WG-WaveNet: On-Demand Speech Synthesis on CPU

Dr. V. Dhanakoti, Dr. B.Muthusenthil, Dr.L.Karthikeyan
Department of CSE
SRM Valliammai Engineering College
Chennai, India

Abstract**:** This paper proposes WG-WaveNet, a generative model for compact, accurate, and quick, waveform generation model. It comprises a lightweight flow-based model with an additional post-filter. The proposed process involved joint training of the two components by maximising the conformity to the training data and refining the loss functions tuned to their frequency. The proposed model demands significantly fewer computations against various other generative models on both training and testing, as it’s designed as a heavily compressed flow-based model. Despite the heavy compression of the model, the quality of the generated waveform is maintained by the post-filter. On an NVIDIA 1080 Ti GPU, the PyTorch based implementation can allow training with under 8 GB of GPU memory and produces audio waveforms at above 960 kHz. Additionally, it has been demonstrated that the suggested approach can generate 44.1 kHz speech waveforms 20% quicker as opposed to real-time even when synthesised on CPUs.

Keywords: Text-To-Speech, TTS, Raw Waveform Synthesis, Neural Vocoder

1. Introduction

Models based on neural network are constantly demonstrating progressively cutting-edge performances at speech-based operations like and voice conversion text-to-speech [1,2,3,4]. A major set of the above models are generally made up of two parts. The first component performs speech based operations and produces acoustic properties [5] like spectrograms [1,2], F_0 frequencies, etc. The second component, called vocoder, represents the implementation of heuristic algorithms in a generative manner that converts acoustic data into their corresponding audio samples [6,7,8].

WaveNet [5] was initially developed to be a neural network based vocoder in order to generate natural-sounding human-like audio [1]. WaveNet's auto-regressive architecture allows it to generate high-quality audio, but it also results in a considerably high inference time.

This paper aims to design a waveform generating model that is efficient, high-quality, and has minimal footprint. We begin by compressing WaveGlow using the weight-sharing method[10], thereby considerably reduces the vocoder size. In order to mitigate the adverse effects of compression on the speech quality, a post-filter that is based on WaveNet is used. Training phase is performed using frequency domain loss functions. Given that the post-filter just necessitates modification of the initial compressed WaveGlow’s output, a smaller variant of WaveNet is adequate, enabling the resulting model to be quick and compact. The proposed model, WG-WaveNet, has the advantage of being simple in terms of network design as well as loss function. Furthermore, as opposed to other neural vocoding methods, it has a substantially lower computing cost at both training and testing.This work’s contributions are summarised below:

A hybrid vocoder model comprising a heavely compressed WaveGlow and a post-filter based on WaveNet is proposed. The model, referred to as WG-WaveNet, has shown to be effective both in terms of computational cost and performance during training. In the original WaveGlow paper, 8 NVIDIA GV100 GPUs were stated to

https://doi.org/10.17758/DIRPUB12.DIR1122412

be used for Training [9], while WG-WaveNet can be trained in 4 days with an NVIDIA 1080 Ti GPU requiring under 8 GB of GPU memory. The proposed model significantly boosts the generating efficiency. Notably, WG-WaveNet's inference speed is above 960 kHz when with an NVIDIA 1080Ti GPU and 50% quicker than real-time with just a CPU. Subjective evaluation experiments reveal that the model can create speech of comparable fidelity to WaveNet, Squeeze-Wave , WaveGlow, and Parallel WaveGAN .

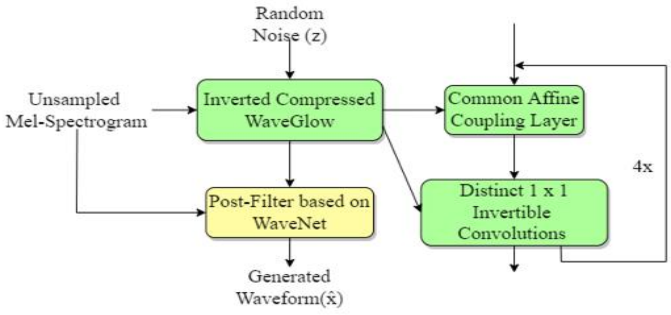
Random
Noise (z)
Unsampled
Mel-Spectrogram
Inverted Compressed WaveGlow
Common Affine Coupling Layer
4x
Post-Filter based on WaveNet
Distinct 1 x 1 Invertible Convolutions
Generated Waveform($\hat{x}$)

