ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446352

STYLESPEECH: SELF-SUPERVISED STYLE ENHANCING WITH VQ-VAE-BASED PRE-TRAINING FOR EXPRESSIVE AUDIOBOOK SPEECH SYNTHESIS

Xueyuan Chen[1,2,†], *Xi Wang*[3], *Shaofei Zhang*[3], *Lei He*[3], *Zhiyong Wu*[1,2,*], *Xixin Wu*[1,*], *Helen Meng*[1,2]

[1] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3] Microsoft, Beijing, China

{xychen,zywu,wuxx, hmmeng}@se.cuhk.edu.hk, {xwang, shazh, helei}@microsoft.com

ABSTRACT

The expressive quality of synthesized speech for audiobooks is limited by generalized model architecture and unbalanced style distribution in the training data. To address these issues, in this paper, we propose a self-supervised style enhancing method with VQ-VAE-based pre-training for expressive audiobook speech synthesis. Firstly, a text style encoder is pre-trained with a large amount of unlabeled text-only data. Secondly, a spectrogram style extractor based on VQ-VAE is pre-trained in a self-supervised manner, with plenty of audio data that covers complex style variations. Then a novel architecture with two encoder-decoder paths is specially designed to model the pronunciation and high-level style expressiveness respectively, with the guidance of the style extractor. Both objective and subjective evaluations demonstrate that our proposed method can effectively improve the naturalness and expressiveness of the synthesized speech in audiobook synthesis especially for the role and out-of-domain scenarios.[1]

Index Terms— expressive speech synthesis, self-supervised style enhancing, VQ-VAE, pre-training

1. INTRODUCTION

Recent text-to-speech (TTS) models, e.g., Tacotron 2 [1], TransformerTTS [2], FastSpeech 2 [3], have been developed with the capability to generate high-quality speech with a neutral speaking style. However, limited expressiveness persists as one of the major gaps between synthesized speech and real human speech, which draws growing attention to expressive speech synthesis studies [4, 5, 6, 7]. Synthesizing long-form expressive datasets, e.g., audiobooks, is still a challenging task, since wide-ranging voice characteristics tend to collapse into an averaged prosodic style.

There are a lot of works focusing on audiobook speech synthesis [8, 9, 10]. Recently, [11] proposes to use the neighbor sentences to improve the prosody generation. To make better use of contextual information, a hierarchical context encoder that considers adjacent sentences with a fixed-size sliding window is used to predict a global style representation directly from text [12]. Besides, [13] tries to consider as much information as possible (e.g., BERT embeddings, text embeddings and sentence ID) to improve style prediction. On top of these, a multi-scale hierarchical context encoder is proposed to predict both global-scale and local-scale style embeddings from context in a hierarchical structure [14]. All these existing works mainly focus on how to use the semantic information of contextual text to predict the expressiveness through an additional style encoder module. Too much information (phoneme, timbre, style, etc.) is simply mixed in the encoder part, leading to challenges for mel-spectrogram decoder. In addition, another serious problem for audiobook synthesis is the unbalanced style distribution in audiobook dataset. Most sentences are relatively plain narration voices, and only a small part is role voices with rich style variations, which brings a great challenge to modeling of style and expressiveness representation with limited audiobook training data, especially for role and out-of-domain scenarios.

