Unsupervised Multi-scale Expressive Speaking Style Modeling with Hierarchical Context Information for Audiobook Speech Synthesis

Xueyuan Chen[*]
Shenzhen International
Graduate School
Tsinghua University
Shenzhen, China
chenxuey20@mails.tsinghua.edu.cn

Shun Lei
Shenzhen International
Graduate School
Tsinghua University
Shenzhen, China
leis21@mails.tsinghua.edu.cn

Zhiyong Wu[†]
Shenzhen International
Graduate School
Tsinghua University
Shenzhen, China
zywu@se.cuhk.edu.hk

Dong Xu
Tencent Music
Entertainment Group
Shenzhen, China
huberyxu@tencent.com

Weifeng Zhao
Tencent Music
Entertainment Group
Shenzhen, China
ethanzhao@tencent.com

Helen Meng
The Chinese
University of Hong Kong
Hong Kong SAR, China
hmmeng@se.cuhk.edu.hk

Abstract

Naturalness and expressiveness are crucial for audiobook speech synthesis, but now are limited by the averaged global-scale speaking style representation. In this paper, we propose an unsupervised multi-scale context-sensitive text-to-speech model for audiobooks. A multi-scale hierarchical context encoder is specially designed to predict both global-scale context style embedding and local-scale context style embedding from a wider context of input text in a hierarchical structure. Likewise, a multi-scale reference encoder is introduced to extract reference style embeddings at both global and local scales from the reference speech, which are used to guide the prediction of speaking styles. On top of these, a bi-reference attention mechanism is used to align both local-scale reference style embedding sequence and local-scale context style embedding sequence with corresponding phoneme embedding sequence. Both objective and subjective experiment results on a real-world multi-speaker Mandarin novel audio dataset demonstrate the excellent performance of our proposed method over all baselines in terms of naturalness and expressiveness of the synthesized speech[1].

1 Introduction

Recently, some text-to-speech (TTS) models, such as Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), Deep voice (Ping et al., 2017), TransformerTTS (Li et al., 2019) have been proposed to generate speech autoregressively from text input, and can achieve performance very close to human quality. In order to increase inference speed and generate more robust speech, non-autoregressive TTS models such as FastSpeech (Ren et al., 2019) and FastSpeech 2 (Ren et al., 2020) are emerged with robust and fast parallel generation.

However, limited expressiveness of synthesized audio persists as one of the major gaps between synthesized speech and real human speech, which draws growing attention to expressive speech synthesis studies. Synthesizing long-form expressive datasets (such as audiobooks) is still a challenging task, since wide-ranging voice characteristics are collapsed into an averaged prosodic style. To address this issue, style transfer TTS has been a popular strategy in recent years (Skerry-Ryan et al., 2018). The global style token (GST) model (Wang et al., 2018) adopts multi-head attention mechanism and several learnable style tokens to extract the global style from reference audio in an unsupervised way. Further more, a hierarchical GST architecture is proposed to learn hierarchical embedding information implicitly, which contains several GST layers with residual connection (An et al., 2019).

[*] Work conducted when the first author was intern at Tencent Music Entertainment Group.

[†] Corresponding author.

[1] Synthesized speech samples are available at: https://thuhcsi.github.io/COLING2022-MSHCE-TTS

To not only learn to represent a wide range of speaking styles, but also synthesize expressive speech without the need of auxiliary inputs at inference time, some methods attempt to predict speaking style directly from text, which is more practical and flexible. The text-predicted global style token (TP-GST) extends the GST by predicting style embedding or style token weights from text only (Stanton et al., 2018). Considering that style and semantic information of sentences are closely related and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) shows its effectiveness in language representation, pre-trained BERT embedding is adopted as an additional contextual information input to Tacotron 2 model to improve the pronunciation and expressiveness of the generated speech (Hayashi et al., 2019) (Fang et al., 2019). In addition, different ways of incorporating linguistic features and BERT-based features are investigated and compared in various application domains (news, chat and audiobooks). Results show that character embedding is the most effective one (Xiao et al., 2020). Some works also attempt to predict fine-grained speaking styles from text, such as word level (Zhang and Ling, 2021) and phoneme level (Lei et al., 2021).

