Towards Multi-Scale Speaking Style Modelling with Hierarchical Context Information for Mandarin Speech Synthesis

Shun Lei†‡, Yixuan Zhou†, Liyang Chen†, Jiankun Hu2, Zhiyong Wu†,‡, Shiying Kang2, Helen Meng4

1 Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
2 Huya Inc., Guangzhou, China
3 XVerse Inc., Shenzhen, China
4 The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract

Previous works on expressive speech synthesis focus on modelling the mono-scale style embedding from the current sentence or context, but the multi-scale nature of speaking style in human speech is neglected. In this paper, we propose a multi-scale speaking style modelling method to capture and predict multi-scale speaking style for improving the naturalness and expressiveness of synthetic speech. A multi-scale extractor is proposed to extract speaking style embeddings at three different levels from the ground-truth speech, and explicitly guide the training of a multi-scale style predictor based on hierarchical context information. Both objective and subjective evaluations on a Mandarin audiobooks dataset demonstrate that our proposed method can significantly improve the naturalness and expressiveness of the synthesized speech.

Index Terms: text-to-speech, expressive speech synthesis, speaker style modelling, multi-scale, BERT

1. Introduction

Text-to-speech (TTS) aims to generate intelligible and natural speech from text. With the development of deep learning, TTS models are now endowed with the ability to synthesize high-quality speech with a neutral speaking style [1–3]. However, the speaking style with limited expressiveness remains a clear gap between synthesized speeches and human recordings, which blocks the development of speech synthesis technology in many application scenarios such as audiobooks, podcasts, and voice assistants. Therefore, how to model expressive speaking style is a hot research topic in academia and industry recently.

One of the general approaches is to extract the speaking style representation from given reference audio [4–7], which synthesizes speech conditioned on the extracted representation. Compared with the reference audio-based methods, another line that directly predicting the speaking style from text without auxiliary inputs is more practical and flexible. The text-predicted global style token (TP-GST) model [8] is proposed to predict the global-level style representation from text alone. Benefiting from the great semantic representation ability of the pre-trained language models, such as bidirectional encoder representations from Transformer (BERT) [9], text representations derived from the pre-trained language model have been used to predict speaking style and shown gains in performance [10, 11].

Some early text-predicted methods concentrate only on the current sentence, which fails to capture the style information influenced by the different context of neighbor sentences [12, 13].

To avoid this problem, [14] proposes to use the neighbor sentences to improve the prosody generation. Our preliminary work [15] utilizes the hierarchical context encoder (HCE) to further consider the hierarchical structure of context and predicts the global-scale speaking style in an explicit way. These studies demonstrate that taking a wider range of contextual information into account is helpful for expressive speech synthesis.

However, HCE still suffers from the absence of local-scale style modeling (e.g., intonation, rhythm, stress). To model and control local prosodic variations in speech, some previous works attempt to predict finer-grained speaking styles from text, such as word level [16, 17] and phoneme level [18, 19]. It is more widely accepted that the style expressions of human speech are multi-scale in nature [20, 21], where the global-scale style is usually observed as emotion and the local-scale is more close to the prosody variation [22, 23]. These styles from different levels work together to produce rich expressiveness in speech. Towards this, some latest researches on similar tasks, such as emotional speech synthesis [24, 25] and style transfer [22], devote effort to performing a multi-scale style modelling which however require auxiliary labels besides text. To our best knowledge, there is currently no work investigating on multi-scale speaking style prediction just from context.

In this paper, we propose a multi-scale speaking style modelling method to capture and predict multi-scale speaking style from hierarchical context information for expressive TTS. Our model contains a multi-scale style extractor, a multi-scale style predictor and a FastSpeech 2-based acoustic model. The extractor is used to extract style embeddings at global level, sentence level and subword level from the ground-truth speech, and to explicitly guide the training of the multi-scale style predictor. The predictor is based on HCE, and we exploit the hierarchical context information of HCE in a more efficient way to predict style embeddings at above three levels. To reduce the interference or overlapping between speaking styles at different levels, residual style embedding is introduced to represent effective style variations in speech. Both subjective and objective evaluations on a Mandarin audiobook dataset demonstrate that the proposed method can improve the naturalness and expressiveness of generated speech, benefiting from its ability to accurately predict both global-scale and local-scale speaking styles from context.

2. Methodology

The architecture of our proposed model is illustrated in Fig.1. It can be mainly divided into three parts, a multi-scale style extractor, a multi-scale style predictor and a FastSpeech 2 [3] based acoustic model. The acoustic model predicts the mel-spectrogram of the current sentence with the assistance of the

---

1 Work conducted when the first author was intern at Huya Inc.
2 Equal contribution. * Corresponding author.
3 Speech samples: https://thuhcsi.github.io/interspeech2022-msc-tts

---
extractor or the predictor. The extractor is used to extract the style embeddings at three different levels, and the predictor is used to predict these style embeddings from the context.

