Monotonic Attention for Robust Text-to-Speech Synthesis in Large Language Model Frameworks

Yike Zhang[1,], *Yiming Li*[1,2], *Jie Chen*[1,], *Qinghua Wu*[1,], *Songjun Cao*[1,], *Long Ma*[1,]

[1]Tencent YouTu Lab, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, China
yikezhang@tencent.com

Abstract

Text-to-speech (TTS) synthesis using large language models (LLMs) has demonstrated promising performance and has recently garnered significant attention. Despite their impressive naturalness, these methods often lack monotonic alignment constraints, resulting in issues such as repetition, omissions, and misalignment in the synthesized output. This paper introduces a stepwise monotonic attention algorithm tailored for LLM-based architectures, to enhance the robustness of TTS synthesis and effectively address these issues. Compared with the best existing model, VALL-E R, the proposed approach requires no additional forced aligners and exhibits greater robustness on out-of-domain test sets. Furthermore, experiments show that the proposed method scales well to large model sizes and large-scale training sets. Audio samples are available at https://zhangyike.github.io/llm_with_sma/.
Index Terms: Large-scale zero-shot TTS model, robust TTS synthesis, LLM-based TTS model, monotonic attention

1. Introduction

Text-to-speech (TTS) synthesis aims to generate human-like voices from given text and has extensive applications in our daily lives including voice navigation, virtual humans, and more. Neural TTS models have significantly improved the naturalness and quality of synthesized speech [1]. However, traditional neural TTS models often rely on high-quality studio recordings for training to generate high-quality single-speaker or multi-speaker speech [2, 3]. This limits the scalability of TTS systems and makes them unsuitable for scenarios where high-quality data is difficult to obtain, such as live streaming environments.

Large-scale neural TTS models have emerged as a solution to these limitations [4, 5]. These models can be trained using a large corpus of ordinary-quality speech from the Internet and exhibit zero-shot capabilities. Among the mainstream large-scale neural TTS frameworks, LLM architectures [4, 6–9] and flow-matching techniques [5, 10, 11] are prominent. Additionally, LLMs are often used to predict discrete acoustic tokens [4, 6–9, 12], further reducing the dependency on high-quality data.

Large-scale TTS methods based on LLM frameworks have garnered significant attention from researchers, with notable examples including the VALL-E series [4,13,14], VoiceCraft [12], CosyVoice [8], and others. However, decoder-only architectures of LLMs lack monotonic alignment constraints, which can lead to issues such as repetition, omissions, and misalignment in the synthesized speech. Various methods have been proposed to impose monotonic attention constraints within traditional neural TTS frameworks to improve system robustness [15–17]. However, these methods are typically designed for encoder-decoder structures and are not directly applicable to decoder-only architectures.

This paper aims to improve the robustness of large-scale TTS systems based on LLM frameworks. The main contributions of this paper are as follows:

- A stepwise monotonic attention algorithm tailored for LLM-based large-scale TTS systems is proposed. This proposed approach eliminates the need for forced alignment information and demonstrates robust performance on out-of-domain evaluation sets.
- An automatic selection approach is proposed to choose heads with alignment characteristics from multi-head attention modules. The proposed stepwise monotonic attention algorithm is only applied to the selected heads.
- An efficient parallel computing approach is proposed to handle variable text lengths in batch processing. This significantly speeds up the training process.

2. Prior Work

The attention mechanism [18] is widely used in sequence-to-sequence modeling with encoder-decoder architectures [19, 20] as well as in LLMs with decoder-only architectures [21,22]. Enhancing the monotonicity of attention alignment can effectively improve the robustness of attention-based models. Monotone alignment algorithms have been extensively studied in encoder-decoder architectures [23–25]. However, these methods are not suitable for LLM-based TTS models with decoder-only architectures.

