VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network

Jinhyeok Yang, Junmo Lee, Youngik Kim, Hoon-Young Cho, Injung Kim

Presented by: Akshay Ram Bhat | s2835509

01/04/2021

Universiteit Leiden
The Netherlands

Index

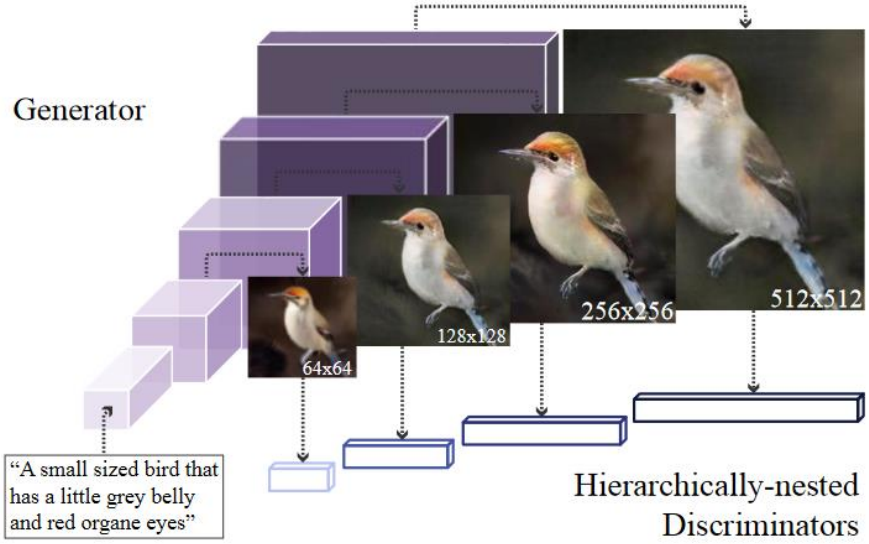
Introduction

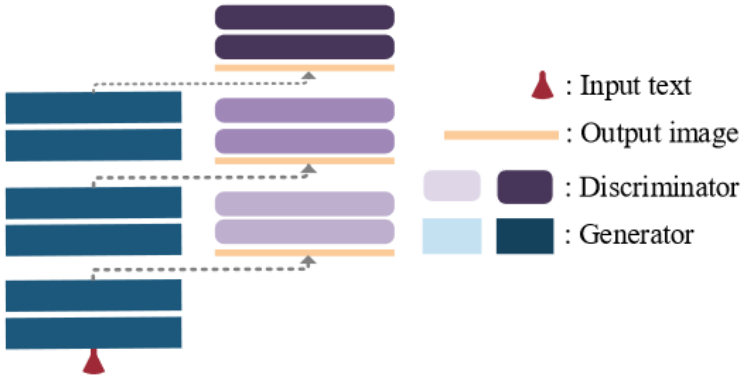
- Neural Vocoder – VocGAN[1]
- Improved performance over MelGAN
- Multi-scale waveform generator
- Hierarchically-nested adversarial network
- Joint conditional and unconditional objective loss (JCU)

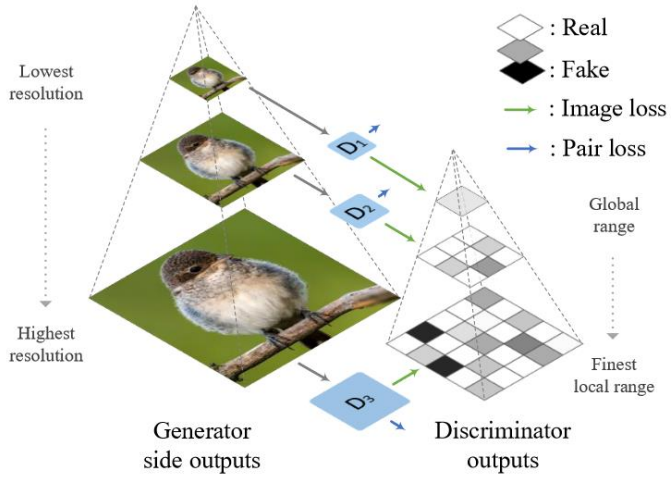
Background Research

Hierarchically Nested GAN[2]

- Used in image generation
- Multiple discriminators at different resolutions
- Backpropagates the real and paired image loss
- Trained end-to-end on a single stream