2.1. Multi-Scale Style Extractor

To extract the multi-scale style embedding from the reference speech, we specifically design a multi-scale style extractor, as shown in Fig.2. The reference encoders and style token layers corresponding to three different levels make up this module. All the reference encoders and style token layers have the same architecture and hyperparameters as those of the GST [5].

Let \( L \) be the number of sentences considered in the past and future context. The mel-spectrograms corresponding to all \( 2L + 1 \) sentences are concatenated and passed to the global reference encoder to extract global reference embedding \( E_g \). The sentence reference encoder is used to extract sentence reference embedding \( E_s \) from the mel-spectrogram of the current sentence. Then the mel-spectrogram of the current sentence is divided by the subword boundaries which are obtained from the forced alignment phoneme boundaries and the subword-to-phoneme alignments. The mel-spectrogram of each subword goes through the style reference encoder and the output is denoted as a subword reference embedding \( E_w \). The lower-level embedding \( E_w \) may contain redundant style information which has already been covered in the higher-level embedding \( E_s \) and similarly for \( E_s \) and \( E_g \). To reduce such overlapping, the residuals between three reference embeddings are represent as the style variation, which can be described as:

\[
\begin{align*}
R_g &= E_g \\
R_s &= E_s - E_g \\
R_w &= E_w - E_s
\end{align*}
\]

where \( R_g, R_s \) and \( R_w \) is the residual style embedding of global-level, sentence-level and subword-level respectively.

The residual reference embeddings are passed to the corresponding style token layers to be decomposed into a fixed number of style tokens respectively, which helps memorize stylistic information at each level and reduce the difficulty of prediction. After style token layers, the global-level style embedding \( S_g \), sentence-level style embedding \( S_s \), and subword-level style embedding \( S_w \) are obtained. Finally, for each subword in the current sentence, the multi-scale style embedding is calculated as the summation of these three levels of style embeddings.

2.2. Multi-Scale Style Predictor

To better model the multi-scale speaking style, we extend the HCE in our preliminary work [15] and design a multi-scale style predictor. The lower-level style is derived by being conditioned on the higher-level style. This structure is symmetrical with the residual strategy in the style extractor.

The structure of the multi-scale style predictor is shown in Fig.3, which consists of the HCE and three extra style predictors. Each style predictor is composed of a linear layer and Tanh activation. Besides the current sentence, the multi-scale predictor also considers \( L \) sentences in the past or future. We firstly concatenate all the \( 2L + 1 \) sentences to form a new long text, and then pass them to a pre-trained BERT model to obtain the subword-level semantic embedding sequence. The HCE contains two levels of attention network, the inter-subword, and inter-sentence, each of them contains a bidirectional GRU [26] and a scaled dot-product attention module [27]. The bidirectional GRU is used to get the context embedding by considering temporal relationships, and the attention module is used to aggregate the context embedding sequence into a higher-level embedding. We denote the output of inter-subword level bidirectional GRU as subword context embedding \( C_{sw} \), the output of inter-sentence level bidirectional GRU as sentence context embedding \( C_s \), and the output of inter-sentence level attention module as global context embedding \( C_g \).

The higher-level style embedding that closer to the global-scale is firstly generated and then utilized as the conditional input to the lower-level style predictor. In this way, the style embeddings at three levels including \( S_g, S_s, \) and \( S_w \) are sequentially generated from the multi-scale style predictor. The training targets of the predictor come from the corresponding ground-truth style embeddings in the extractor. It is noteworthy that these three style embeddings of different scales attempt to restore the multi-scale speaking style in the human speech by considering the context information of different levels. Finally, these embeddings are added together to form the multi-scale style embedding of each subword in the current sentence.

2.3. Acoustic Model

As shown in Fig.1, the backbone of the proposed method is based on FastSpeech 2 [3]. The multi-scale style embedding of each subword in an utterance is provided by the multi-scale style extractor or the multi-scale style predictor. Then, accord-
Chinese subword-level BERT-base model\(^1\) is used in our experiments. The context of current sentence is made up of its two past sentences, itself and its two future sentences.

We train all the models for 220k iterations with a batch size of 16 on an NVIDIA V100 GPU. For our proposed model, we take 180k iterations to the first train step (60k iterations for each extractor), 20k iterations to the second train step and 20k iterations to the third train step. The Adam optimizer is adopted with \(\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}\) and the warm-up strategy is employed before 4000 iterations. In addition, we use a well-trained HiFi-GAN [29] as the vocoder to generate waveform.

### 3.2. Compared Methods

To demonstrate the performance of our proposed multi-scale model, three baseline models are implemented for comparison: **FastSpeech 2** An open-source implementation\(^2\) of FastSpeech 2 [3].

WSV* Word-level style variations (WSV) model. For a fair comparison, instead of Tacotron2 [1] used in the original version of WSV [16], FastSpeech 2 was adopted as the backbone in our implementation. In addition, an extra bidirectional GRU is used to consider the context information.

HCE Hierarchical context encoder (HCE) [15] model, which predicts the style on global-level from the context.