Recently, some researchers have explored ways to improve the robustness of LLM-based TTS models. Song et al. rearranged the inputs of neural codec language models by interleaving phoneme and acoustic tokens, forcing the generation process to focus on the current phoneme [26]. Han et al. aligned phonemes with acoustic tokens and combined them to train neural codec language models [14]. Both methods rely on an additional aligner to obtain forced alignment information for the speech-transcription pairs before the training stage. When the training data comes from diverse sources, it is usually difficult to obtain a robust aligner that can provide accurate alignment information for the entire dataset. Du et al. adopted transducer loss [27] to help the model implicitly learn monotonic alignment constraints during training [28]. This method is extremely memory-consuming.

In this paper, we propose a stepwise monotonic attention algorithm for decoder-only models, which forces the model to explicitly learn monotonic alignment constraints. The proposed approach is alignment-free during both training and inference, and consumes almost no additional memory.

3. Monotonic Attention

Inspired by [15], we adopted stepwise monotonic attention in LLM-based TTS models. Notably, the proposed method can be applied to any decoder-only models with an attention mechanism, such as GPT [21] or LLaMA [22] based models, and is not limited to TTS tasks.

The proposed approach is implemented on CosyVoice [8]. To improve the robustness of pronunciation, a phoneme converter is used to replace the Whisper BPE tokenizer [29]. Given a speech utterance X and the corresponding transcript Y, a speech tokenizer converts X into a sequence of acoustic tokens $a_1, a_2, ..., a_T$ while a phoneme converter transforms Y into a sequence of phonemes $p_1, p_2, ..., p_N$. The inputs of LLM are:

$$[\langle S\rangle, v, \{p_n\}_{n\in[1:N]}, \langle T\rangle, \{a_t\}_{t\in[1:T]}, \langle E\rangle] \tag{1}$$

where $\langle S\rangle$, $\langle T\rangle$, $\langle E\rangle$ represent the start of sequence, a separator between phonemes and acoustic tokens, and the end of sentence respectively, v is the speaker embedding.

3.1. Stepwise monotonic attention

In the LLM component, self-attentions within each transformer layer compute attentions among phonemes, among acoustic tokens, and between phonemes and acoustic tokens. The attentions between phonemes and acoustic tokens are analogous to the cross-attention mechanism in encoder-decoder models.

Formally, the energies in self-attention are computed as:

$$E = \frac{QK^T}{\sqrt{d_k}} \tag{2}$$

where Q and K are the query and key matrices with dimensions $(N+T+3)\times d_k$. To compute the monotonic attention between phonemes and acoustic tokens, a cross mask M_x with dimensions $(N+T+3)\times(N+T+3)$ is defined as:

$$M_x[N+3:N+T+3, 2:N+2] = 0 \tag{3}$$

All other positions are masked when computing monotonic attention.

The proposed monotonic attention is computed step by step. At each step $i \in [N+3, N+T+3)$, the selection probability of phoneme p_j is estimated from the energies E using a sigmoid function:

$$P[i,j] = M_x[i,j]\sigma(M_c[i,j]E[i,j]) \tag{4}$$

where $j \in [2, N+2)$ and M_c is a causal mask. Since the sigmoid function is particularly sensitive to the scale of the energies E, or correspondingly, the scale of Q and K, weight normalization is applied to Q and K. To encourage the selection probabilities to approximate binary values, zero-mean, unit-variance Gaussian noise is added to the pre-sigmoid activations.

Similar to [15, 23], we use a recursive method to efficiently compute the attention scores:

$$\alpha[i] = \alpha[i-1]\cdot P[i] + [0; \alpha[i-1, :-1]\cdot(1-P[i, :-1])] \tag{5}$$

where $[0;]$ denotes zero-padding. The initial state $\alpha[N+2]$ is defined as a one-hot vector with $\alpha[N+2, 2] = 1$.

Attentions among phonemes and among acoustic tokens are computed in the standard way as proposed in [20].