All aforementioned methods only take the single sentence to be synthesized into consideration. Some studies demonstrate that considering a wider range of contextual information contributes to expressive speech synthesis (Tan et al., 2021) (Li et al., 2022). Recently, a hierarchical context encoder that considers adjacent contexts within a fixed-size sliding window is used to predict sentence-level style representation directly from text (Lei et al., 2022a). Although the overall performance is improved, it is still an averaged result that lacks some local fine-grained expressiveness information and rhythmic fluctuations such as pauses and emphasis.

The expressiveness of human speech can be perceived as a compound of multi-scale acoustic factors. One is the global-scale speech style, which includes but not limited to timbre and emotion of the speaker. Styles at this level are supposed to be consistent throughout the entire utterance. The other is the local-scale speech style, which consists of speed, energy, pitch, pause and other acoustic features (Li et al., 2021). Therefore, it is insufficient to model speech style from a single aspect. Some latest studies are observed to devote efforts to performing multi-scale modeling on some specific tasks, in particular emotional speech synthesis (Lei et al., 2021) (Lei et al., 2022b). This further demonstrates the importance and necessity of multi-scale modeling.

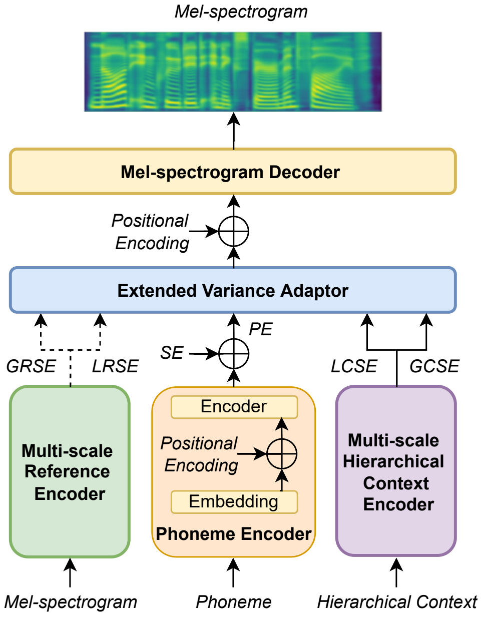


Figure 1: The overall architecture of the proposed model

With all listed imperfections taken into consideration, this paper proposes an unsupervised multi-scale expressive speech synthesis model for audiobooks with hierarchical context information as input. A multi-scale hierarchical context encoder is designed to predict both global-scale context style embedding (GCSE) and local-scale context style embedding (LCSE) from the context in a hierarchical structure. Meanwhile, a multi-scale reference encoder is introduced to extract both global-scale reference style embedding (GRSE) and local-scale reference style embedding (LRSE) from the reference speech, which are used to guide the prediction of speaking styles. On top of these, a bi-reference attention mechanism is adopted to align both the quasi-phoneme-level LRSE sequence and the character-level LCSE sequence with the corresponding phoneme embedding (PE) sequence. Both objective and subjective experiments on a real-world multi-speaker Mandarin novel audio dataset demonstrate the excellent performance of our proposed model over all baseline approaches in terms of naturalness and expressiveness of the synthesized speech. Ablation studies are further conducted to investigate the influences of several main modules in our proposed model.

2 Method

The architecture of our proposed model is illustrated in Figure 1. It consists of three major parts: a multi-scale reference encoder, a multi-scale hierarchical context encoder and a sequence-to-sequence expressive TTS system based on Fastspeech 2 (Ren et al., 2020) with extended variance adaptor. The multi-scale reference encoder is used to extract reference style embeddings at both global and local scales (i.e. GRSE and LRSE) from reference speech. While the multi-scale hierarchical context encoder is used to predict the context style embeddings at global and local styles (i.e. GCSE and LCSE) from hierarchical context. The extended variance adaptor is used to align and fuse the style embeddings at different scales with the phoneme embeddings.