### 3.3. Subjective Evaluation

We conduct mean opinion score (MOS) test to evaluate the naturalness and expressiveness of the synthesized speech. 25 native Mandarin speakers are recruited as subjects to rate the given speeches on a scale from 1 to 5 with 1 point interval. As shown in Table 1, the results demonstrate the effectiveness of our proposed methods over the baselines. There exists a large gap between FastSpeech 2 and Ground Truth, indicating that it is difficult to model multiple speech variations without enough input information. Our proposed approach achieves the best MOS of 4.058, exceeding FastSpeech 2 by 0.485, WSV* by 0.377 and HCE by 0.273.

Table 1: The MOS on naturalness and expressiveness of different models with 95% confidence intervals.

| Model   | MOS         |
|---------|-------------|
| Ground Truth | 4.065 ± 0.074 |
| FastSpeech 2 | 3.573 ± 0.094 |
| WSV*     | 3.681 ± 0.080 |
| HCE      | 3.785 ± 0.084 |
| Proposed | 4.058 ± 0.074 |

ABX preference test is also conducted to ask subjects to give their preferences in terms of naturalness and expressiveness between a pair of speeches generated by different models. We compare the proposed model with each of the three baseline models. As shown in Fig.4, the preference rate of our proposed model exceeds FastSpeech 2 by 54.8%, WSV* by 38% and HCE by 26% respectively. Especially, some subjects report the speech synthesized by the proposed model has richer expressiveness than WSV*, and performs better than HCE on the local style properties, such as intonation and stress.

Both MOS and ABX preference tests demonstrate that our proposed approach significantly outperforms the three baselines in terms of naturalness and expressiveness. Compared with the

---

\(^1\)Available at: https://github.com/google-research/bert

\(^2\)Available at: https://github.com/ming024/FastSpeech2
In this paper, we propose a multi-scale speaking style modelling method to capture and predict multi-scale speaking style from hierarchical context information for expressive TTS. Experimental results demonstrate that our proposed method achieves better performance on expressive speech synthesis with the ability to predict both global-scale and local-scale speaking styles from context accurately.

5. Acknowledgements
This work is supported by National Key R&D Program of China (2020AAA0104500), National Natural Science Foundation of China (62076144), National Social Science Foundation of China (13&ZD189) and Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004).
6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Sherry-Ryan et al., “Natural its synthesis by conditioning wavenet on mel spectrum predictions,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4779–4783.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning,” in International Conference on Learning Representations, 2018.

[3] 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 International Conference on Learning Representations, 2020.

[4] R. Sherry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in International Conference on Machine Learning. PMLR, 2018, pp. 4693–4702.

[5] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in International Conference on Machine Learning. PMLR, 2018, pp. 5180–5189.

[6] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6945–6949.

[7] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019, pp. 623–627.

[8] D. Stanton, Y. Wang, and R. Sherry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 595–602.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[10] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, “Pre-trained text embeddings for enhanced text-to-speech synthesis,” in INTERSPEECH, 2019, pp. 4430–4434.

[11] Y. Xiao, L. He, H. Ming, and F. K. Soong, “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). IEEE, 2020, pp. 6704–6708.

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

[13] R. Clark, H. Silen, T. Kenter, and R. Leith, “Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs,” in Proc. 10th ISCA Speech Synthesis Workshop, pp. 99–104.

[14] Q. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, “Improving prosody modelling with cross-interutter bert embeddings for end-to-end speech synthesis,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6079–6083.

[15] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7922–7926.

[16] Y.-J. Zhang and Z.-H. Ling, “Extracting and predicting word-level style variations for speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1582–1593, 2021.

[17] Y. Ren, M. Lei, Z. Huang, S. Zhang, Q. Chen, Z. Yan, and Z. Zhao, “Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7577–7581.

[18] D. Tan and T. Lee, “Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement,” arXiv preprint arXiv:2110.03943, 2020.

[19] C. Du and K. Yu, “Mixture density network for phone-level prosody modelling in speech synthesis,” arXiv preprint arXiv:2102.00851, 2021.

[20] E. Selkirk, “On derived domains in sentence phonology,” Phonology, vol. 3, pp. 371–405, 1986.

[21] M. Liberman and A. Prince, “On stress and linguistic rhythm,” Linguistic inquiry, vol. 8, no. 2, pp. 249–336, 1977.

[22] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, “Towards multyscale style control for expressive speech synthesis,” arXiv preprint arXiv:2104.03521, 2021.

[23] C.-y. Tseng, S.-h. Pin, Y. Lee, H.-m. Wang, and Y.-c. Chen, “Fluent speech prosody: Framework and modeling,” Speech communication, vol. 46, no. 3-4, pp. 284–309, 2005.

[24] Y. Lei, S. Yang, and L. Xie, “Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 423–430.

[25] Y. Lei, S. Yang, X. Wang, and L. Xie, “Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in NIPS 2014 Workshop on Deep Learning, December 2014, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in Interspeech, vol. 2017, 2017, pp. 498–502.

[29] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in Neural Information Processing Systems, vol. 33, pp. 17022–17033, 2020.