DISCOURSE-LEVEL PROSODY MODELING WITH A VARIATIONAL AUTOENCODER FOR NON-AUTOREGRESSIVE EXPRESSIVE SPEECH SYNTHESIS

Ning-Qian Wu, Zhao-Ci Liu, Zhen-Hua Ling

National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, P.R.China
{wunq,zcliu8}@mail.ustc.edu.cn, zhling@ustc.edu.cn

ABSTRACT

To address the issue of one-to-many mapping from phoneme sequences to acoustic features in expressive speech synthesis, this paper proposes a method of discourse-level prosody modeling with a variational autoencoder (VAE) based on the non-autoregressive architecture of FastSpeech. In this method, phone-level prosody codes are extracted from prosody features by combining VAE with FastSpeech, and are predicted using discourse-level text features together with BERT embeddings. The continuous wavelet transform (CWT) in FastSpeech2 for F0 representation is not necessary anymore. Experimental results on a Chinese audiobook dataset show that our proposed method can effectively take advantage of discourse-level linguistic information and has outperformed FastSpeech2 on the naturalness and expressiveness of synthetic speech.

Index Terms— speech synthesis, prosody modeling, FastSpeech, discourse-level modeling, variational autoencoder

1. INTRODUCTION

Statistical parameter speech synthesis (SPSS) typically contains two components, the acoustic model which converts text features into frame-level acoustic features, and the vocoder which converts acoustic features into speech waveforms. With the development of deep learning techniques, SPSS has made rapid progress in recent years, and neural networks have been widely applied to both the acoustic model and the vocoder.

Among the neural network based acoustic models, autoregressive models such as Tacotron2 [1] and Transformer-TTS [2] can generate high-quality speech. However, these models suffer from the unsatisfactory robustness of the attention mechanism, especially when the training data is highly expressive. Besides, these autoregressive models have to predict acoustic features frame by frame, resulting in a low inference efficiency. Non-autoregressive models, such as FastSpeech [3, 4] and Parallel Tacotron [5], have been proposed to address these issues. These models adopt a duration prediction module to increases the robustness of acoustic feature prediction and a non-autoregressive decoder to improve the inference efficiency. Recently, acoustic models based on Normalizing Flows [6] and Diffusion Probabilistic Models [7] have also pushed the naturalness of speech synthesis to a new level.

Prosody, including the intonation, accent, speed, rhythm of speech, provides important information beyond the phonetic segments of a sentence. Prosody modeling is essential for expressive speech synthesis, whose application scenarios include audiobook production, dubbing, etc. One challenge in expressive speech synthesis is the issue of one-to-many mapping from phoneme sequences to acoustic features, especially the prosody variations in expressive speech with multiple styles. One approach to address this issue is resorting to the latent representations learned from acoustic features. Variational autoencoder (VAE) [8, 9], as a deep neural generative model with a latent distribution, has been applied to expressive speech synthesis not only for style transfer [10], but also for prediction or sampling [11, 12]. Another approach to address the one-to-many issue is utilizing the textual information in a context range wider than the current sentence, e.g., at the paragraph or discourse level. In previous studies, the prosodic representation and text information of the previous sentence were used when synthesizing the current sentence to improve the coherence between sentences [13, 14]. The text embeddings of neighbouring sentences were also employed to improve the prosody generation [15, 16]. However, these two approaches were usually studied separately, and all existing studies mentioned above were based on an autoregressive architecture, e.g., Tacotron2 [1].

Therefore, this paper proposes a method of discourse-level prosody modeling with a VAE for non-autoregressive expressive speech synthesis. In this method, a VAE is combined with FastSpeech [3] to extract phone-level latent prosody representations, i.e., prosody codes, from the fundamental frequency (F0), energy and duration of the speech. Then, a Transformer-based model is constructed to predict prosody codes, taking discourse-level linguistic features and BERT [17] embeddings as input. Our experiments on a Chinese audiobook dataset demonstrate that our proposed method achieved better naturalness and expressiveness of synthetic speech than FastSpeech2 [4]. To the best of our knowledge, our proposed method is the first attempt at performing discourse-level modeling with a non-autoregressive architecture.