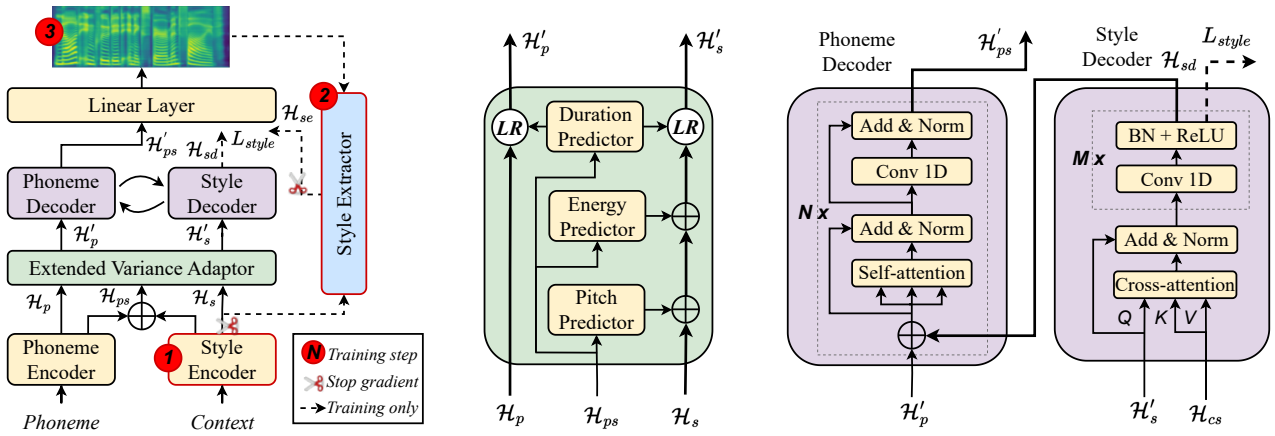
To solve the above-mentioned poor expressiveness problem in audiobook speech synthesis caused by generalized model architecture and unbalanced style distribution in the training data, this paper proposes a self-supervised style enhancing method with VQ-VAE-based pre-training for expressive audiobook synthesis. Firstly, a text style encoder is pre-trained with the help of a large amount of easily obtained unlabeled text-only data. Secondly, a spectrogram style extractor based on VQ-VAE is pre-trained using plenty of audio data that covers multiple expressive scenarios in other domains. On top of these, a special model architecture is designed with two encoder-decoder paths with the guidance of style extractor. To summarize, the main contributions of this paper are:

- We propose a VQ-VAE-based style extractor to model a better style representation latent space and relieve the unbalanced style distribution issues, which is pre-trained by plenty of easily obtained audio data that can cover complex style variations in a self-supervised manner.
- We design a novel TTS architecture with two encoder-decoder paths to model the pronunciation and high-level style expressiveness respectively, so as to enrich the expressive variation of synthesized speech in complex scenarios by strengthening both the encoder and decoder of TTS model.
- Both objective and subjective experimental results show that our proposed style enhancing approach achieves an effective improvement in terms of speech naturalness and expressiveness especially for the role and out-of-domain scenarios.

† Work conducted when the first author was intern at Microsoft.

* Corresponding authors. This research is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030), the CUHK Stanley Ho Big Data Decision Analytics Research Centre and the Centre for Perceptual and Interactive Intelligence.

[1] Audio samples: https://Chenxuey20.github.io/StyleSpeech



(a) Overall architecture of proposed model (b) Extended Variance Adaptor (c) Mel-spectrogram Decoder

Fig. 1. Proposed model structure, where (a) shows the overall architecture with two encoder-decoder paths, (b) shows details of Extended Variance Adaptor and (c) shows Mel-spectrogram Decoder containing Phoneme Decoder and Style Decoder with interaction.

2. RELATED WORK

Our work is related to Context-aware Augmented Deep Embedded Clustering (CADEC) [15] and Vector Quantized-Variational AutoEncoder (VQ-VAE) [16].

2.1. CADEC

CADEC is a two-stage style learning approach from abundant unlabeled plain text in a self-supervised manner. Firstly, it uses contrastive learning [17] to pre-train style embedding to distinguish similar and dissimilar utterances. To this end, a similar utterance is created by replacing an emotional word, determined by an emotion lexicon, with a similar one, while the other utterances in the randomly sampled minibatch are treated as dissimilar utterances. Secondly, the training samples in style embedding space are clustered by minimizing deep clustering loss [18], reconstruction loss and contrastive loss together. Compared with BERT [19], CADEC style embedding is more effective in learning styles other than content.

2.2. VQ-VAE

VQ-VAE is a powerful representation learning framework that can make effective use of the latent space. It combines VAE framework with discrete latent representations through a parameterisation of the posterior distribution of (discrete) latents given an observation. It can successfully model important features that usually span many dimensions in data space (e.g., objects span many pixels in images, phonemes in speech, the message in a text fragment, etc.) as opposed to focusing or spending capacity on noise and imperceptible details which are often local. Many extension models have been proposed, leading to high performance in various tasks, e.g., prosody learning [20] and speaker diarization [21].

3. METHODOLOGY

The overall architecture of our proposed model is illustrated in Fig. 1 (a). It mainly consists of two encoder-decoder paths with interaction. The first and primary one is the fine-grained phoneme path, while the second one is the coarse-grained style path.

