LVCNET: EFFICIENT CONDITION-DEPENDENT MODELING NETWORK FOR WAVEFORM GENERATION

Zhen Zeng, Jianzong Wang*, Ning Cheng, Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd.

ABSTRACT

In this paper, we propose a novel conditional convolution network, named location-variable convolution, to model the dependencies of the waveform sequence. Different from the use of unified convolution kernels in WaveNet to capture the dependencies of arbitrary waveform, the location-variable convolution uses convolution kernels with different coefficients to perform convolution operations on different waveform intervals, where the coefficients of kernels is predicted according to conditioning acoustic features, such as Mel-spectrograms. Based on location-variable convolutions, we design LVCNet for waveform generation, and apply it in Parallel WaveGAN to design more efficient vocoder. Experiments on the LJSpeech dataset show that our proposed model achieves a four-fold increase in synthesis speed compared to the original Parallel WaveGAN without any degradation in sound quality, which verifies the effectiveness of location-variable convolutions.

Index Terms— speech synthesis, waveform generation, vocoder, location-variable convolution

1. INTRODUCTION

In rencet years, deep generative models have witnessed extraordinary success in waveform generation, which promotes the development of speech synthesis systems with human-parity sounding. Early researches on autoregressive model in waveform synthesis, such as WaveNet [1] and WaveRNN [2], have shown much superior performance over traditional parameters vocoders. However, low inference efficiency of autoregressive neural network limits its application in real-time scenarios.

In order to address the limitation and improve the generation speed, many non-autoregressive models have been studied to generate waveforms in parallel. One family relies on knowledge distillation, including Parallel WaveNet [3] and Clarinet [4], where an parallel feed-forward network is distilled from an autoregressive WaveNet model based on the inverse auto-regressive flows (IAF) [5]. Although the IAF models is capable of generating high-fidelity speech in real time, the requirement for a well-trained teacher model

and the intractable density distillation lead to a complicated model training process. The other family is flow-based generation models, including WaveGlow [6] and WaveFlow [7]. They are implemented by a invertible network and trained using only a single likelihood loss function on the training data. While inference is fast on high-performance GPU, the large size of mode limits their application in memory-constrained scenarios. Meanwhile, as a family of generation models, Generative Adversarial Network (GAN) [8] is also applied in waveform generation, such as MelGAN [9], Parallel WaveGAN [10] and Multi-Band MelGAN [11], in which a generator is designed to produce samples as close as possible to real speech, and a discriminator is implemented to distinguish generated speech from real speech. They have a very small amount of parameters, achieve a synthesis speech far exceeding real-time. Impressively, the Multi-band MelGAN [11] runs at more than 10x faster than real-time on CPU. In addition, WaveGrad [12] and DiffWave [13] apply diffusion probabilistic models [14] for waveform generation, which converts the white noise signal into structured waveform in an interative manner.

These models are almost implemented by an wavenet-like network, in which the dilated causal convolution is applied to capture the long-term dependencies of waveform, and the mel-spectrum is used as the local conditional input for the gated activation unit. In order to efficiently capture time-dependent features, a large number of convolution kernels are required in wavenet-like network. In this work, we propose the location-variable convolution to model time-dependent features more efficiently. In detail, the location-variable convolution uses convolution kernels with different coefficients to perform convolution operations on different waveform intervals, where the coefficients of kernels is predicted by a kernel predictor according to conditioning acoustic features, such as mel-spectrograms. Based on location-variable convolutions, we design LVCNet for waveform generation, and apply it in Parallel WaveGAN to achieve more efficient vocoder. Experiments on the LJSpeech dataset [15] show that our proposed model achieves a four-fold increase in synthesis speed without any degradation in sound quality. [1]

And the main contributions of our works as follow:

---

*Corresponding author: Jianzong Wang, jzwang@188.com

[1]Audio samples in `https://github.com/ZENGZHEN-TTS/` `LVCNet`

- A novel convolution method, named location-variable convolution, is proposed to efficiently model the time-dependent features, which uses different convolution kernels to perform convolution operations on different waveform intervals;

- Based on location-variable convolutions, we design a network for waveform generation, named LVCNet, and apply it in Parallel WaveGAN to achieve more efficient vocoder;

- A comparative experiment was conducted to demonstrate the effectiveness of the location-variable convolutions in waveform generation.

2. PROPOSED METHODS

In order to model the long-term dependencies of waveforms more efficiently, we design a novel convolution network, named location-variable convolution, which is applied to the Parallel WaveGAN to verify its performance. The design details are described in the section.

