FloWaveNet : A Generative Flow for Raw Audio

Sungwon Kim [1] Sang-gil Lee [1] Jongyoon Song [1] Sungroh Yoon [1]

Abstract

Most of modern text-to-speech architectures use a WaveNet vocoder for synthesizing a high-fidelity waveform audio, but there has been a limitation for practical applications due to its slow autoregressive sampling scheme. A recently suggested Parallel WaveNet has achieved a real-time audio synthesis by incorporating Inverse Autoregressive Flow (IAF) for parallel sampling. However, the Parallel WaveNet requires a two-stage training pipeline with a well-trained teacher network and is prone to mode collapsing if using a probability distillation training only. We propose FloWaveNet, a flow-based generative model for raw audio synthesis. FloWaveNet requires only a single maximum likelihood loss without any additional auxiliary terms and is inherently parallel due to the flow-based transformation. The model can efficiently sample the raw audio in real-time with a clarity comparable to the original WaveNet and ClariNet. Codes and samples for all models including our FloWaveNet is available via GitHub:
`https://github.com/ksw0306/FloWaveNet`

1. Introduction

End-to-end waveform audio synthesis model is a core component of text-to-speech(TTS) systems. With a striking success of deep learning-based raw audio synthesis architectures, modern text-to-speech systems leverage them as a vocoder which can synthesize a hyper-realistic waveform signal that is nearly indistinguishable to the real-world natural sound.

Current state-of-the-art text-to-speech architectures commonly use the WaveNet vocoder with a Mel-spectrogram as an input for a high-fidelity audio synthesis (Shen et al., 2018; Arik et al., 2017b;a; Ping et al., 2017; Jia et al., 2018). However, there has been a limitation for practical applications because the WaveNet requires an autoregressive sampling scheme, which becomes a major bottleneck for a real-time waveform generation. A number of variations have been proposed to overcome the slow sampling of the original WaveNet. Parallel WaveNet (Oord et al., 2017) has achieved a real-time audio synthesis by incorporating Inverse Autogressive Flow (IAF) (Kingma et al., 2016) for the parallel audio synthesis. A recently suggested ClariNet (Ping et al., 2018) presented an alternative formulation by using a single Gaussian which has a closed-form KL divergence in contrast to a high variance Monte Carlo approximation from the Parallel WaveNet.

Despite the success of the real-time high-fidelity waveform audio synthesis, however, all of the aforementioned approaches require a two-stage training pipeline with a well-performing pre-trained teacher network for a probability distillation training strategy. Furthermore, in practical terms, these models are harder to train without using highly-tuned additional auxiliary losses. For example, if using the probability distillation loss only, the Parallel WaveNet is prone to a mode collapsing problem where the student network converge to a certain mode of the teacher distribution, resulting in a sub-optimal performance.

Here we present FloWaveNet, an alternative flow-based approach of a real-time parallel generative model for the raw audio synthesis. The FloWaveNet requires only a single maximum likelihood loss without any auxiliary loss terms while maintaining the stability of training. It features a simplified single-stage training scheme because it does not require a teacher network and can be trained end-to-end. The model is inherently parallel due to the flow-based transformation, which enables the real-time waveform generation. The FloWaveNet can act as a drop-in replacement for the WaveNet vocoder used in a variety of text-to-speech architectures.

Along with all the advantages described above, the quality and fidelity of samples from the FloWaveNet are comparable to the two-stage models. Codes and samples for all models including our FloWaveNet are available via GitHub:
`https://github.com/ksw0306/FloWaveNet`

[1]Electrical and Computer Engineering, Seoul National University, Seoul, Korea. Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