Fig. 1. WG-WaveNet Architecture

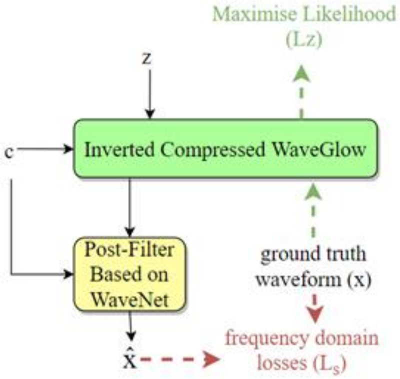

Fig. 2. WG-WaveNet Training

The paper also investigates the accuracy of 44.1 kHz high-fidelity audio waveforms synthesised by neural vocoders. The effects of varying sampling rates and varying parameters used for the underlying Short-Time Fourier Transform (STFT) on the performance of the audio are studied. The proposed method system not only enables synthesising audio samples at 44.1 kHz 20% faster than real-time with a single CPU, but also achieves a MOS score of 4.01, which outperforms 16 kHz recordings

2. Proposed System

The proposed model, is made up of two sections, as shown in Fig. 1, hereto referred as WG-WaveNet.

2.1. Heavily Compressed WaveGlow

WaveGlow [9] was designed as a reversible network that was trained primarily to model real-world voice data distribution. The model is trained using audio inputs and Mel-spectrogram as condition. It is trained to convert the training dataset's audio sample distribution into a zero-mean spherical Gaussian distribution. The goal of training stage is to maximise the conformity of training data. In the testing stage, an inverted variant of WaveGlow obtains an input of random sampled Gaussian noise and produces a speech waveform that are conditioned on a Mel-spectrograms.

The model comprises multiple transformations that are used for incremental mapping speech data into Gaussian space. Every transformation comprises one affine coupling layer along with a 1x1 invertible convolution layer. WaveGlow's affine coupling layers each use a module that is similar to WaveNet. The resulting model is massive and is very difficult to train.

To minimise the parameter count and thereby further compact the model, cross-layer parameter sharing of parameters is employed. The cross-layer sharing of parameters has been established to be beneficial in the Natural Language Processing task pre-training as well as source separation. As portrayed in Fig. 1, a common affine coupling layer is shared by the Transformations in the compressed WaveGlow, achieved by removing the early-output mechanism and keeping the output shape constant across layers. This method prevents the model expanding in size as the network gets deeper.

The training procedure is replicated as described in [9], as indicated by the dotted green flow in Fig. 2. The loss function, represented by L_z, represents the negative log of the likelihood for the given training data. The compression method minimises the WaveGlow parameter count while significantly reducing the GPU memory needed. The following section proposes a post-filter to further increase convergence speed and boost the compressed WaveGlow's performance.

2.2. Post-Filter based on WaveNet

The input for the proposed inverted Compressed WaveGlow is random noise sample z, obtained from a Gaussian distribution. The obtained output is next fed into the post-filter which is based on WaveNet, which parallely generates x̂, conditioned over upsampled Mel-spectrograms. The post-filter is then further trained by minimising the network's loss function represented by $L_s(x, \hat{x})$, where x represents the real world samples and x̂ represents the post-filter's output. To minimise $L_s(x, \hat{x})^2$, both the post-filter as well as the inverted compressed WaveGlow are jointly trained. As WaveNet synthesises audio samples based on the inverted compressed WaveGlow output, it's also feasible to significantly reduce its parameters.