This paper is organized as follows. In Section 2, we introduce related works briefly. In Section 3, the details of the proposed methods are described. Section 4 reports the results of our experiments. Conclusions are given in Section 5.

2. RELATED WORK

2.1. FastSpeech1

FastSpeech1 [3] is the most typical non-autoregressive speech synthesis model. Its text encoder converts the input text features into phoneme-level hidden representations. The hidden representations are expanded to frame-level ones by the length regulator according to phoneme durations and are then fed into the decoder for predicting mel-spectrograms. The phoneme durations are extracted from audio at the training stage and are predicted by a duration predictor at the synthesis stage. The duration predictor takes the hidden representations as input, and is trained jointly with the encoder and the decoder. The encoder and the decoder consist of positional encoding

and multiple feed-forward Transformer (FFT) blocks based on 1D convolution and self-attention.

The structure of FastSpeech1 is straightforward. The parallel model structure and the length regulator guarantee a fast, stable, and controllable generation process. However, due to the one-to-many mapping from phonemes to mel-spectrograms, training FastSpeech1 directly with the ground-truth target usually results in over-smoothed output, which is particularly serious when using expressive training data. In order to alleviate this problem, FastSpeech1 introduces an autoregressive teacher model for knowledge distillation, which leads to a complex training pipeline and still can't achieve satisfactory expressiveness of synthetic speech.

2.2. FastSpeech2

FastSpeech2 [4] provides another approach to solve the one-to-many problem. The additional variation information is introduced as conditional inputs. At the training stage, frame-level pitch and energy extracted from audio data are discretized and then converted into learnable embeddings. The embedding vectors are added to the input of the decoder as the condition. Given the ground-truth prosody information, the decoder can model the mel-spectrograms more accurately than FastSpeech1.

However, FastSpeech2 needs extra modules to predict pitch and energy at the synthesis stage, which still faces the one-to-many problem of mapping phonemes to prosody information. To improve the expressiveness of synthetic speech, pitch values are converted into pitch spectrograms by continuous wavelet transform (CWT). The pitch predictor is trained to fit the ten-dimensional pitch spectrograms and to predict the sentence-level pitch mean and variance. One issue with CWT is that an interpolation procedure at unvoiced segments is always necessary to produce continuous pitch contours for CWT calculation, which distracts the model from modeling the observed pitch values of voiced segments. What's more, FastSpeech2 models pitch spectrograms at the frame level, which may not be appropriate for capturing long-term prosody characteristics at the discourse level.

3. PROPOSED METHOD

In order to ease the one-to-many problem in non-autoregressive speech synthesis, a method of discourse-level prosody modeling with variational autoencoder (VAE) [8] is proposed in this paper. It contains a prosody code extractor that condenses the pitch, energy, and duration information into phone-level prosody codes under a VAE framework, and a prosody code predictor which models the prosody codes at the discourse level.

3.1. VAE-based prosody code extractor

Fig.1 shows the structure and training losses of the prosody code extractor. The backbone of the model structure is FastSpeech1 [3], and the VAE framework is employed to extract phone-level hidden prosody representations, i.e., prosody codes.

The reference encoder predicts the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ of the posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{c}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I}\right)$ from the observed prosody condition $\boldsymbol{c}$. Unlike most VAE-based speech synthesis methods [9–12], we take the variation information (i.e., pitch, energy, and upsampled frame-level duration) instead of mel-spectrograms as the input of the reference encoder, inspired by FastSpeech2 [4]. So that the hidden representation $\boldsymbol{z}$ can focus on prosodic characteristics rather than other spectrum details, such as background noise. The reference encoder consists of six 1-D convolutional layers, two Bi-LSTM layers, and two linear projection layers. The output of the first Bi-LSTM layer is averaged to the phoneme level to facilitate the discourse-level prosody code prediction in Section 3.2. Then, the hidden representation sampled from $q_\phi(\boldsymbol{z}|\boldsymbol{c})$ by the reparameterization trick is concatenated with the output of the text encoder for decoding. The structures of the text encoder, duration predictor, length regulator and mel-spectrogram decoder follows the ones in FastSpeech1 [3].