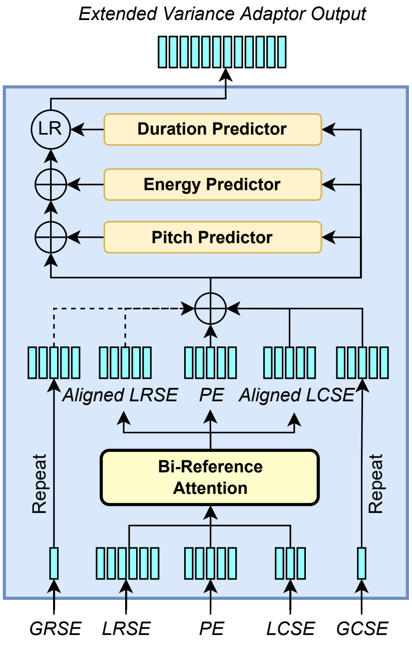


Figure 2: Extended Variance Adaptor

2.1 Multi-scale expressive TTS system

We adopt FastSpeech 2 as the basic acoustic model, of which the phoneme encoder and mel-spectrogram decoder keep the original structure as described in (Ren et al., 2020). On the basis, speaker embedding (SE) is added to the phoneme encoder output to support different timbres. After that, the phoneme embedding (PE) together with GRSE, LRSE, GCSE, LCSE are fed into the extended variance adaptor with several changes as illustrated in Figure 2.

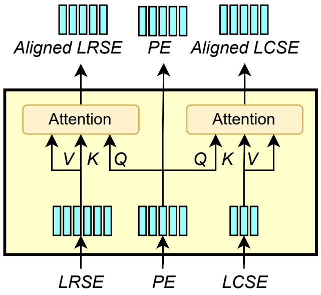


Figure 3: Bi-Reference Attention

2.1.1 Extended variance adaptor

Firstly, the global-scale style inputs of GRSE and GCSE are repeated to the same length as PE. At the same time, the local-scale style inputs of LRSE and LCSE are aligned to phoneme-level sequences by a bi-reference attention mechanism. After that, either the repeated GRSE and aligned LRSE or the repeated GCSE and aligned LCSE are added to PE at different stages, and then the result is passed to the variance predictors containing duration predictor, pitch predictor and energy predictor. Unlike FastSpeech 2, the length regulator is moved after the variance predictors, in order to predict variations at phoneme level rather than frame level, which has been proved to be able to further improve speech quality (Łańcucki, 2021).

2.1.2 Bi-reference attention mechanism

For expressive speech synthesis, fine-grained style embedding sequence is usually reformed into sequence with the same length as phoneme embedding sequence. Inspired by (Lee and Kim, 2019), we propose a bi-reference attention mechanism to align both the quasi-phoneme-scale LRSE sequence and character-scale LCSE sequence with the phoneme-level PE sequence. As shown in Figure 3, the bi-reference attention consists of two scaled dot-product attentions (Vaswani et al., 2017) with the same query input and two groups of different key and value inputs. Here, the phoneme-level PE sequence is fed as the query of the bi-reference attention. Meanwhile, the quasi-phoneme-scale LRSE sequence and the character-scale LCSE sequence are fed as two groups of key and value inputs respectively. Finally, the bi-reference attention outputs the aligned LRSE sequence and aligned LCSE sequence with the same length as the PE sequence. This operation not only could align sequences of different lengths to the phoneme level, but also reduces the difficulty of local style guidance from the multi-scale reference encoder to the multi-scale hierarchical context encoder.

2.2 Multi-scale reference encoder

Inspired by the success of multi-scale emotion transfer task (Li et al., 2021), we introduce a multi-scale reference encoder to extract both the global-scale and local-scale style embeddings from the reference speech. As shown in Figure 4, the multi-scale reference encoder is made up by a stack of 6 convolution layers and 2 scale-specific extractor layers.

For the convolution layers, 1-D convolution along temporal dimension is adopted to help the bi-reference attention to learn the alignment between the output LRSE sequence and the PE sequence. Each of the convolution layers is composed by 3×1 filters, ReLU activation and batch normalization (Ioffe and Szegedy, 2015). In particular, in order to regulate the temporal granularity of the convolution output closer to human vocal perception, the filter strides of 6 convolution layers are set as [2, 1, 2, 1, 2, 2]. This downsampling operation ensures that after the convolution stack, temporal granularity of the intermediate frame-level feature sequence is properly reformed to a quasi-phoneme-scale.

For the global style extractor layer, it is made up of consecutive Gated Recurrent Unit (GRU) layer and global style token (GST) layer (Wang et al., 2018). The GST layer adopts a multi-head attention mechanism and several learnable style tokens to extract the global styles in an unsupervised way. For the local style extractor layer, it consists of GRU layer and full-connected layer with tanh activation. Both scale-specific extractor layers take the above quasi-phoneme-scale feature sequence as input. However, inside the global style extractor layer, only the final state of GRU is fed to the GST layer. The GRSE output by the global style extractor is forced to be a latent sentence-level style embedding vector. On the other hand, the LRSE output by the local style extractor is a quasi-phoneme-scale style embedding sequence.

