Interspeech 2024
1-5 September 2024, Kos, Greece

Synthesizing Long-Form Speech merely from Sentence-Level Corpus with Content Extrapolation and LLM Contextual Enrichment

Shijie Lai[1], *Minglu He*[1], *Zijing Zhao*[1], *Kai Wang*[1*], *Hao Huang*[1,2*], *Jichen Yang*[3]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Key Laboratory of Multi-lingual Information Technology, Urumqi, China
[3]School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou, China

{lsj, hml, mirror_zhao}@stu.xju.edu.cn, {wangkai, huanghao}@xju.edu.cn, yangjichen@gpnu.edu.cn

Abstract

Current text-to-speech (TTS) models can produce natural speech but often fail to synthesize long-form speech properly when only sentence-level corpus is available. The failure is mainly due to 1) poor length generalization of the acoustic model, 2) lack of appropriate pause marks in the inference text, and 3) absence of contextual information during training. We propose Content Extrapolation, which includes introducing Moving Average Equipped Gated Attention (MEGA) to improve the model's generalization for addressing 1) and presenting the Global-information-enhanced Classification Pause Insertion model (GCPI) to address 2). For 3), we propose LLM-based Contextual Enrichment (LLM-CE) to generate multiple sets of different contexts. Experiments show that the proposed methods solve the above issues and successfully generate long-form speech with clear pronunciation and natural prosody using only sentence-level corpus, reducing training costs.

Index Terms: text-to-speech, MEGA, pause insertion, LLM, contextual enrichment

1. Introduction

Neural text-to-speech (TTS) models [1–8] aim to generate human-like speech from text inputs. With the emergence of various TTS applications, there is an increasing demand for long-form speech synthesis. ParaTTS [9], building upon the GMM attention-based Tacotron2 [10], introduces three distinct context encoders to enhance the prosody of synthesized long-form speech. Since the modified Tacotron2 proposed in [10] has good length extrapolation capabilities, ParaTTS achieves sentence-level training and paragraph-level inference, reducing the computational resources required for training. Nonetheless, such autoregressive models suffer from slow inference speed and an inability to control the prosody of synthesized speech. Employing FastSpeech2 as the backbone, [11–18] proposes diverse context encoders to extract contextual information from input paragraphs or previously acoustic features, fostering coherence between synthesized sentences. However, these models face generalization issues and lack length extrapolation ability. When only sentence-level corpus is available, synthesized speech from these models exhibits catastrophic problems such as mispronunciation, erratic duration, and irregular pitch.

Another issue in synthesizing long-form speech is the absence of appropriate pauses in the input text during inference. Human speakers usually insert pauses when delivering long speeches or readings to take a breath or better express their emotions [19]. Depending on the insertion locations, these pauses can be categorized into respiratory pauses (RPs) [20] and punctuation-indicated pauses (PIPs). Phoneme alignment before TTS training adds corresponding RPs to the input text. In the inference phase, the input text lacks these pauses due to the inability to align, diminishing the generated speech's naturalness. [19, 21] utilize BERT [22] and BiLSTM [23] to predict these pauses, but the prediction accuracy is still low.

The final problem in long-form speech synthesis is the lack of contextual information to guide the synthesis process, resulting in synthesized speech that sounds less natural and coherent. Recently, Large Language Models (LLMs) have achieved significant success in natural language processing (NLP) [24–26], leading to increased interest in utilizing LLMs for research tasks in other domains. [26, 27] respectively utilize context information and text descriptions generated by LLMs to enhance the accuracy of visual question answering and zero-shot image classification. [28] employs LLMs to correct errors in automatic speech recognition (ASR) N-best decoding results, significantly reducing word error rates (WER). Meanwhile, [29] explores using ChatGPT for emotion control in dialogue speech synthesis. However, no related work is currently on enhancing long-form TTS using LLMs.

To address the above issues, this paper proposes Content Extrapolation and LLM-based Contextual Enrichment (LLM-CE), enabling TTS models to synthesize long-form speech robustly using only sentence-level corpus. In Content Extrapolation, we introduce one of the state-of-the-art (SOTA) long sequence models, Moving Average Equipped Gated Attention (MEGA), to form the MEGA Encoder and Decoder, enhancing the acoustic model's generalization for synthesizing long-form speech. Additionally, we present the Global-information-enhanced Classification Pause Insertion model (GCPI) to insert pauses into the text during inference. Finally, the proposed LLM-CE utilizes LLM to generate multiple sets of contexts for the sentence-level corpus to guide the model training.