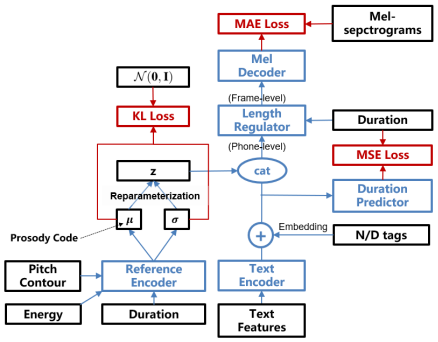

Fig. 1. The structure and training losses of the VAE-based prosody code extractor, where "cat" means concatenation.

The prosody code extractor is trained by optimizing the evidence lower bound (ELBO) loss

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\lambda D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{c}) \| p(\boldsymbol{z})\right) + \mathbb{E}_{q_\phi(\boldsymbol{z}|\mathbf{c})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{t})\right], \tag{1}$$

where $\boldsymbol{x}$ includes the target mel-spectrograms and phone duration, $\boldsymbol{t}$ denotes the input text features, $\boldsymbol{\phi}$ is the parameter set of the reference encoder, $\boldsymbol{\theta}$ is the parameter set of other modules in Fig. 1, D_{KL} calculates the Kullback-Leibler (KL) divergence between two distributions, and $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the prior distribution of $\boldsymbol{z}$. The second term on the right side of Eq. (1) corresponds to the mean absolute error (MAE) loss of mel-spectrograms and the mean square error (MSE) loss of phone durations in Fig. 1. To solve the KL vanishing problem, we combine KL annealing [18] and free bits [19] to ensure stable training. The weight λ of the KL term is close to zero at the beginning of training and then increases to 0.1. The optimization of the KL loss is cut when its value is smaller than 5. The dimension of $\boldsymbol{z}$ is set to 4 according to the results of our preliminary experiments.

A Mandarin Chinese audiobook dataset from the Internet was adopted in our experiments. The anchor used different styles when reading narrations and dialogues. To simplify the task, this paper focuses on synthesizing narration-style texts. Thus, the training sentences were labelled with dialogue/narration (D/N) tags. As shown in Fig. 1, the tags are converted to trainable embeddings and added to the output of the text encoder. Only the narration-style texts were used in our test set for evaluation.

Finally, the four-dimensional mean vector $\boldsymbol{\mu}$ predicted from the reference encoder for each phone in the training set is defined as the prosody code of this phone.

3.2. Discourse-level prosody code predictor

The structure of the discourse-level prosody code predictor is drawn in Fig. 2. It is composed of an encoder and a decoder, which both adopt the same structure as the FastSpeech's encoder. Similar to the

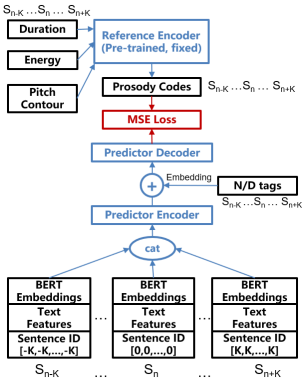

Fig. 2. The structure and training losses of the discourse-level prosody code predictor.