2.3 Multi-scale hierarchical context encoder

To improve expressiveness and naturalness of synthesized speech, we introduce a dedicated multi-scale hierarchical context encoder to predict both global-scale style embedding and local-scale style embedding from the hierarchical context within a fixed size sliding window. As shown in Figure 5, the multi-scale hierarchical context encoder consists of three components: context embedding layer, global style predictor and local style predictor.

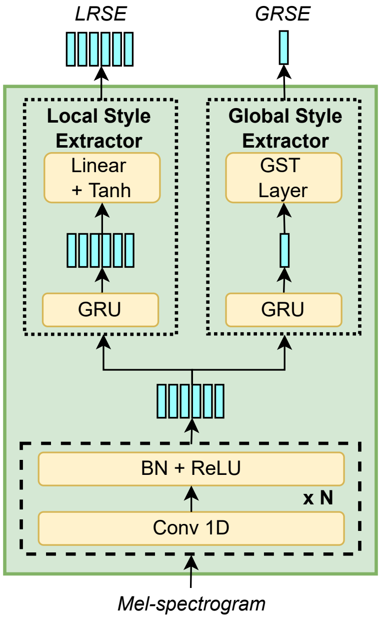


Figure 4: Multi-scale Reference Encoder

2.3.1 Context embedding layer

Let l be the number of sentences considered in the past or future context within the sliding window. U_0 is defined as the current sentence to be synthesized. U_{-l}, U_{1-l}, ..., U_{-1} and U_1, U_2, ..., U_l are the past and future sentences respectively. All these $2l + 1$ sentences are firstly embedded with a well-pretrained character-level BERT model (Devlin et al., 2018) that is composed of a stack of Transformer blocks and pretrained with a huge amount of Chinese text data. Thereafter, role embedding is added to the output of BERT embedding layer to consider the interactions between different roles. Then Bidirectional GRU (BiGRU) is further used to obtain the character-level context-sensitive embedding S_i for each input sentence U_i $(-l \leq i \leq l)$, which can be discribed as:

$$S_i = CEmb(U_i), \tag{1}$$

where $CEmb(\cdot)$ is the operation of context embedding layer.

2.3.2 Global style predictor

The global style predictor contains two levels of attention networks, intra-sentence and inter-sentence respectively.

The intra-sentence attention network is used to abtain a sentence-level representation based on each character and inter-character relations within a sentence. As not all characters contribute equally to the global meaning of the sentence, a scaled dot-product attention is adopted to calculate the weights of each character and aggregate them into a global sentence-level vector G_i for each character-level embedding S_i of a sentence, where S_i is fed as the key and value. The query Q_1 is a 256-dim vector, which is randomly initialized and learnable during the training. It can be seen as a high level representation of a fixed query "to what extent does the character influence the global speaking style". This can be formulated as:

$$K_i = S_i W_k, \tag{2}$$

$$V_i = S_i W_v, \tag{3}$$

$$G_i = A(Q_1, K_i, V_i) = softmax(\frac{Q_1 K_i^\top}{\sqrt{d_{Q_1}}})V_i, \tag{4}$$

where W_k and W_v are linear projection matrices of attention keys and values. d_{Q_1} means the dimension of the query Q_1.

For the inter-sentence attention network, $2l + 1$ sentence-level vectors $G_{-l}, ..., G_0, ..., G_l$ obtained by the intra-sentence network are firstly concatenated along temporal dimension to form a long new sequence G with the length of $2l + 1$, which is further fed to BiGRU to model the correlations among sentences. After that, another scaled dot-product attention is used to predict a global-scale speaking style based on sentence-level embedding G and inter-sentence relations within the hierarchical context. Here, G is the key and value, while query Q_2 is a randomly initialized and learnable 256-dim vector similar to Q_1. Finally, the inter-sentence attention network outputs the GCSE of the current sentence U_0.