3.1. Phoneme encoder-decoder path

The phoneme encoder-decoder path mainly focuses on the pronunciation based on FastSpeech 2 [3]. Both the phoneme encoder and phoneme decoder consist of several feed-forward Transformer (FFT) blocks, which are a stack of self-attention layer and 1D-convolution with residual connection and layer normalization. As shown in Fig. 1 (b), the phoneme hidden embedding $\mathcal{H}_p$ is repeated to frame-level phoneme representation $\mathcal{H}_p^{'}$ by length regulator (LR) in the extended variance adaptor. And it is worth noting that only $\mathcal{H}_p^{'}$ is further fed to the phoneme decoder in this path, not together with the pitch and energy, which is different from FastSpeech 2.

3.2. Style encoder-decoder path

The style encoder-decoder path focuses on the style modeling of synthesized speech. Specifically, a text style encoder and a spectrogram style extractor are designed and pre-trained to learn the style-related representations from contextual text and mel-spectrogram respectively, with a huge amount of unlabeled data. A style decoder is further adopted to make a better fusion of the explicit style features and implicit style representations in the decoding stage.

3.2.1. Style Encoder

We adopt the CADEC encoder [15] as our style encoder. It employs a pre-trained BERT [19] as backbone to extract semantic features, and an emotion lexicon [22] to extract emotion features. By contrastive learning with data augmentation and deep embedded clustering with an autoencoder structure, it can be trained with abundant unlabeled pain text and extract a more style-related representation from context. Finally, by accepting the contextual text C, CADEC encoder can output a global style representation:

$$\mathcal{H}_s = CADEC(C_0) \tag{1}$$

$$\mathcal{H}_{cs} = Concat[CADEC(C_i), i = -k, ..., k] \tag{2}$$

where $Concat[.]$ is the concatenation operation, $\mathcal{H}_s$ and $\mathcal{H}_{cs}$ are the hidden style embeddings for the contextual text C_0 of the current utterance and for the $2k+1$ neighbor utterances respectively.

3.2.2. Extended Variance Adaptor

Based on FastSpeech 2, the extended variance adaptor is designed to explicitly model the style-related features e.g, duration, pitch and energy. As shown in Fig. 1 (b), the phoneme encoder output $\mathcal{H}_p$ and style encoder output $\mathcal{H}_s$ are added together (denoted as $\mathcal{H}_{ps}$) to fed into the pitch predictor, energy predictor and duration predictor respectively to predict the phoneme-level explicit style features,

which are related to both the phoneme and style. Furthermore, the predicted pitch and energy together with the implicit style embedding $\mathcal{H}_s$ are repeated to become the frame-level style embedding $\mathcal{H}_s^{'}$ by the length regulator.

3.2.3. Style Extractor

As shown in Fig. 2, we adopt VQ-VAE [16] as our style extractor to extract a style-related latent representation from mel-spectrogram with a large amount of unlabeled audio data. Specifically, the encoder consists of two 2D-convolution layers with batch normalization (BN) and ReLU activation, followed by several ResBlock [23] layers, while the decoder adopts a symmetrical structure with the encoder. Besides, the one-hot speaker embedding conditions are also fed to decoder to remove the influence of timbres.

Only the low-frequency band of the mel-spectrogram $\mathcal{M}el_{20}$ (first 20 bins in each frame) is taken as input, as it is considered to contain almost complete style and much less content information compared with the full band. Besides, in order to further guide the model to extract style-related latent representations, we also use the frame-level pitch $\mathcal{H}_p$, frame-level energy $\mathcal{H}_e$ and text features $\mathcal{H}_s$ as additional inputs. Finally, a discrete style-related representation $\mathcal{H}_{se}$ can be extracted from the vector quantization layer output of well-pretrained style extractor, which can be described as follows:

$$\mathcal{H}_{se} = VQVAE(\mathcal{M}el_{20}, \mathcal{H}_p, \mathcal{H}_e, \mathcal{H}_s) \tag{3}$$

3.2.4. Style Decoder

The style decoder is designed to further integrate the explicit style features (pitch, energy) and implicit style embeddings in the decoding stage. As shown in Fig. 1 (c), in order to make the style transitions among contextual sentences more natural and smooth, a cross-attention module followed by residual connection is firstly adopted to consider the hierarchical context. Here, the frame-level style embedding $\mathcal{H}_s^{'}$ of current utterance is the query, while the hierarchical context style embedding $\mathcal{H}_{ce}$ from style encoder is the key and value. After that, several 1D-convolution layers with batch normalization and ReLU activation are further used to learn a style-related representation and finally output the style embedding $\mathcal{H}_{sd}$.