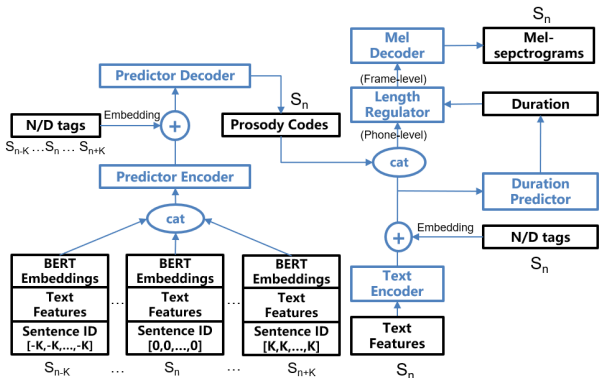

Fig. 3. The generation process of our proposed method.

prosody code extractor, D/N tags are also integrated into the output of the predictor encoder.

One key to solving the one-to-many problem in prosody code prediction is to enrich the input linguistic representations. Therefore, we introduce BERT [17] into the prosody code predictor. BERT embeddings are effective linguistic representations for natural language processing and have been applied to the task of speech synthesis recently. The BERT model is a Transformer-based language model estimated from a large amount of unlabelled text by self-supervised learning. In our model, the open-source Chinese BERT pre-training model [20] is employed to extract word-level representations from the input text. The extracted BERT embeddings are then upsampled to phone-level ones and concatenated with the text features.

Another intuitive idea is to use the context information outside the current sentence. As shown in Fig. 2, the input features of the predictor model are extended by simply concatenating the features extracted from multiple neighbouring sentences. Since the extracted prosody codes are phone-level ones instead of frame-level ones, this discourse-level modeling can be achieved without worrying about the conflict between the lengths of concatenated feature sequences and memory limitations. In order to make better use of the cross-sentence information, we add one-dimensional sentence ID to the input feature to indicate which sentence the current phone belongs to. Each training instance contains $2K+1$ continuous sentences in an audiobook chapter, and their sentence IDs are $[-K, ..., K]$. As shown in Fig. 2, the prosody code predictor is trained to predict the prosody codes of all the $2K+1$ sentences simultaneous with an MSE Loss.

At the generation stage, the prosody code predictor is combined with other modules in the prosody code extractor besides the reference encoder to synthesize the mel-spectrograms of the middle sentence, as Fig. 3 shows.

4. EXPERIMENTS

4.1. Experimental setup

In order to verify the performance of our proposed method on synthesizing expressive speech, we collected a Chinese audiobook with about one thousand chapters from the Internet as training data. It contained 200 hours of recordings from a single male speaker with a highly expressive style. The waveforms had 16kHz sampling rate and 16bits resolution. The recording data was segmented into sentences and manually transcribed into text. Then the dataset was divided into a training set, a validation set, and a test set. The training set contained 149,615 sentences from 1,118 chapters. The validation set and the test set both contained about 1,300 sentences from 10 chapters. In the following experiment, the test set was only used to calculate objective indicators. The other test set was selected from another well-known Chinese online novel with a total of 150 sentences in 4 chapters, which was specifically used to conduct subjective experiments.

The front-end analyzed the text to produce 898-dimensional text feature vectors for each phone. The dimension of BERT embeddings was 1024. 80-dimensional mel-spectrograms were extracted with a frame shift of 12.5ms and frame length of 50ms. The pitch contour was extracted by PyWORLD [21] tool with linear interpolation. L2-norm of the amplitude of each short-time Fourier transform (STFT) frame was defined as the energy. Phoneme durations were extracted by the Montreal forced alignment (MFA) [22] tool. A HiFi-GAN[1] [23] vocoder was trained on the collected 200h training sets for waveform reconstruction. The details of the model are basically the same as the experimental configuration v1 in [23]. The upsampling rates in the vocoder were modified to fit the 16kHz sampling rate.