3.2. Alignment heads selection

In contrast to the single-head cross-attention in encoder-decoder based models, multi-head attentions in LLMs exhibit different characteristics. Only a few heads in specific layers demonstrate alignment characteristics. Therefore, it is crucial to identify these alignment heads for fine-tuning with monotonic attention constraints as described in Eq.(5). This paper proposes a method to automatically identify the alignment heads in a pre-trained LLM by computing a diagonal ratio:

$$\gamma = \frac{\sum_{j=2}^{N+1}\lambda[s_j : e_j, j]}{\sum_{i=N+3}^{N+T+2}\sum_{j=2}^{N+1}\lambda[i,j]} \tag{6}$$

where λ represents the standard attention score, and

$$\begin{aligned} s_j &= \max(0, k\cdot(j-2)-\tau)+N+3 \\ e_j &= \min(k\cdot(j-1)+\tau, T)+N+3 \end{aligned} \tag{7}$$

define the interval, with $k = \lfloor T/N + 0.5\rfloor$ and τ being the overlap window of two adjacent intervals.

In our preliminary experiments, we observed that the absolute value of γ varies among different samples, and the number of heads with explicit alignment characteristics is typically no more than three. We compute the diagonal ratio for each head in every layer across several batches of samples and sum the diagonal ratios corresponding to the same head. Finally, the top two or three heads with the highest summed diagonal ratios are selected as the alignment heads. During the fine-tuning stage, Eq.(5) is applied only to the selected alignment heads, while standard attention is applied to the other heads.

3.3. Variable phoneme numbers in batch processing

As shown in Eq.(1), the number of phonemes varies across training samples within a batch. Consequently, Eq.(5) cannot be computed in parallel since the starting step $N+3$ differs for each sample. To address this issue, we shift each element of P by $N+3$ positions along the first dimension to align the element corresponding to the first acoustic token to the first position, wrapping around when the end of the dimension is reached:

$$P'[i,j] = P[(i-N-3) \bmod (N+T+3), j] \tag{8}$$

Here, mod denotes the remainder operation. Then, Eq.(5) can be applied to the shifted selection probabilities P' with the initial state $\alpha'[-1, 2] = 1$, resulting in the shifted attention scores α'. Similarly, α can by obtained by shifting α' back to the right:

$$\alpha[i,j] = \alpha'[(i+N+3) \bmod (N+T+3), j] \tag{9}$$

3.4. Zero-shot inference

The proposed stepwise monotonic attention algorithm is well-suited for zero-shot inference. During zero-shot inference, the input to the LLM model is structured as follows:

$$[\langle S\rangle, v, \{\hat{p}_n\}_{n\in[1:\hat{N}]}, \{p_n\}_{n\in[1:N]}, \langle T\rangle, \{\hat{a}_t\}_{t\in[1:\hat{T}]}] \tag{10}$$

where p and $\hat{p}$ represent the phonemes of the text to be synthesized and the transcript of the prompt, respectively, and $\hat{a}$ denotes the acoustic tokens corresponding to the prompt. With these inputs and an initial attention state $\alpha[\hat{N}+N+2, 2] = 1$, the initial attention state for the first acoustic token to be generated, $\alpha[\hat{N}+N+\hat{T}+2]$, can be obtained using Eq.(5). The LLM then recursively generates acoustic tokens using Eq.(5) in the alignment heads.

4. Experiments setup

4.1. Dataset

We used two scales of training sets to verify the performance of the proposed approach. The small-scale dataset contains approximately 1,000 hours of internal Chinese speech and 960 hours of English speech from the LibriSpeech corpus. The large-scale dataset consists of approximately 100,000 hours of Chinese speech and 50,000 hours of English speech.

The proposed approach was also evaluated on multiple test sets, including the LibriSpeech test-clean set, SeedTTS sets, and the ELLA-V hard case set [26]. Following [14], a cross-sentence test set (denoted as Lirbi-X) was build with utterances ranging from 4 to 10 seconds from the LibriSpeech test-clean set. The SeedTTS sets consist of three subsets, denoted as Seed-ZH, Seed-ZH-Hard, and Seed-EN respectively. The original ELLA-V hard case set has 100 specially designed English sentences similar to those in Seed-ZH-Hard. To eliminate disturbances caused by prompts, we provided 4 prompt for each sentence, resulting in a total of 400 challenging English sentences (denoted as ELLA-V-Hard).