The model employs loss functions on several frequency domains for L_s. Spectral losses were found to be beneficial for training waveform generating models. The multi-resolution STFT auxiliary loss is modified as below.

$$L_s(x,\hat{x}) = \frac{1}{\mathrm{M}}\sum_{i=1}^{M}(L^i_{sc}(x,\hat{x}) + L^i_{mag}(x,\hat{x}) + L^i_{mel}(x,\hat{x})) \quad (1)$$

here, M represents the various parameter sets' count for STFT; L_{mag} and L_{sc} are the log of the STFT magnitude loss and the spectral convergence loss from:

$$L_{sc}(x,\hat{x}) = \frac{\||STFT(x)| - |STFT(\hat{x})|\|_F}{\||STFT(x)|\|_F} \quad (2)$$

$$L_{mag}(x,\hat{x}) = \frac{1}{\mathrm{N}_{maa}}\|\log|STFT(x)| - \log|STFT(\hat{x})|\| \quad (3)$$

here, $\|\cdot\|_F$ represents Frobenius norm, $\|.\|$ represents L1 norm, |STFT(.)| represents the STFT magnitude, and N_{mag} represents the magnitude's element count. In order to make L_s more compatible to natural human perception, an STFT-magnitude loss on the Mel-scale is utilized:

here, the symbol |MEL(.)| represents the Mel-scaled magnitude of the Short Time Fourier Transform while Nmel represents the magnitude's element count. The Mel band count varies based on the varying sets of STFT parameters.

$$L_{\mathrm{mel}}(x,\hat{x}) = \frac{1}{\mathrm{N}_{mel}}\|\log|\mathrm{MEL}(x)| - \log|\mathrm{MEL}(\hat{x})|\| \quad (4)$$

$$L_{\mathrm{total}} = \lambda L_z + L_s \quad (5)$$

The model's second part, post-filter is jointly trained alongside the aforementioned inverted compressed variant of WaveGlow, depicted using red in Fig . 2. The training process for WG-WaveNet uses a linear combination of L_z and L_s for the loss function. The loss terms are balanced by a scalar coefficient, λ. The loss Ls is periodically calculated at every n iterations. The whole WG-WaveNet model is reduced to one thirty-fifth the size of the original WaveGlow.

3. Experiments

3.1. Datasets

The experiments utilised two datasets, with the LJ Speech Dataset being one of them. This LJ dataset contains 13100 clean audio samples (approximately adding up to 24 hours) in English spoken by a female control. The samples were recorded at 22050 Hz. The other dataset utilized was a corpus of 9,004 Mandarin utterances (summing up to around 6.8 hours) by a female control subject, recorded at a sampling rate of 44100 Hz. From each dataset, 100 utterances were chosen for evaluation. A Mel-spectrogram of 80 bands was used in the audio synthesising phase. WG-WaveNet's STFT has its window size as 800, hop size as 200, and a Fast Fourier Transform size of 2048.

3.2. Model Details

The post-filter is built with seven layers of 64-channel dilated convolution blocks. The original model of WaveGlow employs 12 transformations. The compressed WaveGlow is composed of only four transformations owing to the post-filter. The shared affine coupling layer's WaveNet-like module consists of seven layers of 128 channels. Both the compressed WaveGlow as well as the post-filter contain a reduced count of channels and layers than the initial WaveGlow and WaveNet, thereby allowing the WG-WaveNet to be significantly lighter. WaveGlow uses 87.9 million parameters while WaveNet uses 24.7 million.. On the other hand, WG-WaveNet contains just 2.5 M parameters, which is approximately 3% and 10% of the parameters used by WaveGlow and WaveNet, respectively.

TABLE I. LOSS FUNCTION L_S PARAMETERS

Parameters	Values
Window Size	100,200,400,800,1600
Hop Size	25,50,100,200,400
FTT Size	256,512,1024,2048,4096
Mel Band Count	40,80,160,320,640

The proposed technique was compared against four base models: WaveGlow, Parallel WaveGAN, SqueezeWave, and WaveNet. The proposed model's training consisted of 1 million steps with a batch size of eight while employing the Adam optimizer. The rate of learning was 4e-4 and was halved every 200 thousand steps. On the basis of preliminary studies, the set parameters were λ=1 and n=3. Table I contains the parameters used to calculate L_s in Section 3.2.

