 1: ***Framework Overview of the proposed TacoLM.*** *Gated prefix self-attention (GPSA) layers and gated cross-attention (GCA) layers are adopted in the AR model to generate the first layer of codec codes, while GCA layers are adopted in the NAR model to generate the rest layers of codes.*

- We train the discrete speech language model leveraging on a moving average equipped gated attention mechanism (MEGA), which has a lightweight computation and storage compared to the vanilla multi-head attention.
- We design a novel gated cross-attention layer that efficiently links up information between text and the audio sequence, further improving the performance and efficiency.

2. PROPOSED APPROACH

In this section, we introduce TacoLM, a TTS system with zero-shot speech synthesis capability. Similar to VALL-E, TacoLM consists of four main components: a text encoder, a neural audio codec, an autoregressive language model, and a non-autoregressive language model. In order to improve the training and inference efficiency and reduce the model size, we introduce two updates to the vanilla Transformer-based autoregressive language model, as depicted in Fig. 1. Firstly, a gated attention mechanism (namely MEGA [17]) is used as an efficient drop-in replacement for regular multi-head attention. Moreover, we propose a gated cross-attention layer to improve the accuracy of synthetic speech from the paired text prompts, which further boosts the computation and storage efficiency of the autoregressive language model.

2.1. Model framework

TacoLM operates in two stages. On the input side, we map the input text transcript and audio waveform to a sequence of semantic tokens, where a pre-trained neural audio codec model, EnCodec [19], is used as the audio tokenizer. At the first stage, these tokens are input into the AR language model to generate the codec codes of the first quantizer of EnCodec. Subsequently, in the second stage, the NAR language model predicts the codes of the rest quantizers in parallel. In the inference process, given a target text sequence and a 3-second recording of an unseen speaker as a prompt, the neural codec decoder is able to synthesize high-quality waveform, keeping the acoustic environment and timbre of the speech prompt.

Text encoder. Text encoders are utilized to extract content representations from the text transcription. In this paper, byte pair encoding (BPE) [20] is used to extract discrete text representation. We use text data transcribed from the LibriSpeech 960h training set to directly train a BPE text encoder model, where the word vocabulary is set to 2000 and the Character coverage rate (CCR) is set to 1.0.

Neural audio codec. In this paper, we use the pre-trained neural audio codec model EnCodec [19] as the audio encoder for TacoLM. EnCodec is an efficient real-time neural audio compression model that generates high-fidelity audio samples. Through the RVQ module of EnCodec, speech tokens have a hierarchical structure: speech tokens from the low-level residual quantizers recover acoustic attributes, such as the speaker's identity and the coarse-grained content information, while the rest successive quantizers learn finer acoustic details. Each quantizer computes the residuals from the lower-layer quantizers. The EnCodec encoder converts the 24 kHz input waveform into 75 Hz discrete tokens, effectively reducing the sampling rate to $1/320$. Each frame is represented using a residual vector quantization (RVQ) module with 8 hierarchical quantizers, each containing 1024 entries. In this setting, for each second of a 24 kHZ waveform, EnCodec's encoder synthesizes a matrix of 75×8 entries as a discrete representation of the audio.

Autoregressive language model. For discrete tokens from the first quantizer of the RVQ of EnCodec, we train an autoregressive language model. Its goal is to predict subsequent codewords from the first residual quantizer conditional on the target text sequence and the acoustic tokens. Formally, let $\mathcal{X}$ denote the transcribed text sequence, and $\mathcal{A}_{:,1}$ denotes the acoustic tokens of the first quantizer extracted from the speech $\mathcal{S}$. The autoregressive prediction process of TacoLM can be formulated as:

$$P\left(\mathcal{A}_{:,1} \mid \mathcal{X}; \theta_{AR}\right) = \prod_{t=1}^{T} p\left(\mathcal{A}_{t,1} \mid \mathcal{X}, \mathcal{A}_{<t,1}; \theta_{AR}\right) \quad (1)$$

Non-autoregressive language model. We obtain the codewords of the first quantizer through an autoregressive language model. In order to predict discrete tokens from the second to the last quantizer conditioning on the first layer, we train a non-autoregressive language model. Specifically, the prediction goal of the model can be expressed as:

$$P\left(\mathcal{A}_{:,2:8} \mid \mathcal{X}; \theta_{NAR}\right) = \prod_{l=2}^{8} p\left(\mathcal{A}_{:,l} \mid \mathcal{X}, \mathcal{A}_{:,<l}; \theta_{NAR}\right) \quad (2)$$

The non-autoregressive language model has 8 independent acoustic embedding layers. The discrete speech tokens of the first $l-1$ layers are summed up to be used as inputs to the model to predict the tokens of the l-th quantizer.