Objective and subjective experiments show that our proposed method can robustly synthesize about 1 minute of long-form speech using training data of only 2-7 seconds per sample. Furthermore, through ablation experiments, we validate the effectiveness of each module in the proposed method[1]. The contributions of this study are as follows:

- We propose Content Extrapolation, including the MEGA-based Encoder and Decoder, as well as GCPI.
- MEGA of Content Extrapolation enhances the model's generalization, enabling it to robustly synthesize long-form speech with clear pronunciation and successfully reducing the training costs of long-form TTS.
- GCPI of Content Extrapolation more accurately inserts pauses into text during inference, improving the prosody of synthesized speech.
- LLM-CE is proposed, which enables the model to gain the

[1]Speech sample: https://speechpaper.github.io/is2024/

10.21437/Interspeech.2024-1913

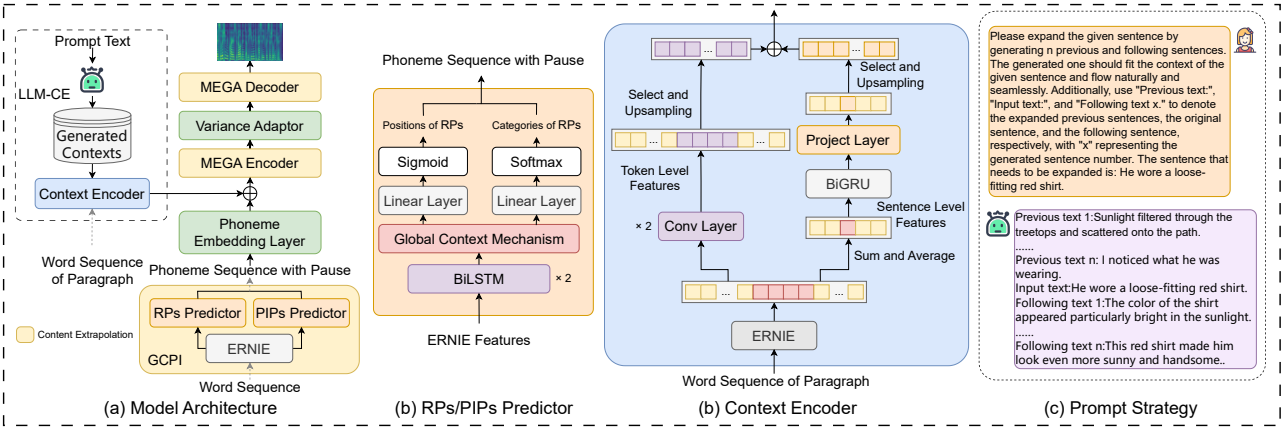

Figure 1: *Overview of the proposed method. The gray dashed line is only executed during inference.*

ability to capture contextual information when only sentence-level corpus is available, reducing training costs while maximizing the naturalness of synthesis.

2. Proposed method

As shown in Fig. 1(a), our proposed model is built upon FastSpeech2. In this architecture, the Content Extrapolation comprises MEGA Encoder, MEGA Decoder, and GCPI. Next, we will introduce the details of the proposed modules respectively.

2.1. MEGA Encoder & Decoder

Transformer has shown remarkable performance across various domains, but it is difficult to handle long sequence tasks [30]. MEGA, first proposed in [31], is a new attention that can model long context sequences, mainly composed of multi-dimensional damped exponential moving average (MDDEMA) and single-head gated attention (SHGA). MEGA has demonstrated SOTA-level performance in long sequence modeling tasks such as original speech classification and image classification.

Specifically, we introduce MEGA to replace the self-attention in the original FastSpeech2, forming the MEGA Encoder and Decoder to process input sequences and achieve length extrapolation. Given an input $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\} \in \mathbb{R}^{L\times D}$ and output $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_L\} \in \mathbb{R}^{L\times D}$, the MDDEMA can be defined as follows:

$$\boldsymbol{u}_t^{(d)} = \boldsymbol{\alpha}_d \odot \boldsymbol{\beta}_d x_t^{(d)} + (1 - \boldsymbol{\alpha}_d \odot \boldsymbol{\delta}_d) \odot \boldsymbol{u}_{t-1}^{(d)} \quad (1)$$

$$y_t^{(d)} = \boldsymbol{\eta}_h^T \boldsymbol{u}_t^{(h)}, \quad (2)$$

where $x_t^{(d)} \in \mathbb{R}$ and $y_t^{(d)} \in \mathbb{R}$ represent the d-th dimension element of the input $\boldsymbol{x}_t$ and output $\boldsymbol{y}_t$, respectively. $\boldsymbol{u}_t^{(d)} \in \mathbb{R}^N$ is the EMA hidden state, and $\odot$ is the element-wise product.