3.3. Speed and Computational Costs

The memory usage along with the speed for training and testing were assessed for various models. The training phase for both WG-WaveNet and Parallel WaveGAN was performed in same server running on Nvidia V100, with 16 GB of GPU memory to evaluate the training stage's computational costs. The testing setup consisted of a PC running on Intel i7-6700K processor, with an Nvidia 1080 Ti GPU.

3.4. Audio Quality Comparison

Subjective evaluations were done by performing MOS (Mean Opinion Score) tests (higher is better) and objective evaluation was done using Mel Cepstral Distortion (MCD) (the lower the better). For the MOS test, testers were instructed to assign a quality rating to utterances out of 5. Each speech was chosen at random from testing set and assessed by a minimum of twenty raters. Despite using the official release models to generate utterances, SqueezeWave and WaveGlow both performed poorly. Subjects reported that the synthesised speech contained noise and reverberation effects. The ablation investigation demonstrates that both L_z and L_s are essential for WG-WaveNet training. It was discovered that training with L_s ($\lambda = 0$, $n = 1$) alone resulted in high-quality voiced speech but substantial high-frequency glitches in unvoiced speech. Improvements on the generating efficiency has shown to result in rapid reduction in the MOS. However, WG-WaveNet is both faster and scores a 4.09 MOS, comparable to Parallel WaveGAN's MOS. These results demonstrate that WG-WaveNet may significantly speed up synthesis while maintaining comparable performance. As a result of WG-WaveNet's high testing speed demonstrated in Sections 3.3 and 3.4, the model is effective in generating high-fidelity audio (sampling rate of 44.1 kHz). The Parallel WaveGAN model and the WG-WaveNet model utilized the same 44.1 kHz speech dataset specified in Section 3.1 for training to evaluate their performance. WG-WaveNet was compared solely against Parallel WaveGAN primarily as only SqueezeWabe and Parallel WaveGAN are capable of synthesising audio at 44.1 kHz, and SqueezeWave's audio fidelity is inferior to Parallel WaveGAN’s fidelity. MOS evaluations were run on both the generated waveform as well as the ground truth data using identical settings as in Section 3.4 but with different sampling rates.

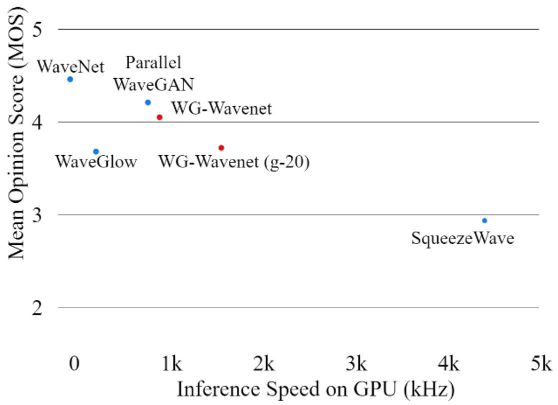

Fig. 2. MOS vs GPU inference speed

TABLE II. The evaluation metrics Mean Opinion Score and Mel Cepstral Distortion against other models and mel-spectrogram generated for ground truth. The Mean Opinion Score results have confidence intervals of 95%.

Model	**Mean Opinion Score**	**Mel Cepstral Distortion**
WaveGlow	3.61±0.149	4.293
Parallel WaveGAN	4.23±0.107	4.025
SqueezeWave	2.95±0.131	3.607
WaveNet	4.47±0.103	4.617
WG-WaveNet		
λ = 0, n = 1	3.66±0.165	2.406
λ = 1, n = 1	3.22±0.161	2.947
λ = 1, n = 3	4.09±0.119	3.782
g-20	3.74±0.125	3.851
Ground Truth	4.62±0.095	-

Table III summarises the findings. "w800" indicates that the window size for Mel-spectrogram extraction has been set at 800. Additionally, the hop size, FTT size, and the Mel band count are identical to those specified in Section 3.1. "w1600" indicates a doubled window size of 1600 pixels, with all other parameters having been doubled as well.