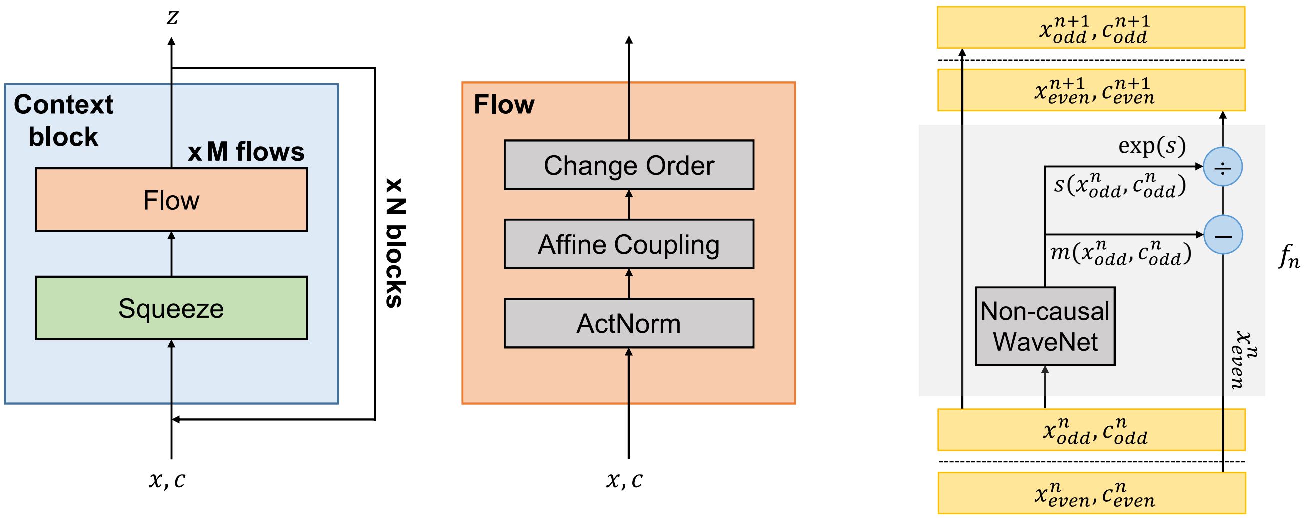

Figure 1. Schematic diagram of FloWaveNet. Left: an entire forward pass of the FloWaveNet consisting of N context blocks. Middle: an abstract diagram of the Flow operation. Right: a detailed version of the affine coupling operation.

2. FloWaveNet

Here we describe our FloWaveNet which learns to maximize the exact likelihood of the data while also not losing the ability of the real-time parallel sampling, instead of the two-stage training from related work. The FloWaveNet is a hierarchical architecture composed of a context block as a higher abstract module and multiple reversible transformations inside the context block.

2.1. Flow based generative model

FloWaveNet is a flow-based generative model using a normalizing flow (Rezende & Mohamed, 2015) to model a raw audio data. Given a waveform audio signal x, assume there is an invertible transformation function $f(x) : x \longrightarrow z$ that directly maps the signal into a known prior z. We can explicitly calculate the log probability distribution of x from the prior z by using a change of variables technique as follows:

$$\log P_X(x) = \log P_Z(f(x)) + \log\det(\frac{\partial f(x)}{\partial x}) \qquad (1)$$

The flow-based generative model can realize the reversible transformation by fulfilling following properties: (i) Calculation of a Jacobian matrix of a determinant of the transformation f should be tractable, (ii) From the randomly sampled known prior z, mapping it back to the data x by applying the inverse transformation $x = f^{-1}(z)$ should be efficient enough to compute. The parallel sampling becomes computationally tractable only if the above properties are satisfied.

To construct a parametric invertible transformation f that fulfills both properties, the FloWaveNet uses an affine coupling layer suggested in (Dinh et al., 2016). To model the data using a transformation complex and flexible enough for audio, the FloWaveNet stacks multiple flow operations composed of the WaveNet affine coupling layers.

The change of variables formula in equation (1) holds for a conditional distribution as well. In practice, the WaveNet (Van Den Oord et al., 2016) additionally receives the Mel-spectrogram as a local condition c for the network to model the conditional probability $p(x|c)$.

2.2. Affine Coupling Layer

A usual flow-based neural density estimator focuses sorely on a density estimation of the data as a main objective. The neural density estimator family has an advantage of using much more flexible transformation such as Masked Autoregressive Flow (Papamakarios et al., 2017) and Transformation Autoregressive Network (Oliva et al., 2018), etc. However, it only satisfies the property (i) and not (ii), making it unusable for our purpose.