Then, the output of MDDEMA is combined with the attention mechanism. Let $\boldsymbol{X}' \in \mathbb{R}^{L\times D}$ be the output of EMA. The SHGA output can be calculated as:

$$\boldsymbol{O} = f\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\tau(\boldsymbol{X})} + \boldsymbol{b}_{rel}\right), \quad (3)$$

where $\boldsymbol{Q}, \boldsymbol{K}$ is computed by $\boldsymbol{X}'$, $\boldsymbol{V}$ is computed by $\boldsymbol{X}$, and $\boldsymbol{b}_{rel}$ is the rotary positional embeddings[26]. The reset gate $\boldsymbol{\gamma}$ and the update gate $\boldsymbol{\varphi}$ are designed to integrate the EMA output $\boldsymbol{X}'$ and the attention output $\boldsymbol{O}$:

$$\boldsymbol{\gamma} = \phi_{silu}\left(\boldsymbol{X}' W_\gamma + b_\gamma\right) \quad (4)$$

$$\boldsymbol{\varphi} = \phi_{sigmoid}\left(\boldsymbol{X}' W_\varphi + b_\varphi\right). \quad (5)$$

Finally, we can compute the candidate output $\hat{\boldsymbol{H}} \in \mathbb{R}^{L\times D}$ and the MEGA output $\boldsymbol{Y} \in \mathbb{R}^{L\times D}$:

$$\hat{\boldsymbol{H}} = \phi_{silu}\left(\boldsymbol{X}' W_h + (\boldsymbol{\gamma} \odot \boldsymbol{O}) U_h + b_h\right) \quad (6)$$

$$\boldsymbol{Y} = \boldsymbol{\varphi} \odot \hat{\boldsymbol{H}} + (1 - \boldsymbol{\varphi}) \odot \boldsymbol{X}. \quad (7)$$

2.2. GCPI

Inserting suitable pauses into long-form speech is crucial for enriching prosody. However, input texts often lack proper pause marks during inference, leading to models' inability to synthesize natural long-form speech. Inspired by [19], we propose GCPI to address this issue. The structure of GCPI, as shown in Fig. 1(a-b), comprises a pre-trained ERNIE3[2] [32] and two predictors. Each predictor consists of two layers of BiLSTM and a global context mechanism (GCM) [33] to predict the positions and categories of RPs/PIPs. Different from [19], we add GCM to enhance the global information of cell states in BiLSTM to improve prediction accuracy. In addition, compared to BERT, ERNIE3's unique model architecture and pre-training tasks enable it to directly handle longer texts without length limitations, making it more suitable for this task.

For GCM, let $\overrightarrow{\boldsymbol{h}}_t$ and $\overleftarrow{\boldsymbol{h}}_t$ represent the hidden states of the forward LSTM and backward LSTM at time step t, and L denotes the sequence length. The BiLSTM output $\boldsymbol{h}_t \in \mathbb{R}^D$ can be defined as follows:

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t \,||\, \overleftarrow{\boldsymbol{h}}_t], \quad (8)$$

where $[\cdot\,||\,\cdot]$ denotes the concatenation operation. So, we can directly obtain the sentence representation $\boldsymbol{G} = [\overleftarrow{\boldsymbol{h}}_1 \,||\, \overrightarrow{\boldsymbol{h}}_L]$. Global weight $\boldsymbol{i}_t^G$ and local weight $\boldsymbol{i}_t^H$ are designed to combine sentence feature with the output at each time step of the BiLSTM:

$$\boldsymbol{i}_t^G = \phi_{sigmoid}\left([\boldsymbol{G}\,||\,\boldsymbol{h}_t] W_G + b_G\right) \quad (9)$$

$$\boldsymbol{i}_t^H = \phi_{sigmoid}\left([\boldsymbol{G}\,||\,\boldsymbol{h}_t] W_H + b_H\right) \quad (10)$$

$$\hat{\boldsymbol{h}}_t = \boldsymbol{i}_t^G \odot \boldsymbol{G} + \boldsymbol{i}_t^H \odot \boldsymbol{h}_t, \quad (11)$$

where $\hat{\boldsymbol{h}}_t \in \mathbb{R}^D$ is the GCM output, which is used to predict the positions and categories of inserted pauses.