Due to the change in sample rate from 22,050 to 44,100, doubling the Short Time Fourier Transform parameters (w1600) results in the same temporal resolution of retrieved features similar to in Section 3.4, whereas "w800" doubles the temporal resolution. Similarly, the parameters used to calculate L_s in "w800" are identical to those in Table I, but are doubled in "w1600." Initially, the ground truth samples’ sampling rates were shown to have a considerable effect on their perceived quality and consequently their perceptual scores. The raters determined that larger sample rates for ground truths were better. Experiments demonstrated that generating 44.1 kHz speech was more difficult than generating 22 kHz speech if the temporal resolution for the acoustic features was set a constant (w1600).

Mel-spectrograms with a higher temporal resolution (w800) were shown to help improve the performance of WG-WaveNet (w800). In both the "w800" and "w1600" instances, WG-WaveNet outperformed Parallel WaveGAN. The faster WG-WaveNet finally achieved a MOS of 4.01, which is superior to even that of ground truth speech at 16kHz. Table IV summarises the MOS testing results and the GPU inference speed for vocoders

TABLE III. High-quality audio synthesis MOS values with confidence of 95%. Mel-spectrogram for 44.1 kHz ground truths were generated.

Model	Mean Opinion Score
Parallel WaveGAN	
w800	3.04±0.125
w1600	3.12±0.133
WG-WaveNet	
w800 (g-20)	4.02±0.110
w800	3.72±0.131
w1600	3.16±0.147
16 kHz Ground Truth	3.73±0.146
22 kHz Ground Truth	4.14±0.126
44.1 kHz Ground Truth	4.43±0.104

TABLE IV. Evaluation of Mean Opinion Score and GPU testing speed (in kHz) against other models. The tacotron 2 model was used to generate the Mel-spectrograms. The MOS results have confidence intervals of 95%. High-Quality Audio Synthesis

Model of Vocoder with Tacotron 2	Mean Opinion Score	Testing Speed
Griffin Lim	2.12±0.138	-
Parallel WaveGAN	3.73±0.122	842
WaveNet	3.97±0.115	0.13
WG-WaveNet	3.67±0.132	966
Ground Truth	4.35±0.107	-

4. Conclusion

This paper proposed WG-WaveNet, a waveform generating model that is quick, lightweight, and produces high-quality waveforms. By combining a heavily compressed WaveGlow with a post-filter based on WaveNet, the resulting WG-WaveNet utilises significantly fewer computer resources during both training and inference than previous parallel synthesis approaches. Experimental results demonstrate WG-WaveNet's capability to generate high-fidelity audio samples at 44.1 kHz and 22 kHz quicker than real-time without the use of a GPU.

5. References

[1] 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", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Issue. IEEE, pp. 4779– 4783, 2018.

https://doi.org/10.1109/ICASSP.2018.8461368

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6706–6713, 2019

https://doi.org/10.1609/aaai.v33i01.33016706

[3] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations", arXiv preprint arXiv:1804.02812, 2018.

https://doi.org/10.21437/Interspeech.2018-1830

[4] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss", 2019.

[5] A. v. d. 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.

[6] H. Dudley, "Remaking speech", J. Acoust. Soc. Amer., vol. 11,no. 2, pp. 169–177, 1939.

https://doi.org/10.1121/1.1916020

[7] M. R. Schroeder, “Vocoders: Analysis and synthesis of speech”, Proc. IEEE, vol. 54, Issue. 5, pp. 720–734, 1966.

https://doi.org/10.1109/PROC.1966.4841

[8] J. L. Flanagan and R. Golden, “Phase vocoder”, Bell Syst. Tech. J., vol. 45, Issue. 9, pp. 1493–1509, 1966.

https://doi.org/10.1002/j.1538-7305.1966.tb01706.x

[9] R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A flow-based generative network for speech synthesis”, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3617–3621, 2019.

https://doi.org/10.1109/ICASSP.2019.8683143

[10] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, Issue. 10, pp. 1533-1545, Oct. 2014.

https://doi.org/10.1109/TASLP.2014.2339736