2.1. Location-Variable Convolution

In the traditional linear prediciton vocoder [16], a simple all-pole linear filter is used to generate waveform in autoregressive way, of which the linear prediction coefficients is calculated according to the acoustic features. This process is similar to the autoregressive wavenet vocoder, except that the coefficients of the linear predictor is variable for different frames while the coefficients of convolution kernels in wavenet is the same in all frames. Inspired by this, we try to design a novel convolution network with variable convolution kernel coefficients in order to improve the ability to model long-term dependencies for waveform generation.

Define the input sequence to the convolution as $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, and define the local conditioning sequence as $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$. An element in the local conditioning sequence is associated with a continuous interval in the input sequence. In order to effectively use the local correlation to model the feature of the input sequence, the location-variable convolution uses a novel convolution method, where different intervals in the input sequence use different convolution kernels to implement the convolution operation. In detail, a kernel predictor is designed to predict multiple sets of convolution kernels according to the local conditioning sequence. Each element in the local conditioning sequence corresponds to a set of convolution kernels, which is used to perform convolution operations on the associated intervals in the input sequence. In other words, elements in different intervals of the input sequence use their related and corresponding convolution kernels to extract features. And the output sequence is spliced by the convolution results on each interval.
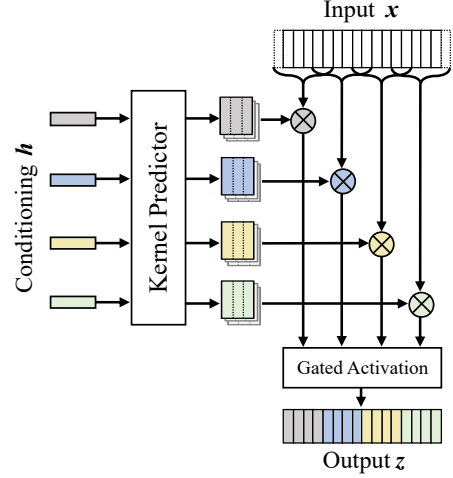

Fig. 1. An example of convolution process in the location-variable convolution. According to the conditioning sequence, the kernel predictor generates multiple sets of convolution kernels, which are used to perform convolution operations on the associated intervals in the input sequence. Each element in the conditioning sequence corresponds to 4 elements in the input sequence.

Similar to WaveNet, the gated activation unit is also applied, and the local condition convolution can be expressed as

$$\{\boldsymbol{x}_{(i)}\}_m = \text{split}(\boldsymbol{x}) \tag{1}$$

$$\{\boldsymbol{W}^f_{(i)}, \boldsymbol{W}^g_{(i)}\}_m = \text{Kernel Predictor}(\boldsymbol{h}) \tag{2}$$

$$\boldsymbol{z}_{(i)} = \tanh(\boldsymbol{W}^f_{(i)} * \boldsymbol{x}_{(i)}) \odot \sigma(\boldsymbol{W}^g_{(i)} * \boldsymbol{x}_{(i)}) \tag{3}$$

$$\boldsymbol{z} = \text{concat}(\boldsymbol{z}_{(i)}) \tag{4}$$

where $\boldsymbol{x}_{(i)}$ denotes the intervals of the input sequence associated with h_i, $\boldsymbol{W}^f_{(i)}$ and $\boldsymbol{W}^g_{(i)}$ denote the filter and gate convolution kernels for $\boldsymbol{x}_{(i)}$.

For a more visual explanation, Figure 1 shows an example of the location-variable convolution. In our opinion, since location-variable convolutions can generate different kernels for different conditioning sequences, it has more powerful capability of modeling the long-term dependency than traditional convolutional network. We also experimentally analyze its performance for waveform generation in next section.

2.2. LVCNet

By stacking multiple layers of location-variable convolutions with different dilations, we design the LVCNet for waveform generation, as shown in Figure 2(a). The LVCNet is composed of multiple LVCNet blocks, and each LVCNet block contains multiple location-variable convolution (LVC) layers with inscreasing factorial dilation coefficients to improve the
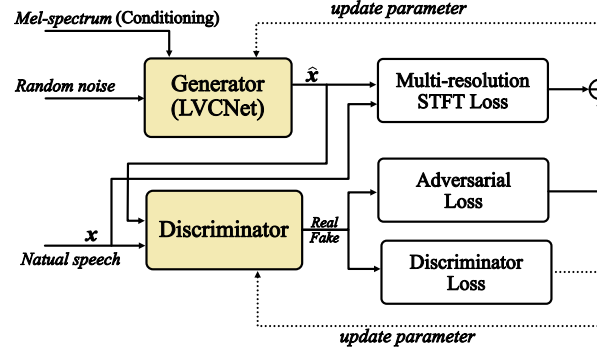
(a) Architecture of LVCNet