4.2. Training configuration

We implemented the proposed monotonic attention algorithm on CosyVoice. We only re-trained the LLM component with the default model configuration, except for replacing the BPE text tokenizer with a phoneme tokenizer. We trained two models of different sizes. The small model (Phn-Cosy-S) consists of 7 transformer blocks, with a hidden size of 2,048 and 8 heads in each attention module, resulting in approximately 100M parameters. The large model (Phn-Cosy-L) consists of 14 transformer blocks, with a hidden size of 4,096 and 16 heads in each attention module, resulting in approximately 300M parameters. Phn-Cosy-S was trained with standard attention for 60 epochs on the small-scale dataset. Phn-Cosy-L was trained with standard attention for 10 epochs on the large-scale dataset. We then identified 2 alignment heads in Phn-Cosy-S and 3 alignment heads in Phn-Cosy-L utilizing Eq.(6). We fine-tuned Phn-Cosy-S for 10 epochs with monotonic attention from the checkpoint of the 50th epoch, and fine-tuned Phn-Cosy-L for 2 epochs with monotonic attention from the checkpoint of the 8th epoch. The fine-tuned models are denoted as Phn-Cosy-S-SMA and Phn-Cosy-L-SMA, respectively. A constant learning rate of 1e-5 was used during the fine-tuning stage.

4.3. Evaluation metrics

We adopted the WER or character error rate (CER) and speaker similarity (SIM) metrics for objective evaluation. For subsets Seed-ZH and Seed-ZH-hard, CERs were computed using the Paraformer model. Whisper Large V3 was used to compute WER on Seed-EN, and the Conformer-Transducer ASR model[1] was used to compute WER on Libri-X and ELLA-V-Hard. For similarity, we used WavLM-large [30] fine-tuned on the speaker verification task to obtain speaker embeddings, which were then used to calculate the cosine similarity of generated speech against the corresponding prompt.

4.4. Baseline

In addition to Phn-Cosy-S and Phn-Cosy-L, we also compared the proposed approach with the best existing algorithm, VALL-E R, for addressing the same problem. To ensure a fair comparison, we implemented the core concepts of VALL-E R on Phn-Cosy-S, enabling the prediction of acoustic tokens alongside aligned phonemes. We also incorporated the sampling strategy for phonemes with monotonic alignment. This modified model is referred to as Phn-Cosy-S-Ali.

[1] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge

5. Results and analysis

5.1. Objective evaluation

We first evaluated the impact of the proposed stepwise monotonic attention algorithm on WER/CER and speaker similarity across all test sets mentioned in Subsection 4.1. The results in Table 1 demonstrate that the proposed method, with both training configurations, can effectively reduce WER/CER on hard case sets, while having little influence on general test sets. This indicates that the proposed method can not only improve the robustness of LLM-based models but also scale well to large model sizes and large-scale training sets. Although Phn-Cosy-S-Ali effectively reduces WER/CER on hard case sets, it leads to significant WER/CER deterioration on Seed-ZH and Seed-EN sets. This indicates that the method proposed in [14] does not perform well on out-of-domain test sets with small-scale training sets. Additionally, the proposed method does not negatively affect speaker similarity.

To demonstrate the effect of the proposed approach more intuitively, a detailed CER/WER comparison on Seed-ZH-Hard and ELLA-V-Hard is presented in Table 2. Insertion and deletion errors are usually associated with word-repeating and word-missing issues. Table 2 shows that Phn-Cosy-L-SMA achieves significant reductions in deletion/insertion error rates on both hard case sets compared to Phn-Cosy-L. These results indicate that the proposed method can effectively alleviate word-repeating and word-missing problems, thereby improving the robustness of LLM-based TTS models.

5.2. Alignment analysis

Based on previous studies, the robustness of attention-based autoregressive generative models is highly related to the monotonicity of certain attention alignments. Therefore, we investigate whether the proposed method can enhance the monotonicity of attention alignment as expected. Using three typical examples from ELLA-V-Hard, we plotted the attention scores from the same alignment head in Phn-Cosy-S-SMA and Phn-Cosy-S in Figure 1. The results demonstrate that the proposed approach can indeed improve the monotonicity of attention scores. Take subfigures (c1) and (c2) as an example, the target text is “In the bustling bustling bustling bustling bustling bustling metropolis”. Phn-Cosy-S failed to align the acoustic token to the correct “bustling” word in the target text, leading to several extra “bustling” in the synthesized speech. On the contrary, Phn-Cosy-S-SMA can align acoustic tokens to the correct “bustling” words and synthesize accurate speech.

