PERIOD VITS: VARIATIONAL INFERENCE WITH EXPLICIT PITCH MODELING FOR END-TO-END EMOTIONAL SPEECH SYNTHESIS

Yuma Shirahata\textsuperscript{1}, Ryuichi Yamamoto\textsuperscript{1}, Eunwoo Song\textsuperscript{2}, Ryo Terashima\textsuperscript{1}, Jae-Min Kim\textsuperscript{2}, Kentaro Tachibana\textsuperscript{1}

\textsuperscript{1}LINE Corp., Tokyo, Japan, \textsuperscript{2}NAVER Corp., Seongnam, Korea

ABSTRACT

Several fully end-to-end text-to-speech (TTS) models have been proposed that have shown better performance compared to cascade models (i.e., training acoustic and vocoder models separately). However, they often generate unstable pitch contour with audible artifacts when the dataset contains emotional attributes, i.e., large diversity of pronunciation and prosody. To address this problem, we propose Period VITS, a novel end-to-end TTS model that incorporates an explicit periodicity generator. In the proposed method, we introduce a frame pitch predictor that predicts prosodic features, such as pitch and voicing flags, from the input text. From these features, the proposed periodicity generator produces a sample-level sinusoidal source that enables the waveform decoder to accurately reproduce the pitch. Finally, the entire model is jointly optimized in an end-to-end manner with variational inference and adversarial objectives. As a result, the decoder becomes capable of generating more stable, expressive, and natural output waveforms. The experimental results showed that the proposed model significantly outperforms baseline models in terms of naturalness, with improved pitch stability in the generated samples.

Index Terms—Text-to-speech, end-to-end model, pitch modeling, variational inference, emotional speech

1. INTRODUCTION

Text-to-speech (TTS) has recently had a significant impact due to the rapid advancement of deep neural network-based approaches [1]. In most previous studies, TTS models were built as a cascade architecture of two separate models—an acoustic model that generates pre-defined acoustic features (e.g., mel-spectrogram) from text [2, 3] and a vocoder that synthesizes waveform from the acoustic feature [4, 5, 6]. Although these cascade models were able to generate speech reasonably well, they typically suffered from an error deriving from the use of pre-defined features and separated optimization for the two independent models. Sequential training or fine-tuning can mitigate the quality degradation [7], but the training procedure is complicated.

To address this problem, several works have investigated the use of fully end-to-end architecture that jointly optimizes the acoustic and vocoding models\footnote{We refer to text-to-wave model as end-to-end TTS throughout this paper. Note that the term is not used for cascade type models.} [8, 9, 10, 11]. One of the most successful works is VITS [10], which adopts a variational autoencoder (VAE) [12] with the augmented prior distribution by normalizing flows [13]. The VAE is used to acquire the trainable latent acoustic features from waveforms, whereas the normalizing flows are used to make the hidden text representation as powerful as the latent features.

However, we found that although VITS generates natural-sounding speech when trained with a reading style dataset, its performance is limited when applied to more challenging tasks, such as emotional speech synthesis, where the dataset has significant diversity in terms of pronunciation and prosody. Specifically, the model generates less intelligible voices with unstable pitch contour. Although the intelligibility problem could be addressed by expanding the phoneme-level parameters of prior distribution to frame-level parameters [14], it is still a challenge to generate accurate pitch information due to the architectural limitation of the non-autoregressive vocoders [15].

To tackle this, we propose Period VITS, a novel TTS system that explicitly provides sample-level and pitch-dependent periodicity when generating the target waveform. In particular, the proposed model consists of two main modules termed the prior encoder and the waveform decoder (hereinafter simply called “decoder”). On the prior encoder side, we employ a frame prior network with a frame pitch predictor that can simultaneously generate the parameters of the prior distribution and prosodic features in every frame. Note that the parameters are used to learn expressive prior distribution with normalizing flows and the prosodic features such as the pitch and the voicing flags are used to produce the sample-level sinusoidal source signal. On the decoder side, this periodic source is fed to every up-sampled representation in the HiFi-GAN-based vocoder [6], to guarantee pitch stability in the target waveform. Note that the training process optimizes the entire model in the end-to-end scheme from the variational inference point of view.

Several works have addressed similar problems by focusing on the periodicity of speech signals in the vocoder context [16, 17, 18]. The proposed model differs in that the conventional methods have used pre-defined acoustic features and only optimized the vocoder part separately. In contrast, the proposed architecture has the benefit of end-to-end training and obtains optimal latent acoustic features guided by the auxiliary pitch information. In addition, another prior work has tackled the pitch stability problem by adopting a chunked autoregressive architecture in the vocoder [15]. Unlike that method, our proposal can generate waveforms in much faster speed, thanks to the fully non-autoregressive model architecture.