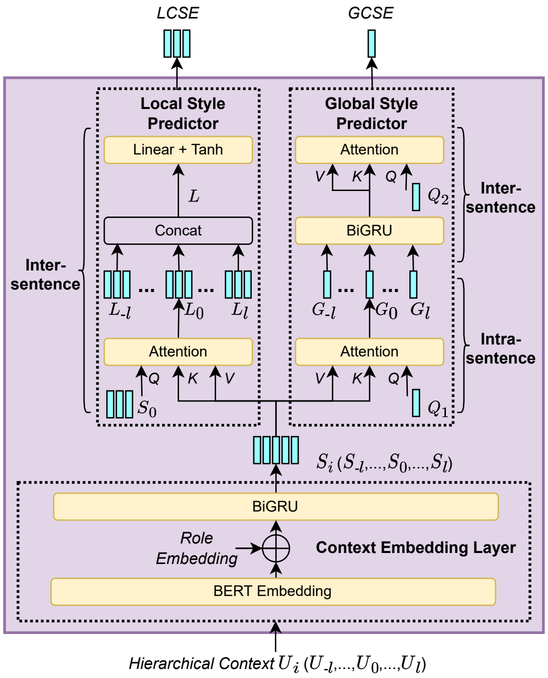


Figure 5: Multi-scale Hierarchical Context Encoder

2.3.3 Local style predictor

The local style predictor is used to obtain the local-scale style embedding of current sentence from the hierarchical context embeddings.

Firstly, a scaled dot-product attention is used to align each character-level embedding S_i of a sentence in the hierarchical context with the character-level embedding S_0 of the current sentence, where S_i is fed as the key and value,S_0 is fed as the query. Here, S_0 can be seen as a fixed local-scale query "to what extent does each character influence the local speaking style of the current sentence" for each sentence in the context. This can be formulated as:

$$K_i = S_i W_k, \tag{5}$$

$$V_i = S_i W_v, \tag{6}$$

$$L_i = A(S_0, K_i, V_i) = softmax(\frac{S_0 K_i^\top}{\sqrt{d_{S_0}}})V_i, \tag{7}$$

where W_k and W_v are linear projection matrices of attention keys and values. d_{S_0} means the dimension of the query S_0.

Since all the $2l + 1$ attention outputs $L_{-l}, ..., L_0, ..., L_l$ have the same length as S_0, a concatenation operation along feature dimension is further implemented, followed by full-connected layer with tanh activation:

$$L = Concat_f(L_{-l}, ..., L_l), \tag{8}$$

In this way, the LCSE of current sentence is obtained.

2.4 Training strategy and inference procedure

During training, to encourage the multi-scale hierarchical context encoder learn style representation better, the proposed model is trained with knowledge distillation strategy in three stages.

i) In the first stage, the acoustic model and the multi-scale reference encoder are jointly trained with paired <utterance, speech> data to get a well-trained multi-scale reference encoder in an unsupervised way. In order to better extract style features at different scales, this training stage is divided into two steps. Firstly, only the global style extractor is inserted into the model to obtain the well-represented global-scale features. After that, the local style extractor is further involved, which can extract more fine-grained local style representations. The multi-scale style embeddings extracted from all speeches in the training set can be regarded as ground-truth speaking style representations.

ii) In the second stage, knowledge distillation strategy is used to transfer the knowledge from the multi-scale reference encoder to the multi-scale hierarchical context encoder. That is, we use ground-truth style embeddings GRSE and aligned LRSE as targets to guide the prediction of speaking style representations GCSE and aligned LCSE from context, for training the multi-scale hierarchical context encoder.

iii) In the third stage, the acoustic model and the multi-scale hierarchical context encoder are jointly trained with a lower learning rate to further improve the expressiveness of the synthesized speech.

During inference, the multi-scale reference encoder is abandoned. Only GCSE and LCSE predicted from the multi-scale hierarchical context encoder are fed to variance adaptor together with PE. Finally, by accepting input text and hierarchical context, the model can synthesize speech with more expressive styles.

3 Experiment

3.1 Dataset and model details

An internal multi-speaker novel audio corpus on Mandarin is employed in our experiment. It contains more than 40 roles and around 15 hours speech spoken by 5 Mandarin native speakers with quite different timbres. The speaking styles vary among roles and utterances, and the speed, pitch and energy fluctuate greatly in an utterance. The dataset has a total of 11,000 audio clips, of which 200 clips are used for validation and 100 clips for test, and the rest for training.

For feature extraction, we transform the raw waveforms into 80-dim mel-spectrograms with sampling rate 16kHz, frame size 1200 and hop size 240. An open-source pre-trained Chinese character-level BERT model[2] with frozen parameters is used in our experiments. The context of current sentence is made up of its two past sentences, two future ones, and itself.