In contrast, the affine coupling layer is a parametric layer suggested in Real NVP (Dinh et al., 2016) and satisfies both (i) and (ii) and can sample from z efficiently in parallel. The layer enables the efficient bidirectional transformation of f by making the transformation function bijective while maintaining the computational tractability. Each layer is the parametric transformation $f_n : x^n \longrightarrow x^{n+1}$ which keeps a half of the channel dimension identical and applies the affine transformation only on the remaining half, as the following:

$$x_{odd}^{n+1} = x_{odd}^{n} \tag{2}$$

$$x_{even}^{n+1} = \frac{x_{even}^{n} - m(x_{odd}^{n}, c_{odd}^{n})}{\exp(s(x_{odd}^{n}, c_{odd}^{n}))}, \tag{3}$$

Similarly, for the inverse transformation $f_n^{-1} : x^{n+1} \longrightarrow x^n$, we have:

$$x_{odd}^{n} = x_{odd}^{n+1} \tag{4}$$

$$x_{even}^{n} = x_{even}^{n+1} \odot \exp s(x_{odd}^{n+1}, c_{odd}^{n+1}) + m(x_{odd}^{n+1}, c_{odd}^{n+1}), \tag{5}$$

where the m and s is a shared neural network with the non-causal WaveNet architecture composed of a stack of dilated convolutional layers. The shared trainable neural network m and s satisfies the property (ii), which endows the system the ability of the efficient sampling from f^{-1}. The Jacobian matrix is lower triangular and the determinant is a product of the diagonal elements, which satisfies the property (i).

$$\log\det(\frac{\partial x^{n+1}}{\partial x^{n}}) = -s, \tag{6}$$

A single affine coupling layer does not alter half of the feature by keeping it identical. To construct more flexible transformation f, the FloWavenet stacks multiple flow operations for each context block. The change order operation at the end of each flow swaps the order of x_{odd} and x_{even} so that all channels can affect each other during subsequent flow operations.

2.3. Context Block

Context block is a main module of the FloWaveNet. Each context block consists of a squeeze operation followed by stacks of flow. The squeeze operation takes the data and condition, then doubles the channel dimension by splitting the time dimension in half. This operation doubles an effective receptive field per block for the WaveNet-based flow, which is conceptually similar to the dilated convolutions of the WaveNet itself. Thus, upper-level blocks can learn a long-term characteristics of audio while lower-level blocks can model a high-frequency information. The flow operation inside the block contains activation normalization, affine coupling layer, and change order operation described above.

2.4. Activation Normalization

Activation normalization (ActNorm) layer suggested in Glow (Kingma & Dhariwal, 2018) stabilizes the training of the network composed of multiple flow operations. The ActNorm layer is a per-channel parametric affine transformation at the beginning of the flow. The layer performs a data dependent initialization of the trainable parameters by scaling the activation channel-wise to have zero mean and unit variance for the first given batch of data.

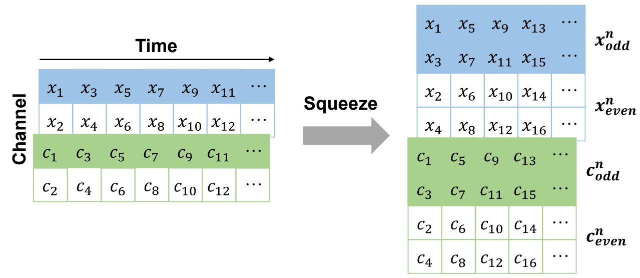

Figure 2. Squeeze operation used in the context block.

3. Experiments

We trained the model using LJSpeech dataset (Ito, 2017) which is a 24-hour waveform audio of a single female speaker with 13,100 audio clips and 22kHz sample rate. We extracted a random 6,400 sample chunks and normalized to $[-1, 1]$ as the input. For the local conditioning with the Mel-spectrogram construction, we used a preprocessing method from Tacotron 2 (Shen et al., 2018). The generated 80-band Mel-spectrogram is used for the network to estimate the conditional probability.

We used the original autoregressive WaveNet (Van Den Oord et al., 2016) and the recently proposed ClariNet (Ping et al., 2018) with a Gaussian IAF as baselines. All models are trained to estimate the probability distribution of raw audio $\log p(x|c)$, given the Mel-spectrogram condition. We performed random sampling of the audio from the estimated probabilities for experiments.

We used an Adam optimizer (Kingma & Ba, 2014) with a learning rate of 10^{-3} for all models same as the ClariNet training configuration. We scheduled the learning rate decay by a factor of 0.5 for every 200K iterations. We used a single NVIDIA TITAN V GPU with a batch size of 8, 4, and 8 for the WaveNet, ClariNet, and FloWaveNet due to a memory constraint of the ClariNet.