3.3. Interaction between phoneme and style paths

Existing expressive speech synthesis works mainly simply introduce style information into the TTS encoder part, leading to challenges to the mel-spectrogram decoder. As shown in Fig. 1 (c), we make the feature interaction between phoneme and style paths not only in the encoder part but also in the mel-spectrogram decoder part. Specifically, the output embedding $\mathcal{H}_{sd}$ of style decoder is fed into each FFT block of phoneme decoder as an additional style input in order to fully integrate the style and pronunciation information. After that, the well-mixed output embedding $\mathcal{H}_{ps}^{'}$ of mel-spectrogram decoder based on the two encoder-decoder paths is finally fed into the post linear layer to reconstruct the mel-spectrogram.

3.4. Training strategy and inference procedure

As shown in Fig. 1(a), our proposed model is trained in three stages.

i) In the first stage, the style encoder is pre-trained with a large amount of text data. Training details are similar to [15].

ii) In the second stage, the style extractor is trained with a large amount of audio data. Consistent with the original VQ-VAE, the total training loss consists of a reconstruction loss for reconstructing the mel-spectrogram, a vector quantisation loss for updating the dictionary and a commitment loss for making sure the encoder commits to an embedding.

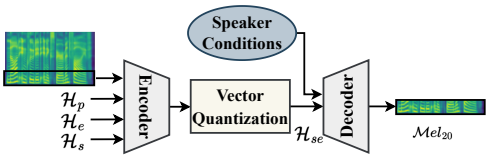


Fig. 2. Style Extractor based on VQ-VAE.

iii) In the third stage, the TTS model is trained with audiobook data. The model parameters of style encoder and style extractor are frozen without gradient update. An additional style loss $\mathcal{L}_{style}$ is adopted to the style encoder-decoder path to give a guidance from the pre-trained style extractor.

$$\mathcal{L}_{style} = MSE(\mathcal{H}_{sd}, \mathcal{H}_{se}) \tag{4}$$

where MSE is the mean square error (MSE) loss, $\mathcal{H}_{sd}$ and $\mathcal{H}_{se}$ are the outputs of style decoder and style extractor respectively. The total loss is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{tts} + \alpha\mathcal{L}_{style} \tag{5}$$

where $\mathcal{L}_{tts}$ is the TTS loss consistent with FastSpeech 2.

During inference, the style extractor is abandoned (as shown by the dotted line in Fig. 1 (a)). By accepting phoneme and context input, the model can synthesize speech with more expressive styles.

4. EXPERIMENTS

4.1. Datasets and system settings

We use 3 types of internal Mandarin datasets to train style encoder, style extractor and TTS model respectively. The style encoder is trained with a large plain text dataset, containing 7.5M audiobook sentences. The style extractor is trained with a large multi-speaker audio corpus, which contains around 400 hours of audios with corresponding text and covers a wealth of application scenarios and style variations. We use an audiobook corpus to train the TTS model. It has around 30-hour speech data with context and is cut into 30,000 audio clips, of which 1000 clips are used for validation and 500 clips for test, and the rest for training. Besides, another small audiobook dataset covering several different categories is further used to evaluate the out-of-domain performance. Details are shown in Table 1.

Table 1. Datasets of different training stages.

Stage	Style Encoder	Style Extractor	TTS model
Type	Text	Audio	Audiobook
Size	7.5M	400 hours	30 hours

For feature extraction, we transform the raw waveforms into 80-dim mel-spectrograms with sampling rate 16kHz, frame size 1200 and hop size 240. The context of current sentence is made up of its two past sentences, two future ones and itself. The codebook size of style extractor is set to 512 and the style loss coefficient α is 1. All the trainings are conducted with a batch size of 16 on a NVIDIA V100 GPU. The Adam optimizer is adopted with $\beta_1 = 0.9, \beta_2 = 0.98$. In addition, a well-trained HiFi-GAN [24] is used as the vocoder to generate waveform. Two FastSpeech 2 based methods are implemented for comparison as follows:

- **FastSpeech 2**: Original FastSpeech 2 [3] is implemented as the first baseline method.
- **FS2-CADEC**: Inspired by [15], we set an end-to-end TTS model by combining CADEC encoder with FastSpeech 2.

Table 2. Subjective and objective evaluation results for different models.