We take 200k steps to train the acoustic model and multi-scale reference encoder, where 100k steps are for global style extractor and the remaining 100k steps for local style extractor. Then we take 20k steps to train the multi-scale hierarchical context encoder and 20k steps to adapt the acoustic model and the multi-scale hierarchical context encoder. All the trainings are conducted with a batch size of 16 on a NVIDIA A100 GPU. The Adam optimizer is adopted with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The warm-up strategy is used before 4000 steps. In addition, a well-trained HiFi-GAN (Kong et al., 2020) is used as the vocoder to generate waveform.

3.2 Compared methods

Three FastSpeech 2 based models are implemented for comparison, and the details are described as follows:

FastSpeech 2: Original FastSpeech 2 (Ren et al., 2020) with minor changes on the variance predictor to be consistent to the proposed model as described in section 2.1.1.

BERT-FS 2: Inspired by (Xiao et al., 2020), we set an end-to-end TTS model by combining BERT with FastSpeech 2, which contains a plain context encoder and only considers the current sentence. The same character embeddings obtained from BERT are directly passed to a GRU layer whose final state is used as a style embedding.

HCE-FS 2: It uses a reference encoder to extract the global style representation, and uses a hierarchical context encoder to predict the global style embedding, which is fed to the variance adaptor of FastSpeech 2 (Lei et al., 2022a).

Model	**S-MOS**	**P-MOS**
Ground Truth	4.705 ± 0.067	4.737 ± 0.073
FastSpeech 2	3.426 ± 0.091	3.432 ± 0.099
BERT-FS 2	3.503 ± 0.096	3.526 ± 0.086
HCE-FS 2	3.589 ± 0.089	3.613 ± 0.073
Proposed	$\mathbf{4.031 \pm 0.068}$	$\mathbf{4.142 \pm 0.071}$

Table 1: The sentence-MOS (S-MOS) and paragraph-MOS (P-MOS) of different models with 95% confidence intervals for subjective evaluation.

[2] https://huggingface.co/bert-base-chinese

Model	F0 RMSE	Energy RMSE	Duration MSE	MCD
FastSpeech 2	70.847	13.228	0.1316	7.198
BERT-FS 2	67.912	13.031	0.1263	7.134
HCE-FS 2	64.975	12.449	0.1254	6.975
Proposed	**62.471**	**11.683**	**0.1224**	**6.843**

Table 2: Objective evaluations for different models.

3.3 Subjective evaluation

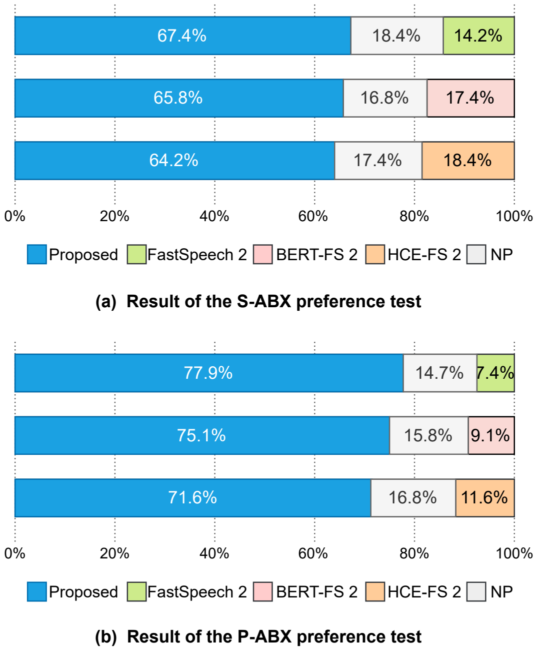
We conduct the mean opinion score (MOS) tests to evaluate the naturalness and expressiveness of the generated speeches. As the task in this paper is paragraph-level audiobook speech synthesis, we conduct two kinds of MOS tests, sentence-MOS (S-MOS) and paragraph-MOS (P-MOS) respectively. The former mainly focuses on the naturalness and expressiveness of the synthesized speech considering only the current single sentence. The latter focuses on the coherence of speech styles of the current sentence in a paragraph considering the context of past and future sentences, where the speeches of the past and future sentences are the resynthesized version of ground truth speech recordings. 10 single sentences and 10 short paragraphs are randomly selected in the test set. 25 native Chinese speakers are asked to listen to the generated speeches and rate on a scale from 1 to 5 with 1 point interval.