Since our data had a rich variety of styles, Tacotron 2 [1] performed unstably on our data. FastSpeech2 [4] , which adopted non-autoregressive architecture, was chosen as our baseline model. The encoder and decoder in the FastSpeech2 baseline and our proposed model had the same structure of 4 layers FFT blocks whose hidden dimension was 256. For a fair comparison, the pitch predictor and energy predictor in the FastSpeech2 baseline also had 4-layer FFT blocks instead of two-layer convolution for better prosody prediction. The N/D tags were introduced to FastSpeech2 baseline as an additional condition. In order to study the effectiveness of BERT embeddings and discourse-level modeling, four models were built for comparison as follows.

- **FS2**: the FastSpeech2 baseline model;
- **K0**: the proposed method without discourse-level modeling (i.e., $K=0$);
- **K10**: the proposed method considering 21 neighbouring sentences in discourse-level modeling (i.e., $K=10$);
- **K10 w/o BERT**:the proposed method considering 21 neighbouring sentences in discourse-level modeling but without using BERT embeddings.

[1]We used the open source implemention of HiFi-GAN vocoder at https://github.com/jik876/hifi-gan.

Table 1. The log F0 distribution distances between the ground-truth speech and the speech generated by different models.

Method	Wasserstein Distance	Energy Distance
FS2	0.0445	0.0588
K0	0.0226	0.0306
K10	0.0240	0.0312
K10 w/o BERT	0.0308	0.0394

Table 2. The log F0 distribution distances between the ground-truth speech and the speech generated by K10 with different contextual information.

Method	Wasserstein Distance	Energy Distance
A	0.0240	0.0312
B	0.0256	0.0340
C	0.0313	0.0444

4.2. Objective evaluation

Fundamental frequency (F0) is the most important prosody feature. The distances between the log F0 distribution of the ground-truth speech and that of the generated speech on the test set were used as objective evaluation metrics. The results of Wasserstein distance [24] and energy distance [25] are shown in Table 1. It can be observed that the log F0 distributions of the speech generated by all models were very close to that of the ground-truth speech. But we still can see that our proposed models K0 and K10 achieved lower distances than the FastSpeech2 baseline model. Comparing K10 with K10 w/o BERT, we can see the effectiveness of introducing BERT embeddings into the prosody code predictor.

To analyze the effectiveness of introducing surrounding sentences, we conducted experiments on our test set using model K10 with different contextual information as follows:

- **A**: using matched surrounding utterances when synthesizing the current sentence;
- **B**: replacing surrounding utterances with the current sentence (i.e. repeating the current sentence 21 times as input);
- **C**: replacing surrounding utterances with mismatched utterances sampled from another novel.

We calculated the distance measures of the generated speech by the 3 methods, the results are shown in Table 2. It can be observed that the distances of speech synthesized by the model increased when mismatched contexts were used. The results demonstrate that under our proposed framework, contextual information can effectively influence the prosody of synthetic speech.

4.3. Subjective evaluation

We conducted one group of mean opinion score (MOS) test and several groups of preference tests for subjective evaluation.[2] Thirteen native Chinese speakers participated in each test. In the MOS test, participants were asked to rate the naturalness of speech samples synthesized from different systems on a scale from 1 to 5. In the preference tests, the participants must listen to two samples and indicate which utterance achieved better expressiveness or there was no preference. For intuitive comparison, the samples generated by all the systems used the same phone durations predicted by the FastSpeech2 baseline.

[2]Demos of generated speech by different models can be found at http://home.ustc.edu.cn/%7Ewunq/DLFS/demo.html

Table 3. The MOS with 95% confidence intervals of the FastSpeech2 baseline (FS2) and the proposed methods.

FS2	K0	K10
3.58±0.17	4.15±0.16	4.25±0.13

Table 4. Average preference scores on speech expressiveness among FastSpeech2 baseline (FS2) and the proposed methods, where N/P means "no preference" and p denotes the p-value of the one-way ANOVA between two methods.