In our preliminary experiments, we found that the selection of alignment heads heavily influences the performance of the proposed stepwise monotonic attention algorithm. If a non-alignment head is selected for stepwise monotonic attention fine-tuning using Eq.(5), the accuracy of LLM prediction of acoustic tokens will be severely compromised. Therefore, we also evaluated the efficiency of the proposed alignment heads selection method. Figure 2 shows attention scores from a non-alignment head in Phn-Cosy-S. All acoustic tokens in this head focus on the speaker embedding, rather than

Table 1: *Results of WER/CER and speaker similarity across different Chinese and English test sets. Libri-X, Seed-EN, and ELLA-V-Hard are English test sets, Seed-ZH and Seed-ZH-Hard are Chinese test sets.*

Model	**Libri-X**		**Seed-ZH**		**Seed-EN**		**Seed-ZH-Hard**		**ELLA-V-Hard**	
	WER (%)	SIM	CER (%)	SIM	WER (%)	SIM	CER (%)	SIM	WER (%)	SIM
Phn-Cosy-S	4.77	0.593	4.03	0.710	5.32	0.587	17.76	0.703	22.98	0.579
Phn-Cosy-S-Ali	4.68	0.594	5.08	0.716	6.40	0.583	12.00	0.707	19.41	0.579
Phn-Cosy-S-SMA	4.48	0.595	4.08	0.709	5.62	0.581	12.51	0.704	19.03	0.578
Phn-Cosy-L	3.12	0.580	2.28	0.752	3.81	0.601	10.42	0.739	14.92	0.602
Phn-Cosy-L-SMA	2.99	0.596	2.33	0.750	3.65	0.600	8.53	0.740	12.01	0.602

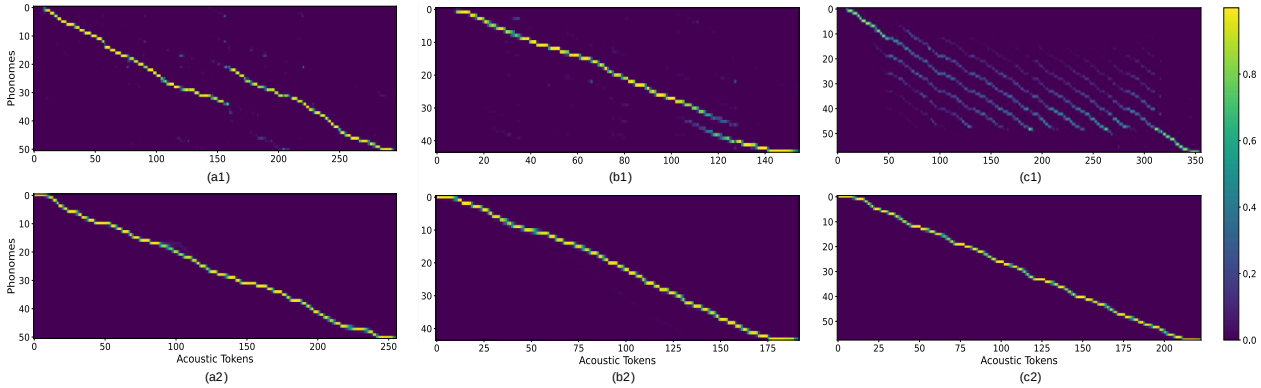

Figure 1: *Attention scores in a selected alignment head. The upper figures show attention scores from Phn-Cosy-S, while the lower figures show attention scores from Phn-Cosy-S-SMA. (a1), (b1), and (c1) represent repetition, omission, and misalignment issues in the synthesized speech, respectively.*