As shown in Table 1, our proposed approach achieves the best S-MOS of **4.031** and P-MOS of **4.142**. The results demonstrate the effectiveness of our proposed methods over all the baselines especially on the paragraph level. There is a big gap between **FastSpeech 2** and **Ground Truth**, indicating that it is difficult to model the multiple speech variations without enough input information.

The ABX preference tests are also conducted on our proposed model and each of the three baselines respectively. Similarly, we also conduct two kinds of ABX preference tests, sentence-ABX (S-ABX) and paragraph-ABX (P-ABX) respectively. The same 25 subjects are asked to choose a preferred speech in terms of naturalness and expressiveness between a pair of methods.

As shown in Figure 6, the preference rate of our proposed model exceeds **FastSpeech 2** by 53.2%, **BERT-FS 2** by 48.4% and **HCE-FS 2** by 45.8% on the S-ABX preference test. Moreover, the gaps between our proposed model and baseline models are more reflected on the preference rate of P-ABX, which are 70.5%, 66.0% and 60.0% respectively.

Both MOS and ABX preference tests demonstrate that our proposed method significantly outperforms the baselines in terms of naturalness and expressiveness especially for the paragraph-level speech synthesis tasks.



(a) Result of the S-ABX preference test

(b) Result of the P-ABX preference test

Figure 6: Results of the sentence-ABX (S-ABX) and paragraph-ABX (P-ABX) preference tests. NP means no preference.

3.4 Objective evaluation

As the common for the objective evaluation of synthesized speech, we employ the root mean square error (RMSE) of pitch and energy, the mean square error (MSE) of duration and mel cepstral distortion (MCD) as the objective evaluation metrics. Specifically, the dynamic time warping (DTW) is firstly used to construct the alignment paths between the ground-truth mel-spectrogram and the predicted one. After that, the F0 sequence and energy sequence are aligned towards ground-truth following the DTW path. We also utilize DTW to compute the minimum MCD by aligning the two sequences. Here, MCD is utilized to calculate the difference between the mel-spectrograms of the synthesized speech and the ground truth. For duration, we compute the MSE between the predicted duration and ground-truth duration.

As shown in Table 2, our proposed model achieves **62.471** for F0 RMSE, **11.683** for Energy RMSE, **0.1224** for Duration MSE and **6.843** for MCD, which outperforms all the baselines on all metrics. This excellent results indicate that our proposed model can predict more accurate style features, such as duration, pitch and energy, than baselines.

3.5 Ablation study

To further investigate the influence of several main modules in our proposed model, we have tried four other settings based on the proposed method:

i) **Proposed - Global-scale Style**: The global style extractor in multi-scale reference encoder and the global style predictor in multi-scale hierarchical context encoder are removed, and only LRSE, LCSE together with PE are fed to variance adaptor.

ii) **Proposed - Local-scale Style**: The local style extractor in multi-scale reference encoder and the local style predictor in multi-scale hierarchical context encoder are removed, and only GRSE, GCSE together with PE are fed to variance adaptor.

iii) **Proposed - Knowledge Distillation**: The knowledge distillation strategy is abandoned, by removing the multi-scale reference encoder. The predicted GCSE and LCSE from multi-scale hierarchical context encoder together with PE are fed to variance adaptor directly throughout the training process.

iv) **Proposed - Role Embedding** The role embedding of context embedding layer in the multi-scale hierarchical context encoder is removed.

Comparison mean opinion score (CMOS) is employed to compare the synthesized speeches in terms of naturalness and expressiveness. The results are shown in Table 3.

Model	**CMOS**
Proposed	0
- Global-scale Style	-0.174
- Local-scale Style	-0.211
- Knowledge Distillation	-0.189
- Role Embedding	-0.153

Table 3: CMOS comparision for ablation study.

Compared with the proposed method, the performance of the four settings removing different main modules is degraded to various degrees respectively. This indicates that all these components have substantial impact on our proposed model. When the global-scale style is removed, the overall style of a sentence lacks expressiveness and there will be a obvious deviation from the ground truth. The proposed model without local-scale style leads to significant performance degradation in some rhythmic aspects, especially pauses and stress. When the knowledge distillation strategy is abandoned, the model needs to predict both the global-scale and local-scale style from context directly without any guidance from the reference encoder. The performance degradation demonstrates the necessity of the multi-scale reference encoder and indicates that learning the speaking style representation from context in an explicit way is more suitable for this style prediction task. The proposed model without the role embedding also causes some performance degradation, which further indicates that considering the interactions between different roles is crucial for the style prediction of novel speech synthesis.