FS2	K10	K0	K10 w/o BERT	N/P	p
14.39%	69.12%	–	–	16.49%	<0.0001
17.89%	–	64.92%	–	17.19%	<0.0001
21.05%	–	–	54.04%	24.91%	<0.0001
–	50.53%	33.68%	–	15.79%	<0.0001
–	54.74%	–	20.00%	25.26%	<0.0001

The MOS test compared the FS2, K0 and K10 methods, and the results are shown in Table 3. It is evident that our proposed model K10 outperformed the FS2 baseline significantly. It was reported that some words synthesized by FS2 suffered from F0 discontinuity, which degraded the naturalness MOS. K10 also performed slightly better than K0, which demonstrates the effectiveness of employing discourse-level linguistic information for prosody code prediction.

Table 4 shows the results of preference tests on speech expressiveness among baseline and proposed methods. Our proposed method with discourse-level modeling (K10) was significantly better than the baseline model on the expressiveness of synthetic speech. It is worth noting that even without discourse-level modeling (K0) or without BERT embeddings (K10 w/o BERT), the performance of the proposed model degraded significantly but still exceeded the baseline. These results clearly show the advantage of introducing VAE-based phone-level prosody codes into the FastSpeech model and utilizing richer contextual information for improving the expressiveness of synthetic speech.

5. CONCLUSION

In this paper, we have presented a non-autoregressive acoustic model based on phone-level prosody codes for expressive speech synthesis. The phone-level prosody codes are extracted from prosody features by combining VAE with FastSpeech. The prosody code predictor takes advantage of discourse-level text features and BERT embeddings effectively to ease the one-to-many mapping problem between text futures and prosodic variations. Our experiments on a Chinese novel dataset indicate that our proposed method can achieve better naturalness and expressiveness than FastSpeech2. In our future work, we will try to address the mismatch between training with true prosody features, and testing with predicted prosody features. We will also consider introducing richer text information to achieve better prosody prediction performance and improving the BERT model to derive embeddings that are more suitable for prosody modeling.

6. ACKNOWLEDGEMENTS

This work was partially funded by the National Key R&D Program of China under Grant 2019YFF0303001 and the National Nature Science Foundation of China under Grant 61871358.

7. REFERENCES

[1] Jonathan Shen, Ruoming Pang, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 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, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[4] 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.

[5] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5709–5713.

[6] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[7] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8599–8608, PMLR.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, pp. 1, 2014.

[9] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al., "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2018.

[10] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua 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.

[11] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. 2018, pp. 3067–3071, ISCA.

[12] Sri Karlapati, Ammar Abbas, Zack Hodari, Alexis Moinet, Arnaud Joly, Penny Karanasou, and Thomas Drugman, "Prosodic representation learning and contextual sampling for neural text-to-speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6573–6577.

[13] Pilar Oplustil-Gallegos and Simon King, "Using previous acoustic context to improve text-to-speech synthesis," *arXiv preprint arXiv:2012.03763*, 2020.

[14] Shubhi Tyagi, Marco Nicolis, Jonas Rohnke, Thomas Drugman, and Jaime Lorenzo-Trueba, "Dynamic Prosody Generation for Speech Synthesis Using Linguistics-Driven Acoustic Embedding Selection," in *Proc. Interspeech 2020*, 2020, pp. 4407–4411.

[15] 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.

[16] Slava Shechtman and Avrech Ben-David, "Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 66–71.

[17] 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.

[18] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio, "Generating sentences from a continuous space," in *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*. Association for Computational Linguistics (ACL), 2016, pp. 10–21.

[19] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling, "Improved variational inference with inverse autoregressive flow," *Advances in neural information processing systems*, vol. 29, pp. 4743–4751, 2016.

[20] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, Online, Nov. 2020, pp. 657–668, Association for Computational Linguistics.

[21] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[22] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*. Aug. 2017, pp. 498–502, ISCA.

[23] 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, 2020.

[24] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, pp. 47, 2017.

[25] Maria L Rizzo and Gábor J Székely, "Energy distance," *wiley interdisciplinary reviews: Computational statistics*, vol. 8, no. 1, pp. 27–38, 2016.