Model	MOS (out-of-domain)	Style MOS (out-of-domain)	Paragraph CMOS (out-of-domain)	F0 RMSE	Energy RMSE	Duration MSE	MCD
Ground Truth	4.22 ± 0.14	4.17 ± 0.11	-	-	-	-	-
FastSpeech 2	4.00 ± 0.10	4.09 ± 0.13	-0.161	58.930	10.848	0.0636	5.969
FS2-CADEC	3.97 ± 0.10	4.08 ± 0.10	-0.022	58.865	10.788	0.0629	5.951
Proposed	$\mathbf{4.08 \pm 0.09}$	$\mathbf{4.17 \pm 0.08}$	**0**	**57.271**	**10.697**	**0.0617**	**5.937**

4.2. Subjective comparison for different systems

Mean Opinion Score (MOS) is first conducted in terms of the comprehensive performance of synthesized speech including sound quality, naturalness, expressiveness, etc, to ensure that all baseline systems are well reproduced. Furthermore, style MOS is used to only focus on the style expressiveness of synthesized speech, and paragraph Comparative MOS (CMOS) is used to evaluate the style transition among sentences within a paragraph. All the tests are conducted on Microsoft UHRS crowdsourcing platform. As our ultimate goal is to synthesize any other given audiobook, we mainly focus on the out-of-domain role performance. 50 single sentences and 20 short paragraphs are randomly selected in an out-of-domain set. Each audio is judged by at least 10 native speakers.

As shown in Table 2, our proposed approach achieves the best MOS of 4.08 and best Style MOS of 4.17 compared with the baseline methods. Specially, our proposed approach achieves comparable results to the ground truth recording on Style MOS. In the paragraph-level comparison, our proposed approach also achieves the best CMOS performance. These results demonstrate the effectiveness of our proposed methods especially on the role style expressiveness in out-of-domain scenarios.

4.3. Objective comparison for different systems

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.

As shown in Table 2, our proposed model achieves 57.271 for F0 RMSE, 10.697 for Energy RMSE, 0.0617 for duration MSE and 5.937 for MCD, which outperforms all the baselines on all metrics. These results indicate that our proposed model can predict more accurate explicit style features, e.g., duration, pitch and energy, and reconstruct more preserved mel-spectrograms, than baselines.

4.4. Analysis for the pre-training strategy

To further verify whether the pre-training strategy with plenty of audio data is helpful for the style representation latent space modeling of the unbalanced audiobook data, we also train a style extractor with only audiobook dataset for comparison. We extract a few role style embeddings with different style categories in the audiobook dataset by the above-mentioned two well-trained style extractors and make a t-SNE visualization respectively.

Fig. 3 (a) shows the extracted style embeddings when only audiobook dataset participates in training, and Fig. 3 (b) shows the extracted style embeddings when we use the large dataset to train the style extractor. Obviously, compared to Fig. 3 (a), there is a better cohesion and distribution differences among different styles in Fig. 3 (b). Note that the style in audiobook dataset is too complex to be divided into several categories, and there may be several slightly different distribution forms even within the same style category by manual annotation. The results show that it's difficult to model the style representation latent space with only unbalanced and limited audiobook data, and our proposed pre-training strategy with a large amount of audio data that covers multiple expressive scenarios in other domains is necessary and beneficial for the style latent space modeling of audiobook speech synthesis.

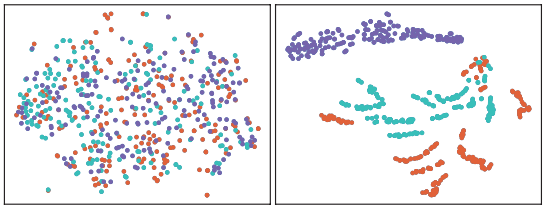


Fig. 3. Visualization of style embedding space trained with different datasets. Each color indicates a ground truth style category. (a) represents style embeddings while training with only audiobook dataset. (b) represents style embeddings while training with large dataset.

4.5. Ablation study for the model architecture

To further investigate the influence of several main modules in our proposed model, we have tried three other settings: i) **Proposed w/o Style Encoder**: The style encoder is removed, only the output $\mathcal{H}_p$ of phoneme encoder is fed to the extended variance adaptor for both the two paths. ii) **Proposed w/o Style Decoder**: The style decoder is removed, both the outputs $\mathcal{H}_p^{'}$ and $\mathcal{H}_s^{'}$ of the extended variance adaptor are fed to phoneme decoder together. iii) **Proposed w/o Style Extractor**: The style extractor is removed, which means the style loss $\mathcal{L}_{style}$ is removed during the TTS training stage.