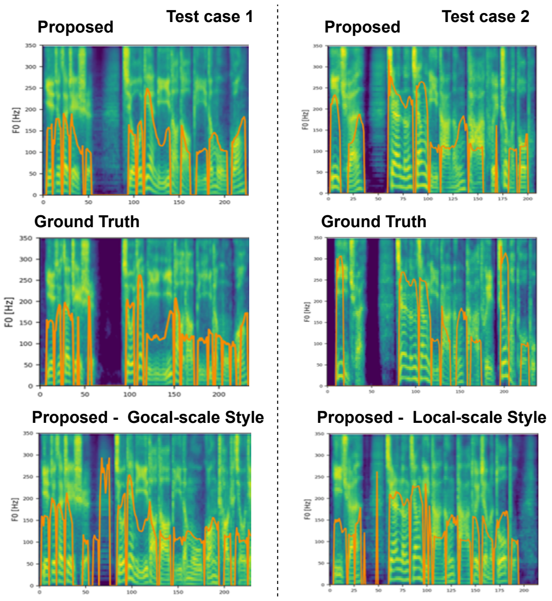


Figure 7: Mel-spectrograms and pitch contours of the speeches for the two test cases.

3.6 Case study

To further explore the impact of multi-scale speaking style on the expressiveness and naturalness of synthesized speech, a case study is conducted to synthesize two example utterances in test set with the operation of removing the global-scale style and local-scale style respectively. The speech synthesized by our proposed model and the ground truth are also provided for reference. The mel-spectrograms and pitch contours of speeches are shown in Figure 7.

When the global-scale style is removed, the pitch fluctuations become larger than others, and the overall style and speed of synthesized speech vary greatly compared with the ground truth. When the local-scale style is removed, some local style characteristics of the synthesized speech are lost, resulting in a relatively averaged style throughout the whole sentence. Compared with the two single-scale results, the speeches synthesized by our proposed multi-scale model are more similar to the ground-truth speech in terms of the overall and fine-grained style properties, such as the trend of intonation and stress patterns. The results demonstrate that modelling the speaking style from hierarchical context information in a multi-scale way is essential and effective to improve the naturalness and expressiveness of the synthesized speech.

4 Conclusion

In this paper, we propose an unsupervised multi-scale context-sensitive text-to-speech model for audiobooks. A multi-scale hierarchical context encoder is designed to predict both the global-scale context style embedding and local-scale context style embedding from hierarchical context with the guidance of a multi-scale reference encoder. Both objective and subjective experiment results on a real-world multi-speaker Mandarin novel audio dataset demonstrate the excellent performance of our proposed multi-scale context-sensitive model over all baseline approaches in terms of naturalness and expressiveness of the synthesized speech.

Acknowledgement: This work is supported by National Natural Science Foundation of China (62076144), the Major Key Project of PCL (PCL2021A06, PCL2022D01), Shenzhen Key Laboratory of next generation interactive mediainnovative technology (ZDSYS20210623092001004), CCF-Tencent Open Research Fund (CCF-Tencent RAGR20210122) and Tencent AI Lab Rhino-Bird Focused Research Program (RBFR2022005).

References

Xiaochun An, Yuxuan Wang, Shan Yang, Zejun Ma, and Lei Xie. 2019. Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 184–191. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wei Fang, Yu-An Chung, and James Glass. 2019. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. *arXiv preprint arXiv:1906.07307*.

Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu. 2019. Pre-trained text embeddings for enhanced text-to-speech synthesis. In *INTERSPEECH*, pages 4430–4434.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Younggun Lee and Taesu Kim. 2019. Robust and fine-grained prosody control of end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE.

Shun Lei, Yixuan Zhou, Liyang Chen, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2022a. Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022b. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yi Lei, Shan Yang, and Lei Xie. 2021. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 423–430. IEEE.

Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. 2022. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7917–7921. IEEE.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 6706–6713.

Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia, and Helen Meng. 2021. Towards multi-scale style control for expressive speech synthesis. *arXiv preprint arXiv:2104.03521*.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: 2000-speaker neural text-to-speech.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan. 2018. Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 595–602. IEEE.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR.

Yujia Xiao, Lei He, Huaiping Ming, and Frank K Soong. 2020. Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708. IEEE.

Ya-Jie Zhang and Zhen-Hua Ling. 2021. Extracting and predicting word-level style variations for speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1582–1593.