2.3. LLM-CE and Context Encoder

In order to equip the model with the ability to capture contextual information during training, thereby enhancing the naturalness and smooth transitions between sentences in synthesizing

[2]https://huggingface.co/nghuyong/ernie-3.0-base-zh

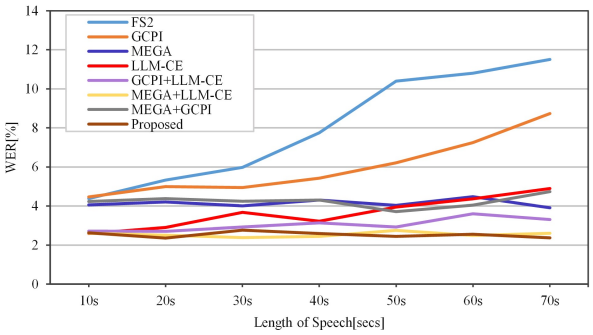

Figure 2: *Comparison of synthesis stability across various modules (measured by WER).*

long-form speech, this paper proposes LLM-CE. Specifically, we employ the latest GLM4[3] to generate context for the training corpus and design a context encoder to extract contextual features. GLM4 shows better performance than GPT4 on Chinese datasets.

2.3.1. Prompt strategy

We instruct GLM4 to generate n sentences of natural and fluent context before and after the utterances in the training data based on the text prompts. As illustrated in Fig. 1(d), the text prompts consist of 1) constraints on the generated context, 2) the form of the returned answers, and 3) the utterances from the training corpus. Additionally, we generate multiple sets of contexts containing different numbers of sentences to explore the effectiveness of various LLM-CE strategies in enhancing naturalness.

2.3.2. Context encoder

Inspired by [15], we design a simple context encoder to validate the effectiveness of LLM-CE. Fig. 1(c) shows that the context encoder consists of a pre-trained ERNIE3, two Conv1d layers, a BiGRU layer, and a projection layer. The ERNIE3 takes the context sequences as input to obtain token-level semantic features. The convolutional layer is used to refine and reduce the dimensionality of the extracted features to get token-level contextual features. Additionally, token-level features for each sentence in the ERNIE3 output are summed and averaged to obtain sentence-level semantic features. Subsequently, the sentence-level semantic features are fed into the BiGRU to extract sentence-level contextual features. Finally, the corresponding portions of token-level and sentence-level contextual features for the current training sentence are selected, upsampled, and added to obtain the complete contextual features, which are then fed into the acoustic model.

2.4. Training and inference

Before training begins, we need to collect and clean the data generated by LLM. During training, we use the available sentence-level corpus as input, coupled with LLM-CE, to train a TTS model capable of extrapolating and capturing contextual information. At the same time, GCPI is trained on the same data. During inference, we first employ the trained GCPI to insert appropriate pauses into the input long text and then directly send it to the acoustic model and context encoder to synthesize natural long-form speech robustly.

[3]https://open.bigmodel.cn/

Table 1: *Results of position and classification prediction.*

Model	Pause	Position Prediction			Category Prediction		
		Precision	Recall	F_β	Precision	Recall	F_β
CPI	RPs	0.536	0.381	$F_{0.5} = 0.496$	0.237	0.227	$F_{0.5} = 0.235$
	PIPs	1.000	1.000	$F_2 = 1.000$	0.724	0.724	$F_2 = 0.724$
GCPI	RPs	**0.657**	**0.454**	$F_{0.5} =$ **0.603**	**0.247**	**0.237**	$F_{0.5} =$ **0.245**
	PIPs	**1.000**	**1.000**	$F_2 =$ **1.000**	**0.729**	**0.729**	$F_2 =$ **0.729**

3. Evaluation and analysis

3.1. Datasets

We evaluated the effectiveness of the model on the BZNSYP[4] database, an open-source Chinese dataset containing around 12 hours of sentence-level speech by a female speaker (about 10,000 utterances, up to 7 seconds). During the acoustic model training, we resampled the audio to 16k Hz and divided 100 utterances for validation and 300 for testing.