Table 2: *Detailed CER/WER comparison on hard case sets. Sub, Del, and Ins refer to Substitution, Deletion, and Insertion error rates respectively.*

Model	**Seed-ZH-Hard** Sub |Del |Ins (%)	**ELLA-V-Hard** Sub |Del |Ins (%)
Phn-Cosy-L	7.14 |1.64 |1.63	8.56 |4.07 |2.29
Phn-Cosy-L-SMA	7.15 |1.14 |0.24	7.92 |2.36 |1.73

the whole phoneme sequence. The focus rate proposed in [3] assigns a large value to this head by computing $F = \frac{1}{T}\sum_{i=N+3}^{N+T+2} \max_{2\leq j\leq N+1} \lambda[i,j]$. Eq.(6) assigns a very low diagonal ratio value to this head, since the proposed alignment heads selection method considers both focus and completeness. The visualization analysis in Figure 2 shows that the proposed alignment heads selection approach can effectively identify heads with truly monotonic characteristics.

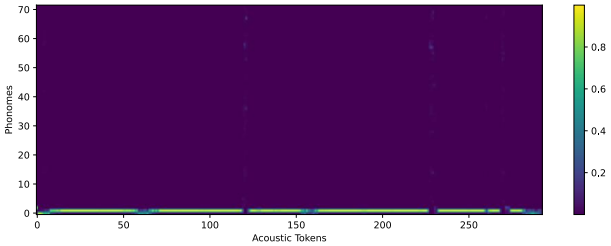

Figure 2: *Attention scores from a non-alignment head in Phn-Cosy-S. The focus rate of this head is 0.87, while its diagonal ratio is 0.03.*

6. Conclusions

This paper proposes a stepwise monotonic attention algorithm to improve the robustness of LLM-based TTS models. The proposed method can also be applied to any autoregressive generative models with a decoder-only transformer architecture. Additionally, this paper introduces an automatic alignment heads selection method in a multi-head attention scenario, as well as a parallel training method. Compared with previous methods to improve the robustness of LLM-based TTS models, the proposed algorithm does not require any additional modules or inputs, such as forced alignment information derived from a pre-trained forced aligner, and performs well on out-of-domain test sets. The proposed algorithm can scale well to large model sizes and large-scale training sets. Experimental results show that the proposed method can effectively alleviate issues such as repetition, omission, and misalignment in the synthesized output. Furthermore, visualization analysis indicates that the proposed method indeed improves the monotonicity of attention alignment, thereby enhancing the robustness of LLM-based TTS models.

7. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, vol. 164, 2017.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[4] 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.

[5] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[6] 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.

[7] Y. Zhou, X. Qin, Z. Jin, S. Zhou, S. Lei, S. Zhou, Z. Wu, and J. Jia, "Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 554–563.

[8] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

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

[10] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[11] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," *arXiv preprint arXiv:2406.18009*, 2024.

[12] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," *arXiv preprint arXiv:2403.16973*, 2024.

[13] 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.

[14] 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.

[15] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," *arXiv preprint arXiv:1906.00672*, 2019.

[16] X. Liang, Z. Wu, R. Li, Y. Liu, S. Zhao, and H. Meng, "Enhancing monotonicity for robust autoregressive transformer tts." in *INTERSPEECH*, 2020, pp. 3181–3185.

[17] P. Neekhara, S. Hussain, S. Ghosh, J. Li, R. Valle, R. Badlani, and B. Ginsburg, "Improving robustness of llm-based speech synthesis by learning monotonic alignment," *arXiv preprint arXiv:2406.17957*, 2024.

[18] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[19] I. Sutskever, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[20] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[21] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[23] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *International conference on machine learning*. PMLR, 2017, pp. 2837–2846.

[24] X. Ma, J. Pino, J. Cross, L. Puzon, and J. Gu, "Monotonic multihead attention," *arXiv preprint arXiv:1909.12406*, 2019.

[25] X. Ma, A. Sun, S. Ouyang, H. Inaguma, and P. Tomasello, "Efficient monotonic multihead attention," *arXiv preprint arXiv:2312.04515*, 2023.

[26] 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.

[27] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[28] 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.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.