(b) Architecture of Parallel WaveGAN with LVCNet

Fig. 2. (a) The architecture of LVCNet, which is composed of multiple LVCNet blocks, and each LVCNet block contains multiple location-variable convolution (LVC) layers with ins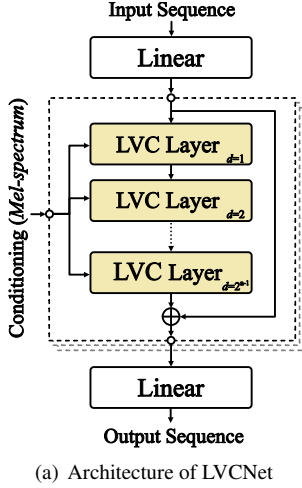creasing factorial dilation coefficients to improve the receptive field. (b) The architecture of Parallel WaveGAN with LVCNet. The generator is implemented using LVCNet.

receptive field. A linear layer is applied on the input and output sides of the network to achieve channel conversion. The residual connection is deployed in each LVCNet block instead of each LVC layer, which can achieve more stable and satisfactory results according to our experimental analysis. In addition, the conditioning sequence is input into the kernel predictor to predict coefficients of convolution kernels in each LVC layer. The kernel predictor consists of multiple residual linear layer with leaky ReLU activation function ($\alpha = 0.1$), of which the output channel is determined by the amount of convolution kernel coefficients to be predicted.

2.3. Parallel WaveGAN with LVCNet

In order to verify the performance of location-variable convolutions, we choose Parallel WaveGAN as the baseline model, and use the LVCNet to implement the network of the generator, as shown in Figure 2(b). The generator is also conditioned on the mel-sepctrum, and transforms the input noise to the output waveform. For a fair comparison, the discriminator of our model maintains the same structure as that of Parallel WaveGAN, and the same loss function and training strategy as Parallel WaveGAN are used to train our model.

Note that, the design of the kernel predictor is based on the correspondence between the mel-spectrum and the waveform. A 1D convolution with 5 of kernel size and zero of padding is firstly used to adjust the alignment between the conditioning mel-spectrum sequence and the input waveform sequence, and subsequent multiple stacked 1×1 convolutional layers to output convolution kernels of the location-variable convolutions. In addition, we remove the residual connection of the first LVCNet block in the generator for better model performance.

3. EXPERIMENTS

3.1. Experimental setup

3.1.1. Database

We train our model on the LJSpeech dataset [15]. The entire dataset is randomly divided into 3 sets: 12,600 utterances for training, 400 utterances for validation, and 100 utterances for test. The sample rate of audio is set to 22,050 Hz. The mel-spectrograms are computed through a short-time Fourier transform with Hann windowing, where 1024 for FFT size, 1024 for window size and 256 for hop size. The STFT magnitude is transformed to the mel scale using 80 channel mel filter bank spanning 80 Hz to 7.6 kHz.

3.1.2. Model Details

In our proposed model, the generator is implemented by the LVCNet, and the network structure of the discriminator is consistent with that in original Parallel WaveGAN. The generator is composed of three LVCNet blocks, where each blocks contains 10 LVC layers, and the residual channels is set to 8. The kernel size of the location-variable convolution is set to three, and the dilation coefficients is the factorial of 2 in each LVCNet block. The kernel predictor consists of one 1×5 convolutional layer and three 1×1 residual convolutional layers, where the hidden residual channel is set to 64. The weight normalization is applied in all convolutional layer.

We choose Parallel WaveGAN [10] as the baseline model, and use the same strategy to train our model and baseline model for a fair comparison. In addition, in order to verify the efficiency of the location-variable convolution in modeling waveform dependencies, we conduct a set of detailed

comparison experiments. Our proposed model is trained and compared with the Parallel WaveGAN under different residual channels (4, 8, 16 for our model and 32, 64, 128 for Parallel WaveGAN).

3.2. Results

3.2.1. Evaluation

In order to evaluate the performance of these vocoder, we use the mel-spectrograms extracted from test utterances as input to obtain synthetic audios, which is rated together with the ground truth audio (GT) by 20 testers with headphones in a conventional mean opinion score (MOS) evaluation. At the same time, the audios generated by Griffin-Lim algorithm [17] are also rated together.