3.1. Autoregressive Wavenet

We trained the best-performing original autoregressive WaveNet from Tacotron 2, which is a 24-layer architecture with four 6-layer dilation cycles. We experimented with both Mixture of Logistics (MoL) and single Gaussian formulation for the output distribution, with the configuration closely following the Tacotron 2 (Shen et al., 2018) and ClariNet (Ping et al., 2018) for each model. We trained the models for a total of 900K iterations.

3.2. Gaussian Inverse Autoregressive Flow (IAF)

For the Gaussian IAF from the ClariNet, we used the best-performing pre-trained Gaussian autoregressive WaveNet from subsection 3.1 as the teacher network for the probability density distillation. We used the transposed convolution parameters from the teacher network without further tuning for upsampling the Mel-spectrogram condition.

The student network has an architecture similar to the teacher network. It has a 42-layer architecture with four stack of IAF modules, each of them having 1, 1, 1, and 4 blocks of a 6-layer dilation cycle.

The ClariNet training requires a regularized KL divergence loss and a spectrogram frame loss. We performed an analysis of an impact of each loss by additionally training models with only one of the two losses.

3.3. FloWaveNet

The FloWaveNet has 8 context blocks. Each block has 6 flows which results in a total of 48 stack of flows. We used the affine coupling layer with a 2-layer non-causal WaveNet (Van Den Oord et al., 2016) with a kernel size of 3 for each flow. We used 256 channels for a residual, skip, and gate channel of the WaveNet architecture with a gated tanh activation unit, along with the Mel-spectrogram condition. The weights for the convolutional layers are initialized with zero which reportedly showed a better training result. (Kingma & Dhariwal, 2018)

We induced the model to learn the long-term dependency of audio by stacking many context blocks instead of increasing the dilation cycle of the affine coupling. We trained the model with a single maximum likelihood loss without any auxiliary terms for 1,000K iterations.

4. Results and Analysis

Our preliminary results with the sampled audio for all models are available via the GitHub link. We are planning a quantitative performance measurement with a Mean Opinion Score (MOS) for a future version of the draft.

The Gaussian IAF and the FloWaveNet generate the waveform audio by applying the normalizing flow from the randomly sampled known prior z. We set the variance of the prior, *i.e.* a temperature, below 1 which generated relatively higher-quality audio. We chose temperature of 0.7 as default which empirically showed the best quality for both models.

4.1. Model Comparisons

Among the WaveNet (MoL or Gaussian), Gaussian IAF, and FloWaveNet, the MoL WaveNet generated the highest fidelity audio. It was difficult to decide a clear winner from the remaining models in terms of the fidelity, each of the models having a distinct characteristic of the generated audio.

For the Gaussian WaveNet, we were able to generate a high-quality sample from the experiment for certain sentences. However, most of the sample contained a high frequency noise and there were high pitch artifacts during silence between words regardless of the overall quality of the sampled speech.

The Gaussian IAF had a tendency of a high fidelity and clear sound, but with a constant white noise across all the samples. The amount of the white noise varied for different values of the temperature.

The FloWaveNet did not incur the white noise and had a clear sound quality unlike the Gaussian IAF, but instead there was a periodic artifact perceived as a trembling voice which also varied for different temperatures.

Overall, the sound quality of the FloWaveNet was comparable to the previous approaches as well as having the advantages of the single maximum likelihood loss and the single-stage training. Note that although the MoL WaveNet showed the highest fidelity, it requires the ancestral sampling (50 minutes of GPU time for a 7 second audio clip). The parallel models such as the ClariNet and our FloWaveNet only require approximately 0.5 seconds for the 7 second clip.

4.2. Temperature Effect on Audio Quality Trade-off

To analyze the characteristics of the sampled audio according to the temperature, we performed a parameter sweep from 0 to 1 with a 0.1 step size during sampling from the Gaussian IAF and our FloWaveNet. For the Gaussian IAF, increasing the temperature showed a trade-off between the higher fidelity vs. the volume of the background white noise. Similarly, the FloWaveNet had the trade-off when increasing the temperature between the higher fidelity vs. the presence of the periodic artifact. Finding the appropriate temperature had a noticeable effect on sampling the high quality audio with a minimal white noise or artifact.

4.3. Analysis of ClariNet Loss Terms

The Parallel WaveNet (Oord et al., 2017) and the ClariNet (Ping et al., 2018) are parallel waveform synthesis models based on IAF as discussed earlier. They use the KL divergence in tandem with additional auxiliary loss terms to maximize the quality of the sampled audio. Here we present an empirical analysis of the role of each term. We decomposed the two losses of the ClariNet and trained separate Gaussian IAF models trained with only one of the losses: the KL divergence and the spectrogram frame loss.