2.2. Gated Attention

The key design in TacoLM is an autoregressive decoder-only network component based on a gated attention mechanism, which contains a gated prefix self-attention (GPSA) layer and a gated cross-attention (GCA) layer.

Gated prefix self-attention layer. Unlike previous works [15, 16], the decoder-only network of the TacoLM autoregressive

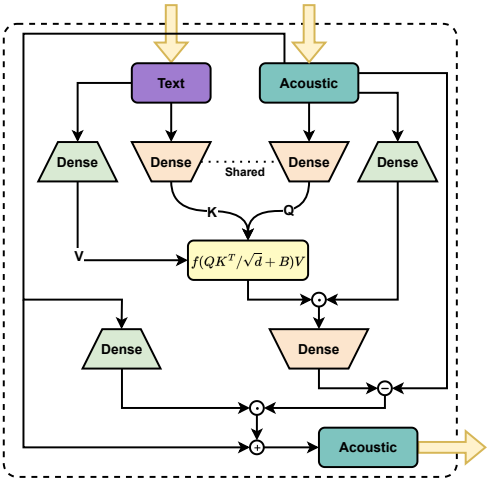

Figure 2: ***Illustration of the details of gated cross-attention layer.*** *B in the yellow box refers to the relative position bias, where we use RoPE [21] for position encoding. d is the dimension of Q and K.*

language model first introduces the Moving Average Mechanism Equipped with Gated Attention (MEGA) [17] as the drop-in replacement of traditional multi-head attention. The key idea of MEGA is to incorporate the classical exponential moving average (EMA) method into an attention mechanism for the whole sequence. With the benefit of single-head gated attention, MEGA has higher time and memory efficiency compared to multi-head self-attention. In this paper, we use bidirectional self-attention for text prefixes and unidirectional self-attention for discrete speech tokens to ensure the causality of audio generation.

Gated cross-attention layer. In the decoder-only causal language model, since unidirectional attention is applied to both the source sequence and the target sequence, less and less attention is focused on the source sequence as the length of the target sequence grows [22]. For the discrete speech language model, the attention degradation problem results in text mismatches when generating long sequences. In order to alleviate the problem, we propose a gated cross-attention layer after the GPSA layer, the structure of which is shown in Fig.2. The gated cross-attention computes the key and value matrices of attention for text sequences and the query matrix of attention for acoustic sequences, which makes it focus on the text part and effectively mitigates the attention degradation problem since the key matrix is not affected by the growth of the target audio sequence.

2.3. Inference

In the inference process, we first encode the text transcription into a sequence of discrete codewords using BPE and the recorded audio prompts into an acoustic matrix using the EnCodec encoder. Both prompts are used for AR and NAR models. For the AR model, we employ the sample-based decoding method since beam search may cause the language model to enter an infinite loop, and greedy decoding is very unstable. In addition, the sample-based decoding method can significantly enhance the diversity of speech output. On the other hand, for the NAR model, we use greedy decoding to select tokens with the highest probability. Finally, we use the decoder of the neural codec to synthesize waveforms conditioned on a sequence of eight residual quantizer codewords.

Table 1: ***Objective and subjective evaluation for zero-shot TTS.*** *For a fair comparison, we train VALL-E on the LibriSpeech dataset. † indicates the audio is fed through EnCodec's encoder and decoder to eliminate the interference of the codec on the experimental results. The subjective evaluations are shown with a 95% confidence interval (CI).*

Models	SPK (↑)	WER (↓)	CMOS (↑)	SMOS (↑)
Ground Truth†	0.9130	1.6211	0.30±0.04	4.50±0.05
VALL-E	0.8617	6.2461	0±0	3.50±0.11
TacoLM *(ours)*	**0.8696**	**5.9560**	**0.12±0.08**	**3.75±0.11**

Table 2: ***Model size, peak memory consumption, training speed, and inference latency comparison*** *of the autoregressive model with input length of 4K. All the inference experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.*

Models	**#Param.**	**Mem.**	**Training Speed** (↑) (epochs/hour)	**Inference Latency** (↓) (RTF)
VALL-E	154.3M (1×)	9.7GiB (1×)	2.31 (1×)	7.54 (1×)
TacoLM	**15.8M** (**0.10**×)	**3.0GiB** (**0.31**×)	**5.45** (**2.37**×)	**1.45** (**5.20**×)

3. EXPERIMENTS

3.1. Experimental Setup

3.1.1. Data

We train TacoLM on the multi-speaker English speech dataset LibriSpeech [18]. Specifically, we use `train-clean-100`, `train-clean-360`, and `train-other-500` as the training set, which contains a total of 960 hours of 16kHz English speech. For the evaluation, we select only the test samples from `test-clean` and `test-other` with a duration between 4 and 10 seconds following VALL-E [16].