The experimental results show that the proposed model performed significantly better than all the baseline models including end-to-end and cascade models in terms of naturalness in the multi-speaker emotional TTS task. Moreover, the proposed model achieved comparable scores to the recordings for neutral and sad style with no statistically significant difference.
During inference, the latent feature from the posterior distribution and the predicted pitch distribution on text. The VAEB is composed of a posterior encoder and a decoder, whereas the prior distribution is modeled by a prior encoder. In addition, we introduce a latent variable to represent the periodicity generator described in Section 2.3. As discussed in Section 2.1, these features are optimized using the L2 norm, as follows:

$$L_{\text{pitch}} = \| \log F_0 - \log \hat{F}_0 \|_2 + \| v - \hat{v} \|_2.$$  

The prior distribution is augmented by normalizing flow $f$ to enhance the modeling capability, as in VITS:

$$p(z|c) = \mathcal{N}(f(z); \mu(c), \sigma(c)) \left| \frac{\partial f(z)}{\partial z} \right|,$$

where $\mu(c)$ and $\sigma(c)$ represent trainable mean and variance parameters calculated from text representation, respectively.

### 2.3. Decoder with periodicity generator

It has been reported that GAN-based vocoder models typically produce artifacts when reconstructing waveforms from acoustic features due to their inability to estimate pitch and periodicity [15]. We found that these artifacts are also observed in end-to-end TTS models, particularly when trained on a dataset with large pitch variance, such as an emotional one. To address this problem, we use a sine-based source signal to explicitly model the periodic component of speech waveforms, which has proven to be effective in some previous works [16, 17, 19]. However, it is not straightforward to incorporate it into the HiFi-GAN-based decoder (i.e., vocoder) architecture.

This assumption is reasonable considering the fact that the pitch and the VAE latent feature are obtained from the waveform separately (tail-to-tail connection in the graphical model).

In our preliminary experiment, we also investigated cross-entropy loss for $v$ instead of L2 loss and found that both criteria performed equally well.
3. EXPERIMENTS

3.1. Experimental setup

3.1.1. Database and feature extraction settings

For the experiments, we used phonetically-balanced emotional speech corpora recorded by five male and ten female Japanese professional speakers. We sampled speech signals at 24 kHz with 16 bit quantization. All the 15 speakers’ data contained three speaking styles: neutral, happy, and sad. The number of utterances for each style was 4000, 1000, and 1000, for each speaker, respectively. For every speaker–style pair, 50 utterances were kept for each validation and test, respectively. The rest were used for training, which amounted to 84.08 hours. We extracted a 513-dimensional linear-spectrogram for end-to-end models and a 80-dimensional log-mel spectrogram for cascade models with a 10 ms frame shift and a 40 ms window length. We also extracted continuous log $F_0$ and voicing flags using the improved time-frequency trajectory excitation vocoder [23, 24]. Acoustic features for the cascade models were constructed by concatenating these pitch features to the log-mel spectrogram. The phoneme durations to train the supervised duration model were manually labeled by professional annotators. We normalized the acoustic features so that they had a zero mean and unit variance for each dimension using the statistics of the training data.

3.1.2. Model details

The training configurations of the proposed method basically followed those of the original VITS [10]. Regarding the loss criteria in Section 2.4, we set the weights of $L_{\text{recon}}, L_{\text{kl}}, L_{\text{pitch}}, L_{\text{dur}}, L_{\text{adv}}$ and $L_{\text{fm}}$ empirically to 45, 1, 1, 1, and 2, respectively. As our model is a multi-speaker emotional TTS model, we used speaker and emotion embeddings as the global conditions. In addition, we used accent information as external input to synthesize speech with natural prosody for the Japanese language [25].

The frame prior network was composed of six 1-D residual convolutional stacks with a kernel size of 17. The frame pitch predictor was composed of a stack of five 1-D convolutions with a kernel size of 5 and a dropout rate of 0.3. Conditional embedding was also added to predict the speaker-dependent pitch contour. The sine-wave from $F_0$ was generated using an open-source implementation of neural source-filter models.Voicing flags in the periodicity generator were up-sampled from the frame level to the sample level using nearest neighbor up-sampling. We used up-sample rates [6, 5, 2, 2, 2] for the decoder, as one frame contains 240 samples. During training, we used a dynamic batch size with an average of 26 samples to create a minibatch [27].

Baseline systems are summarized in Table 1. We prepared several variants of VITS: VITS is the original VITS; FPN-VITS is a model that introduces a frame prior network to VITS; and CAT-P-VITS is similar to FPN-VITS, but additionally concatenated the frame-level $F_0$ and voicing flags to the latent variable $z$. Note that the periodicity generator was not used in this model. Sine-P-VITS was similarly configured to the proposed method P-VITS, but it omitted voicing flags and Gaussian noise from the input of the periodicity generator.