The scoring results of our proposed model and Parallel WaveGAN with different residual convolutional channels are shown in Table 1, where the real-time factor implemented (RTF) on CPU is also illustrated. We find that our proposed model achieves almost the same results as Parallel Wave-GAN, but the inference speed is increased by four times. The reasan for the speech increase is that small number fo residual channels greatly reduces the amount of convolution operations in our model. Meanwhile, our unoptimized model can synthesizes multiple utterances at approximately 300 MHz on an NVIDIA V100 GPU, which is much faster than 35 MHz of Parallel WaveGAN.

In addition, as the residual channels decreases, the rate of performance degradation of our model is significantly slower than that of Parallel WaveGAN. In our opinion, even if the residual channels is very small (such as 4, 8), the convolution coefficients are adjusted according to mel-spectrums in our model, which still guarantees effective feature modeling.

3.2.2. Text-to-Speech

To verify the effectiveness of the proposed model as the vocoder in the TTS framework, we combine it with Transformer TTS [18] and AlignTTS [19] for testing. In detail, according to the texts in the test dataset, Transformer TTS and AlignTTS predict mel-spectrums respectively, which is used as the input of our model (with 8 of residual channels) and Parallel WaveGAN to generate waveforms for MOS evaluation. The results are shown in Table 2. Compared with Parallel WaveGAN, our model significantly improves the speed of speech synthesis without degradation of sound quality in feed-forward TTS systems.

In our opinion, due to the mutual independence of the acoustic features (such as mel-specturms), we can use difference convolution kernels to implement convolution operations on difference time intervals to obtain more effective feature modeling capabilities. In this work, we just use the LVC-Net to design a new generator for waveform generation, and

Table 1. The comparison between our proposed model (LVC-Net) and Parallel WaveGAN (PWG) with different residual channels.

Method	Size	MOS	RTF (CPU)
GT		4.56 ± 0.05	−
Griffin-Lim		3.45 ± 0.24	−
PWG-32	0.44 M	3.62 ± 0.23	2.16
PWG-48	0.83 M	4.03 ± 0.10	3.05
PWG-64	1.35 M	4.15 ± 0.08	3.58
LVCNet-4	0.47 M	3.96 ± 0.15	0.53
LVCNet-6	0.84 M	4.09 ± 0.12	0.62
LVCNet-8	1.34 M	4.15 ± 0.10	0.73

Table 2. The comparison between our proposed model (LVC-Net) and Parall WaveGAN (PWG) in TTS systems.

Method	MOS	Time (s)
GT	−	−
Transformer + PWG	4.08 ± 0.14	2.76 ± 0.94
AlignTTS + PWG	4.07 ± 0.09	0.09 ± 0.01
Transformer + LVCNet	4.07 ± 0.15	2.70 ± 0.88
AlignTTS + LVCNet	4.09 ± 0.11	$\mathbf{0.06 \pm 0.01}$

obtain a faster vocoder without degradation of sound quality. Considering our previous experiments [20], the effectiveness of the location-variable convolution has been sufficiently demonstrated, and there is potential for optimization.

4. CONCLUSION

In this work, we propose the location-variable convolution for time-dependent feature modeling. which uses different kernels to perform convlution operations on different intervals of input sequence. Based on it, we design LVCNet and implement it as the generator of Parallel WaveGAN framework to achieve more efficient waveform generation model. Experiments on LJSpeech dataset show that our proposed model is four times faster than the base Parallel WaveGAN model in inferece speed without any degradation in sound quality, which verifies the effectiveness of the location-variable convolution.

5. ACKNOWLEDGMENT

6. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016.

[2] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning (ICML)*, 2018.

[3] Aaron van den Oord, Yazhe Li, and et. al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning (ICML)*, 2018.

[4] Wei Ping, Kainan Peng, and Jitong Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations (ICLR)*, 2018.

[5] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems (NIPS)*. 2016.

[6] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[7] Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song, "WaveFlow: A compact flow-based model for raw audio," in *International Conference on Machine Learning (ICML)*, 2020.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[9] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems (NIPS)*. 2019.

[10] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on genera-tive adversarial networks with multi-resolution spectrogram," in *International conference on acoustics, speech and signal processing (ICASSP)*, 2020.

[11] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[12] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations (ICLR)*, 2021.

[13] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[15] Keith Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[16] Bishnu S Atal and Suzanne L Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[17] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[18] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and M Zhou, "Neural speech synthesis with transformer network," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[19] Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao, "Aligntts: Efficient feed-foward text-to-speech system without explicit alignment," in *International conference on acoustics, speech and signal processing (ICASSP)*, 2020.

[20] Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao, "MelGlow: Efficient waveform generative network based on location-variable convolution," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.