3.1.2. Implementation Details

To extract acoustic tokens, we use the official open-source EnCodec [19] checkpoint, which is trained on a variety of 24 kHz monophonic audio data. Since LibriSpeech is a 16 kHZ speech dataset, we first upsample the speech data to 24kHZ to feed EnCodec. For text tokens, We use SentencePiece [23] as the text tokenizer to extract the byte pair encoding from the transcription of the speech corpus. In the AR model, we interleave 6 GPSA layers and 6 GCA layers. More specifically, the AR model has 6 blocks, each of which contains a GPSA layer and a GCA layer. The NAR model is similar to AR, except it uses a bidirectional attention mask and stacks 12 GPSA layers. For both AR and NAR, we set embedding dimension to 384, hidden state dimension to 384, feed-forward layer dimension to 768, damped EMA dimension to 24, key and value projection dimension to 240, and a dropout of 0.1. The activation function used is silu (sigmoid linear unit) [24]. All models are trained in parallel using 8 NVIDIA GeForce RTX 3090 GPUs with a batch size of 8192 tokens per GPU, learning a total of 240k steps. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$. For the learning rate schedule, we linearly increase the learning rate from zero to a peak of 10^{-3} for the first 12k updates, followed by a linear decay. The weight decay is 0.05 and clip-norm is 1.0.

Table 3: ***Detailed results of ablation study for TacoLM modules.***

Models	SPK(↑)	WER(↓)	PPL(↓)	#Param.	Mem.	Training Speed(↑) (epochs/hour)	Inference Latency(↓) (RTF)
TacoLM	**0.8696**	**5.9560**	**13.76**	**15.8M** (**0.10**×)	**3.0GiB** (**0.31**×)	**5.45** (**2.37**×)	**1.45** (**5.20**×)
TacoLM w/o GCA	0.8632	6.5157	13.85	18.5M (0.12×)	4.4GiB (0.46×)	3.16 (1.37×)	2.36 (3.19×)
TacoLM w/o GCA&GPSA	0.8617	6.2461	13.95	154.3M (1×)	9.7GiB (1×)	2.31 (1×)	7.54 (1×)

3.1.3. Baselines

We compare zero-shot speech synthesis performance with the recent state-of-the-art zero-shot TTS system VALL-E, which was trained on the 60k hours Librilight dataset [25] in the original paper. We reproduce VALL-E[1] and re-train it on the LibriSpeech 960h training set. For a fair comparison, we process the ground truth speech by feeding through EnCodec's encoder and decoder, to eliminate the interference of the audio codec. For each test sample, we provide the first 3 seconds of speech as the speech prompt and ask the model to synthesize the speech of the specified transcription.

3.1.4. Evaluation Metrics

We use speaker similarity (SPK) and word error rate (WER) as the objective metrics, and conduct the mean opinion score (MOS) evaluation as the subjective metrics. Specifically, for SPK evaluation, we use the state-of-the-art speaker verification model, WavLM-TDNN [26], to assess speaker similarity between the speech prompt and the synthesized speech. WavLM-TDNN predicts similarity scores in the range of [-1,1], with larger values indicating higher similarity between the synthesized speech and the speech prompt. When the similarity is greater than 0.86, WavLM-TDNN claims that two speech samples come from the same speaker. We also evaluated the accuracy of the synthesized speech to the text prompt. We perform Automatic Speech Recognition (ASR) on the generated speech and calculate the WER relative to the original transcription. In this work, we use the current advanced model Conformer-Transducer [27] as the ASR model. For the subjective evaluation, we selected 60 testing pairs randomly, each listened to by a minimum of 15 listeners. The listeners are then asked to rate the audio samples based on naturalness or speaker similarity. We analyze the comparative mean option score (CMOS) in terms of naturalness, and the similarity mean option score (SMOS) which measures speaker similarity. SMOS is rated on a scale of 1 to 5 on a 0.5-point increment, where higher is better. CMOS ranges from -1 to 1, and higher scores indicate better system performance compared to the baseline.

For our main concern of efficiency issue, we recorded the number of model parameters, training memory usage, training speed, and inference latency of the models. The inference latency is measured using the real-time factor (RTF) metric, which means how much time the model needs to generate one second of audio. The inference length is fixed at 4500 tokens, which support EnCodec with a target bandwidth of 24kbps to synthesize 15 seconds of speech.