We also prepared the cascade-type TTS model FS2+P-HiFi-GAN, as the baseline system. Specifically, we adopted a FastSpeech 2-based acoustic model [8] and a HiFi-GAN-based vocoder [6]. The acoustic model used six Transformer layers and six lightweight convolution blocks to compose the encoder and decoder, respectively. Note that the lightweight convolution blocks perform comparably to or better than the Transformer ones in the TTS context [28]. For each block, we set the hidden sizes of the self-attention and convolution layers to 384. The hidden size of feed-forward layers in Transformer was 768. To bring the condition closer to end-to-end models, pitch and energy predictors in a variance adaptor were

![Decoder architecture with periodicity generator. MRF denotes multi-receptive field fusion in HiFi-GAN [6]. Up/down-sampling is performed by transposed and normal 1-D convolution, respectively.](https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts)
We proposed Period VITS, a fully end-to-end TTS system with explicit pitch modeling capable of high-quality speech synthesis even when the dataset contained significant prosodic diversity. Specifically, the proposed method introduced a periodicity generator that generated sample-level pitch representation from extracted frame-level pitch. The representation was successively down-sampled and added to the HiFi-GAN-based decoder part to stabilize the pitch of the target waveform. The proposed model was optimized in the end-to-end schema from a variational inference point of view. The experimental results showed that our proposed method outperformed the baseline methods in terms of emotional TTS. Future work should include implicit pitch modeling to avoid pitch extraction errors and to improve the sample quality for highly expressive styles such as the happy style.

5. ACKNOWLEDGMENTS

This work was supported by Clova Voice, NAVER Corp., Seongnam, Korea.

6https://yshira116.github.io/period_vits_demo/
6. REFERENCES

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech synthesis,” arXiv preprint arXiv:2106.15361, 2021.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech: Fast, robust and controllable text to speech,” in Proc. NeurIPS, 2019, vol. 32.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “TacoTone: Towards end-to-end speech synthesis,” in Proc. Interspeech, 2017, pp. 4006–4010.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[5] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020, pp. 6199–6203.

[6] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in Proc. NeurIPS, 2020, pp. 17022–17033.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. ICASSP, 2018, pp. 4779–4783.

[8] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech 2: Fast and high-quality end-to-end text-to-speech,” in Proc. ICLR, 2021.

[9] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, “End-to-end adversarial text-to-speech,” in Proc. ICLR, 2021.

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

[11] D. Lim, S. Jung, and E. Kim, “JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech,” in Proc. Interspeech, 2022.

[12] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in Proc. ICLR, 2014.

[13] D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in Proc. ICML, 2015, pp. 1530–1538.

[14] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, “ViSinger: Variational inference with adversarial learning for end-to-end singing voice synthesis,” in Proc. ICASSP, 2022, pp. 7237–7241.

[15] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, “Chunked autoregressive GAN for conditional waveform synthesis,” in Proc. ICLR, 2022.

[16] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components,” in Proc. ICASSP, 2021, pp. 6049–6053.

[17] M.-J. Hwang, R. Yamamoto, E. Song, and J.-M. Kim, “High-fidelity Parallel WaveGAN with multi-band harmonic-plus-noise model,” in Proc. Interspeech, 2021, pp. 2227–2231.

[18] R. Yoneyama, Y.-C. Wu, and T. Toda, “Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation,” in Proc. Interspeech, 2022, pp. 848–852.

[19] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter-based waveform model for statistical parametric speech synthesis,” in Proc. ICASSP, 2019, pp. 5916–5920.

[20] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, “Period-HiFi-GAN: Fast and fundamental frequency controllable neural vocoder,” in Proc. Acoustic Society of Japan (in Japanese), 2022, pp. 901–904.

[21] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in ICCV, 2017, pp. 2794–2802.

[22] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in Proc. ICML, 2016, pp. 1558–1566.

[23] K. Yu and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” IEEE Trans. on Audio, Speech, and Lang. Process., vol. 19, no. 5, pp. 1071–1079, 2010.

[24] E. Song, F. K. Soong, and H.-G. Kang, “Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems,” IEEE/ACM Trans. Audio, Speech, and Lang. Process., vol. 25, no. 11, pp. 2152–2161, 2017.

[25] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language,” in Proc. ICASSP, 2019, pp. 6905–6909.

[26] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter waveform models for statistical parametric speech synthesis,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process., vol. 28, pp. 402–415, 2019.

[27] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in Proc. ICASSP, 2020, pp. 7654–7658.

[28] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, “Parallel Tacotron: Non-autoregressive and controllable TTS,” in Proc. ICASSP, 2021, pp. 5709–5713.

[29] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” in Proc. ICLR, 2020.

[30] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. ICLR, 2019.