The model trained with the KL divergence showed a mode collapsing problem, where the student network converged to a certain mode of the distribution of the teacher network. In effect, the generated audio had a very low volume and a peak amplitude at approximately 10% of the maximum. The samples were hardly audible even if we manaully amplified the volume by 8x.

On the other hand, the model trained with the spectrogram frame loss showed relatively fast estimation of an acoustic content of the original audio earlier in training (10K iterations). However, the generated samples showed a high amount of noise, and the noise did not diminish without the KL divergence loss along the remaining training iterations. Thus, the Gaussian IAF training requires both complementary loss terms and is harder to converge with only one of them.

4.4. Context Block and Long-term Dependency

To understand the role of context blocks in modeling the long-term dependency of audio, we performed a comparative study by reducing the number of context blocks of the FloWaveNet from 8 to 6 with the same training procedure. We found that this smaller network brought consistently higher amount of the artifact we described earlier. The artifact is correspondent to the results from Glow (Kingma & Dhariwal, 2018) in that if using less blocks, the network learn to generate a local high-frequency data (such as eyes or lips in facial images) but cannot model a global context (face shape, hair style, etc.). The higher amount of the periodic audio artifact suggests that the smaller model is harder to learn the long-term dependency of the speech audio.

4.5. Causality of WaveNet Dilated Convolutions

The original WaveNet achieved the autoregressive sequence modeling by introducing causal dilated convolutions. However, for our FloWaveNet, the causality is no longer a requirement because the flow-based transformation is inherently parallel in either directions. We compared both approaches, and the non-causal version of the WaveNet structure exhibits better sound quality.

5. Conclusion

In this paper we proposed FloWaveNet, a flow-based generative model that can achieve a real-time parallel audio synthesis that are comparable in fidelity to the two-stage approaches. Thanks to the simplified single loss function and single-stage training, the model can mitigate the needs for highly-tuned auxiliary loss terms while maintaining the stability of training which are useful for practical applications.

References

Arik, Sercan, Diamos, Gregory, Gibiansky, Andrew, Miller, John, Peng, Kainan, Ping, Wei, Raiman, Jonathan, and Zhou, Yanqi. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*, 2017a.

Arik, Sercan O, Chrzanowski, Mike, Coates, Adam, Diamos, Gregory, Gibiansky, Andrew, Kang, Yongguo, Li, Xian, Miller, John, Ng, Andrew, Raiman, Jonathan, et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017b.

Dinh, Laurent, Sohl-Dickstein, Jascha, and Bengio, Samy. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

Jia, Ye, Zhang, Yu, Weiss, Ron J, Wang, Quan, Shen, Jonathan, Ren, Fei, Chen, Zhifeng, Nguyen, Patrick, Pang, Ruoming, Moreno, Ignacio Lopez, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*, 2018.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, Diederik P and Dhariwal, Prafulla. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Kingma, Diederik P, Salimans, Tim, Jozefowicz, Rafal, Chen, Xi, Sutskever, Ilya, and Welling, Max. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Oliva, Junier B, Dubey, Avinava, Póczos, Barnabás, Schneider, Jeff, and Xing, Eric P. Transformation autoregressive networks. *arXiv preprint arXiv:1801.09819*, 2018.

Oord, Aaron van den, Li, Yazhe, Babuschkin, Igor, Simonyan, Karen, Vinyals, Oriol, Kavukcuoglu, Koray, Driessche, George van den, Lockhart, Edward, Cobo, Luis C, Stimberg, Florian, et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*, 2017.

Papamakarios, George, Murray, Iain, and Pavlakou, Theo. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.

Ping, Wei, Peng, Kainan, Gibiansky, Andrew, Arik, Sercan O, Kannan, Ajay, Narang, Sharan, Raiman, Jonathan, and Miller, John. Deep voice 3: 2000-speaker neural text-to-speech. *arXiv preprint arXiv:1710.07654*, 2017.

Ping, Wei, Peng, Kainan, and Chen, Jitong. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*, 2018.

Rezende, Danilo Jimenez and Mohamed, Shakir. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerrv-Ryan, Rj, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Van Den Oord, Aäron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew W, and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio. In *SSW*, pp. 125, 2016.