[1]https://github.com/Ereboas/TacoLM/tree/valle

3.2. Result

We present the zero-shot TTS evaluation results in Table 1. TacoLM outperforms the baseline with statistical significance (p-value < 0.05 from the Wilcoxon signed-rank test). From Table 2, TacoLM has only 0.10 times the model size of VALL-E and 0.31 times the training memory, and achieves 2.37 and 5.20 times the training and inference speeds, respectively. Furthermore, due to the use of Rotational Position Embedding (RoPE) [21] for training, TacoLM can also be naturally extrapolated to any longer sequence than the training sequences during inference. To summarize, TacoLM demonstrates very competitive performance in terms of both effectiveness and efficiency.

3.3. Abaltion Study

In this section, we conduct ablation experiments to evaluate the effectiveness of the modules in TacoLM. Specifically, we analyze two variants of the model: TacoLM with GCA layers replaced by GPSA layers (TacoLM w/o GCA), and TacoLM with both GCA and GPSA layers replaced by vanilla multi-head attention layers (TacoLM w/o GCA&GPSA). For all three settings, we use a total of 12 attention layers with the same hyperparameters as TacoLM or VALL-E. In addition, we compared the perplexity (ppl) of the three AR language models. The results are presented in Table 3. We can observe that the inclusion of the GPSA layer leads to a significant reduction in the time and space costs, while causing little or no degradation in performance. On the other hand, when replacing the GCA layers, we observe a degradation in the speaker similarity (SPK) and speech accuracy (WER) performance, which indicates that cross-attention plays a crucial role in enhancing the accuracy of synthesized speech with respect to the source text prompt.

4. Conclusion & Limitation

In this paper, we proposed TacoLM, a zero-shot text-to-speech approach towards computational efficiency and synthesis accuracy, built upon the foundation of the AR + NAR two-stage neural codec language modeling. We introduced the gated prefix self-attention layer to reduce the memory and storage requirements while improving training and inference speeds. We further proposed a novel gated cross-attention layer to increase the accuracy of synthesized speech with respect to the source text. Our experimental results demonstrated that TacoLM not only outperformed the existing state-of-the-art zero-shot TTS system across a range of evaluations, but also significantly reduced memory and storage usage, and enhanced computational efficiency. This study still has some limitations. TacoLM relies on two-stage modeling, limiting its convenience for end-to-end training. In addition, limited to resources, TacoLM has not yet been trained and tested on a larger corpus. This prevents us from verifying its scalability, and is left for future works.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62206171 and No. U23B2018), Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, the International Cooperation Project of PCL and Alibaba Innovative Research Program.

6. References

[1] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Proc. NeurIPS*, 2020.

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021.

[3] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *IEEE Proc. ICASSP*, 2021.

[4] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier *et al.*, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, 2023.

[5] E. Casanova, C. Shulby, E. Gölge *et al.*, "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," in *Proc. Interspeech*, 2021.

[6] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022.

[7] N. Kumar, A. Narang, and B. Lall, "Zero-shot normalization driven multi-speaker text to speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[8] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," *Proc. NeurIPS*, 2022.

[9] K. Shen, Z. Ju, X. Tan, Y. Liu *et al.*, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. ICLR*, 2024.

[10] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Proc. NeurIPS*, 2024.

[11] R. Huang, M. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *IJCAI International Joint Conference on Artificial Intelligence*, 2022.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. NeurIPS*, 2020.

[13] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. ICLR*, 2023.

[14] L. Yang, Z. Zhang, Y. Song *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, 2023.

[15] Z. Borsos, R. Marinier, D. Vincent *et al.*, "AudioLM: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

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

[17] X. Ma, C. Zhou, X. Kong, J. He *et al.*, "Mega: Moving average equipped gated attention," in *Proc. ICLR*, 2022.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE Proc. ICASSP*, 2015.

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

[20] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, 2016.

[21] J. Su, M. Ahmed, Y. Lu, S. Pan *et al.*, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, 2024.

[22] Z. Fu, W. Lam, Q. Yu, A. M.-C. So, S. Hu, Z. Liu, and N. Collier, "Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder," *arXiv preprint arXiv:2304.04052*, 2023.

[23] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. EMNLP*, 2018.

[24] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2018.

[25] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov *et al.*, "Libri-Light: A benchmark for ASR with limited or no supervision," in *IEEE Proc. ICASSP*, 2020.

[26] S. Chen, C. Wang, Z. Chen, Y. Wu *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. Interspeech*, 2020.