For the GCPI model, we used 224 sentences each for the validation and testing sets. Since the training requires pause labels for classification, we adopted the method in [19] to obtain three distinct categories of pauses: brief pause ($<$ 100ms, denoted as "sp1"), medium pause (100-200ms, denoted as "sp2"), and long pause ($>$ 200ms, denoted as "sp3").

3.2. Model configuration

For LLM-CE, we generate five different context sets for each sentence in the training corpus, each containing three previous and following context sentences. We randomly select one set for each training sample during training to extract contextual features.

In proposed TTS model, the MEGA Encoder/Decoder consists of MDDEMA with 16 heads, SHGA, and 2-layer Conv1D. The dimensions of queries, keys, and values in SHGA were set to 128, 128, and 1024, respectively. The two Conv1D layers in the Context Encoder have dimensions of 512 and 256, the BiGUR has a dimension of 768, and the projection layer is a linear layer with a dimension of 256. The remaining configurations for the acoustic model followed [4]. A pre-trained HiFi-GAN vocoder [34] was used for waveform synthesis.

For the GCPI model, we utilize ERNIE3 to extract 768-dimensional hidden features from the input text and use a 2-layer 512-dimensional BiLSTM with GCM to refine the features further. We use the Adam optimizer [35] with an initial learning rate of 5×10^{-5} for training. The maximum training epoch is set to 50. Like [19], we use the F_β score to measure the accuracy of the predicted pause (with $\beta = 0.5$ for RPs and $\beta = 2$ for PIPs) and select the best model to save. All models are trained on an NVIDIA 4090 GPU.

3.3. Performance of GCPI

We compare our GCPI with the CPI [19] model. We evaluate the performance of these two models in predicting RPs and PIPs using precision, recall, and F_β score. The experimental results are shown in Table 1. It can be observed that the proposed GCPI achieves higher accuracy in predicting insertion position and classification than CPI. Therefore, in the subsequent experiments, we use GCPI to insert appropriate pauses into the speech.

[4]https://www.data-baker.com/open_source.htm

Table 2: *The MOS with 95% confidence intervals, WER, MCD, F0 RMSE, and Duration MSE of different models in synthesizing short speech.*

Model	MOS ↑	WER ↓	MCD ↓	F0 RMSE ↓	Duration MSE ↓
Ground Truth	4.10 ± 0.08	0.67%	-	-	-
FastSpeech2	3.84 ± 0.12	1.67%	4.5787	55.7324	0.0868
Proposed	$\mathbf{3.96 \pm 0.09}$	1.67%	**4.4973**	**53.6253**	**0.0705**

Table 3: *MOS experiment and ablation experiment results of the proposed model on synthesizing long-form speech.*

Model	MOS / CMOS	WER
FastSpeech2	3.34 ± 0.13	7.47%
Proposed	$\mathbf{3.88 \pm 0.07}$	**0.73**%
w/o GCPI	-0.092	0.75%
w/o LLM-CE	-0.126	1.77%
w/o MEGA	-0.347	2.00%

3.4. Stability analysis of synthesizing long-form speech

We first evaluate the performance of each module in the proposed method on handling data of different lengths to explore which module plays a crucial role in improving the stability of synthesis. Serious pronunciation errors occur when the model fails to synthesize longer speech stably. Therefore, we use the word error rate (WER) to measure the performance of each module. We randomly concatenate 300 short sentences from the test set to form texts of different lengths (corresponding to 10 to 70 seconds of speech). We use iFlytek ASR API[5] to evaluate the WER, and the results are presented in Fig. 2. As the length of the input text increases, the WER of FastSpeech2 continues to rise. Introducing the GCPI module leads to a slight decrease in the WER, but it still shows an upward trend. In contrast, models with MEGA and LLM-CE achieve lower WERs. It should be noted that LLM-CE performs better when synthesizing speech less than 60 seconds, but WER still rises as the input length increases. While MEGA has a slightly higher WER than LLM-CE, it maintains a consistent level throughout. The model that combines MEGA and LLM-CE achieves the lowest WER and remains stable. Therefore, we can conclude that MEGA is the key to synthesizing long-form speech stably, while LLM-CE further reduces pronunciation errors.