CMOS is employed to compare the synthesized speech in terms of naturalness and expressiveness. As shown in Table 3, the performance of the three settings is degraded to various degrees respectively compared with the proposed method. This indicates that all these components have substantial impact on our proposed model. Furthermore, the results also indicate that both our proposed style pre-training strategy and the novel TTS architecture with two encoder-decoder paths can alleviate the role and out-of-domain expressiveness deterioration problem caused by unbalanced style distribution and insufficient model generalizability.

Table 3. CMOS comparison for ablation study.

Model	CMOS
Proposed	**0**
w/o Style Encoder	-0.142
w/o Style Decoder	-0.133
w/o Style Extractor	-0.150

5. CONCLUSION

This work addresses the problem of poor expressiveness in audiobook speech synthesis due to generalized model architecture and unbalanced style distribution in the training data. We propose a pre-trained VQ-VAE-based style extractor and a novel TTS architecture with two encoder-decoder paths. Both objective and subjective experiments demonstrate the effective performance of our proposed method in terms of naturalness and expressiveness of the synthesized speech, especially for the role and out-of-domain scenarios.

6. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP 2018*. IEEE, 2018, pp. 4779–4783.

[2] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 6706–6713.

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

[4] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif 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.

[5] Xueyuan Chen, Changhe Song, Yixuan Zhou, Zhiyong Wu, Changbin Chen, Zhongqin Wu, and Helen Meng, "A character-level span-based model for mandarin prosodic structure prediction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7602–7606.

[6] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif 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.

[7] Xueyuan Chen, Qiaochu Huang, Xixin Wu, Zhiyong Wu, and Helen Meng, "Hilvoice: Human-in-the-loop style selection for elder-facing speech synthesis," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 86–90.

[8] Marcela Charfuelan and Ingmar Steiner, "Expressive speech synthesis in mary tts using audiobook data and emotionml.," in *INTERSPEECH*, 2013, pp. 1564–1568.

[9] Daisy Stanton, Yuxuan Wang, and RJ Skerry-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.

[10] Aghilas Sini, Damien Lolive, Gaëlle Vidal, Marie Tahon, and Élisabeth Delais-Roussarie, "Synpaflex-corpus: An expressive french audiobooks corpus dedicated to expressive speech synthesis," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[11] Guanghui Xu, Wei Song, Zhengchen Zhang, Chao Zhang, Xiaodong He, and Bowen Zhou, "Improving prosody modelling with cross-utterance 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.

[12] Shun Lei, Yixuan Zhou, Liyang Chen, Zhiyong Wu, Shiyin Kang, and Helen Meng, "Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis," in *ICASSP 2022*. IEEE, 2022, pp. 7922–7926.

[13] Ning-Qian Wu, Zhao-Ci Liu, and Zhen-Hua Ling, "Discourse-level prosody modeling with a variational autoencoder for non-autoregressive expressive speech synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7592–7596.

[14] Xueyuan Chen, Shun Lei, Zhiyong Wu, Dong Xu, Weifeng Zhao, and Helen Meng, "Unsupervised multi-scale expressive speaking style modeling with hierarchical context information for audiobook speech synthesis," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 7193–7202.

[15] Yihan Wu, Xi Wang, Shaofei Zhang, Lei He, Ruihua Song, and Jian-Yun Nie, "Self-supervised context-aware style representation for expressive speech synthesis," *arXiv preprint arXiv:2206.12559*, 2022.

[16] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[18] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen Mckeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang, "Supporting clustering with contrastive learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5419–5430.

[19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[20] Yi Zhao, Haoyu Li, Cheng-I Lai, Jennifer Williams, Erica Cooper, and Junichi Yamagishi, "Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction," in *Interspeech 2020*, 2020.

[21] Jennifer Williams, Yi Zhao, Erica Cooper, and Junichi Yamagishi, "Learning disentangled phone and speaker representations in a semi-supervised vq-vae paradigm," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7053–7057.

[22] Sven Buechel, Susanna Rücker, and Udo Hahn, "Learning and evaluating emotion lexicons for 91 languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1202–1217.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] Jungil Kong, Jaehyeon Kim, and Jaekyoung 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.