3.5. Evaluation of synthesized short and long-form speech

To evaluate the performance of the proposed model, subjective and objective experiments were conducted separately on synthesized speech from both in-domain short texts and out-of-domain long-form texts. In the objective comparison, mel cepstral distortion (MCD), F0 RMSE, and Duration MSE are utilized to measure the difference between the synthesized results and ground-truth speech. In subjective evaluation, mean opinion score (MOS) and comparative mean opinion score (CMOS) test were conducted to evaluate the naturalness of speech synthesized. Twenty native Chinese speakers participated in the test and were asked to give their MOS and CMOS scores.

For in-domain short texts, we randomly selected 20 audios from the test set for evaluation. As shown in Table 2, the synthesized speech by our proposed method achieves better naturalness compared to FastSpeech2, with lower values of WER, F0 RMSE, and Duration MSE. For out-of-domain long-form text, we collected 100 sentences from China News Service through the Internet (corresponding to 50-70 seconds of speech). We randomly selected 20 texts to synthesize speech for MOS and CMOS ablation experiments. Due to the lack of ground truth in long-form speech synthesis evaluation, we only tested WER in objective metrics. In Table 3, **w/o GCPI** denotes removing the GCPI module, while the rest of the labels follow a similar convention. It can be observed that the results in the first two rows demonstrate the effectiveness of the proposed method, which can stably synthesize natural long speech using only sentence-level corpora. The results of the ablation experiments in the last three rows also validate the effectiveness of each component.

[5]https://www.xfyun.cn/services/lfasr

Table 4: *AB preference test results for various LLM-CE strategies, bold indicates significant speech synthesis differences.*

A-WER	Method A	Perfence	Method B	B-WER	p-value
0.81%	Intercross	**0.428 vs. 0.572**	No intercross	0.81%	<0.005
0.88%	1 context sets	**0.352 vs. 0.648**	3 context sets	**0.71**%	<0.005
0.71%	3 context sets	**0.457 vs. 0.543**	5 context sets	0.73%	<0.05
0.73%	5 context sest	0.503 vs. 0.497	7 context sets	0.75%	0.89
0.75%	7 context sets	0.525 vs. 0.475	10 context sets	0.81%	0.24
0.77%	1 context utts	0.518 vs. 0.482	2 context utts	0.74%	0.19
0.74%	2 context utts	**0.443 vs. 0.557**	3 context utts	**0.73**%	<0.05
0.73%	3 context utts	0.533 vs. 0.467	4 context utts	0.78%	<0.1

3.6. Analysis of various LLM-CE strategies

We conducted AB preference testing to explore the effects of various LLM-CE strategies on synthesizing long-form speech. In Table 4, **Intercross** means that the sentences from the training corpus can appear at anywhere within the context, at the beginning, middle, or end, while **No intercross** implies the opposite. **M context sets** denotes generating **M** different context sets from the same sentence using LLM. In contrast, **N context utts** means that each sentence from the training corpus is preceded and followed by **N** generated contexts. As shown in Table 4, **No intercross** obtained higher preference scores. This is primarily because the speech prosody varies among the paragraph's beginning, middle, and end positions. Randomly placing the fixed-style sentences from the training corpus can lead to unstable model training, resulting in unnatural prosody on synthesized long-form speech. Meanwhile, as **M** and **N** increase, the preference of the corresponding LLM-CE strategy first rises and then falls, eventually becoming nearly the same. This may be due to the diversity of LLM-generated results and the poorer quality of longer contexts.

4. Conclusion

This paper proposes a novel TTS framework, comprising Content Extrapolation and LLM-CE, which can synthesize natural long-form speech stably only using sentence-level corpus, significantly reducing the training cost of long-form TTS. Both objective and subjective experiments show that the MEGA-based encoder and decoder in Content Extrapolation improve the length generalization of the acoustic model, allowing it to synthesize longer speech. The GCPI of Content Extrapolation accurately inserts appropriate pauses into the inference text, enriching the speech prosody. Meanwhile, LLM-CE uses context generated by LLM to equip the model to capture contextual information during training, improving the naturalness of long-form speech. Finally, AB preference testing concludes that an appropriate amount of context in LLM-CE can achieve optimal synthesis results.

5. Acknowledgements

This work was supported by Opening Project of Key Laboratory of Xinjiang, China (2020D04047), NSFC (61663044).

6. References

[1] 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: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017, pp. 4006–4010.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[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," in *NeurIPS*, 2019, pp. 3171–3180.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2020.

[5] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP*, 2021, pp. 6588–6592.

[6] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021, pp. 5530–5540.

[7] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *arXiv:2307.16430*, 2023.

[8] C. Du, Y. Guo, F. Shen, Z. Liu, Z. Liang, X. Chen, S. Wang, H. Zhang, and K. Yu, "Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding," *arXiv:2306.07547*, 2023.

[9] L. Xue, F. K. Soong, S. Zhang, and L. Xie, "Paratts: Learning linguistic and prosodic cross-sentence information in paragraph-based tts," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, pp. 2854–2864, 2022.

[10] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *ICASSP*, 2020, pp. 6194–6198.

[11] S. Lei, Y. Zhou, L. Chen, Z. Wu, S. Kang, and H. Meng, "Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis," in *ICASSP*, 2022, pp. 7922–7926.

[12] S. Lei, Y. Zhou, L. Chen, Z. Wu, S. Kang, and et al, "Context-aware coherent speaking style prediction with hierarchical transformers for audiobook speech synthesis," in *ICASSP*, 2023, pp. 1–5.

[13] D. Xin, S. Adavanne, F. Ang, A. Kulkarni, S. Takamichi, and H. Saruwatari, "Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts," in *ICASSP*, 2023, pp. 1–5.

[14] Y.-J. Zhang, W. Song, Y. Yue, Z. Zhang, Y. Wu, and X. He, "MaskedSpeech: Context-aware Speech Synthesis with Masking Strategy," in *INTERSPEECH*, 2023, pp. 4803–4807.

[15] Y. Xiao, S. Zhang, X. Wang, X. Tan, L. He, S. Zhao, F. K. Soong, and T. Lee, "ContextSpeech: Expressive and Efficient Text-to-Speech for Paragraph Reading," in *INTERSPEECH*, 2023, pp. 4883–4887.

[16] D. Guo, X. Zhu, L. Xue, T. Li, Y. Lv, Y. Jiang, and L. Xie, "Hignn-tts: Hierarchical prosody modeling with graph neural networks for expressive long-form tts," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[17] S. Lei, Y. Zhou, L. Chen, Z. Wu, X. Wu, S. Kang, and H. Meng, "Msstyletts: Multi-scale style modeling with hierarchical context information for expressive speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3290–3303, 2023.

[18] X. Chen, X. Wang, S. Zhang, L. He, Z. Wu, X. Wu, and H. Meng, "Stylespeech: Self-supervised style enhancing with vq-vae-based pre-training for expressive audiobook speech synthesis," *arXiv:2312.12181*, 2023.

[19] D. Yang, T. Koriyama, Y. Saito, T. Saeki, D. Xin, and H. Saruwatari, "Duration-aware pause insertion using pre-trained language model for multi-speaker text-to-speech," in *ICASSP*, 2023, pp. 1–5.

[20] G. Bailly and C. Gouvernayre, "Pauses and respiratory markers of the structure of book reading," in *INTERSPEECH*, 2012, pp. 2218–2221.

[21] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis," *INTERSPEECH*, p. 3126–3120, 2021.

[22] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[23] "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005.

[24] 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.

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *34th Conference on Neural Information Processing Systems(NeurIPS)*, pp. 1877–1901, 2020.

[26] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 320–335.

[27] S. Zhao, Z. Li, Y. Lu, A. Yuille, and Y. Wang, "Causal-cog: A causal-effect look at context generation for boosting multi-modal language models," *arXiv:2312.06685*, 2023.

[28] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E.-S. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] Y. Saito, S. Takamichi, E. Iimori, K. Tachibana, and H. Saruwatari, "ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-derived Context Word Embeddings," in *INTERSPEECH*, 2023, pp. 3048–3052.

[30] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *ICLR*, 2021.

[31] X. Ma, C. Zhou, X. Kong, J. He, L. Gui, G. Neubig, J. May, and L. Zettlemoyer, "Mega: Moving average equipped gated attention," in *ICLR*, 2022.

[32] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv:2107.02137*, 2021.

[33] C. Xu, K. Shen, and H. Sun, "A global context mechanism for sequence labeling," *arXiv:2305.19928*, 2023.

[34] J. Kong, J. Kim, and J. Bae, "Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis," in *34th Conference on Neural Information Processing Systems(NeurIPS)*, 2020, pp. 17 022–17 033.

[35] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *ICLR*, 2014.