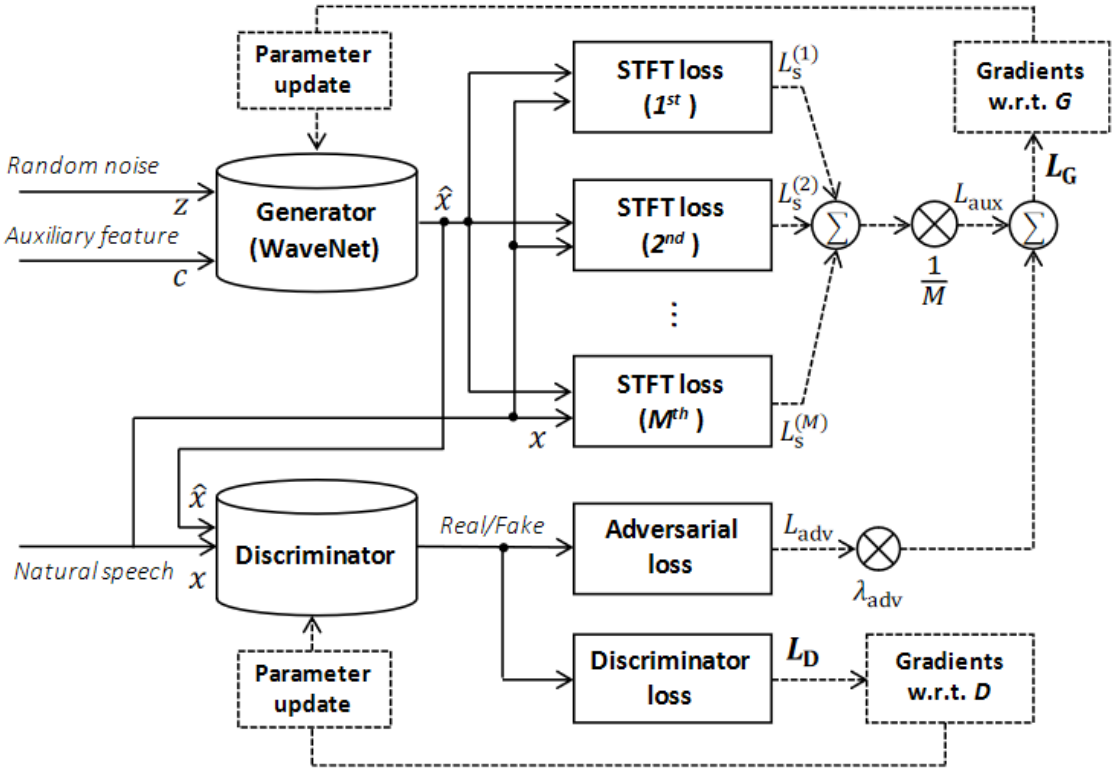




Background Research - Contd

Parallel WaveGAN[3]

- Multi-resolution STFT loss + adversarial loss on the generator
- Discriminator loss on the discriminator



Cons

- Real time performance on CPU is expensive

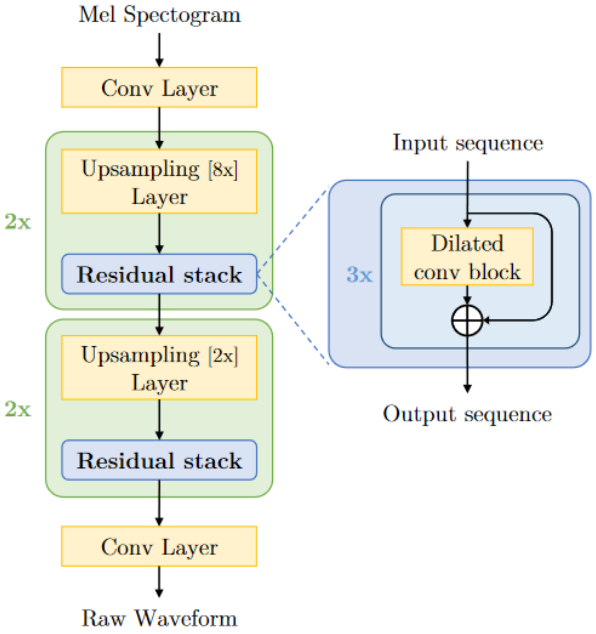
Background Research - Contd

MelGAN[4]

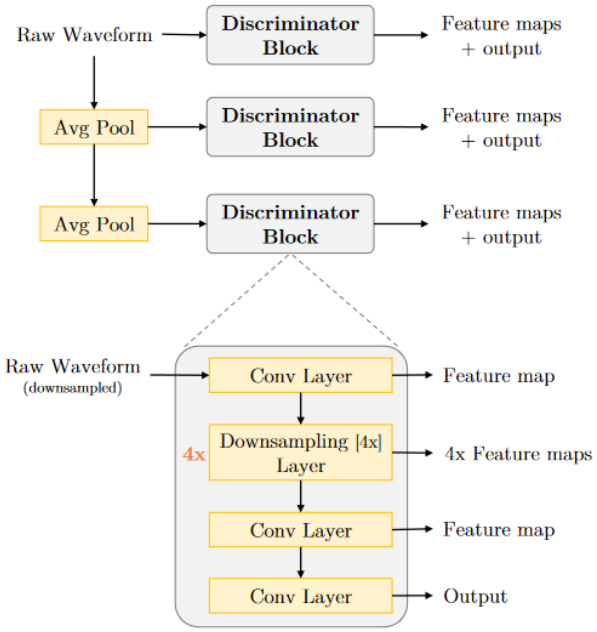
- Carefully tuned, light network
- Window based objective loss function
- Multi-scale discriminator
- Weight normalization

Cons

- Sounds metallic at times
- Network might be too light to learn acoustic features



(a) Generator



(b) Discriminator

Methodology

Extension of MelGAN. Adds the following components:

- Hierarchically nested structure and loss
- JCU loss
- STFT loss

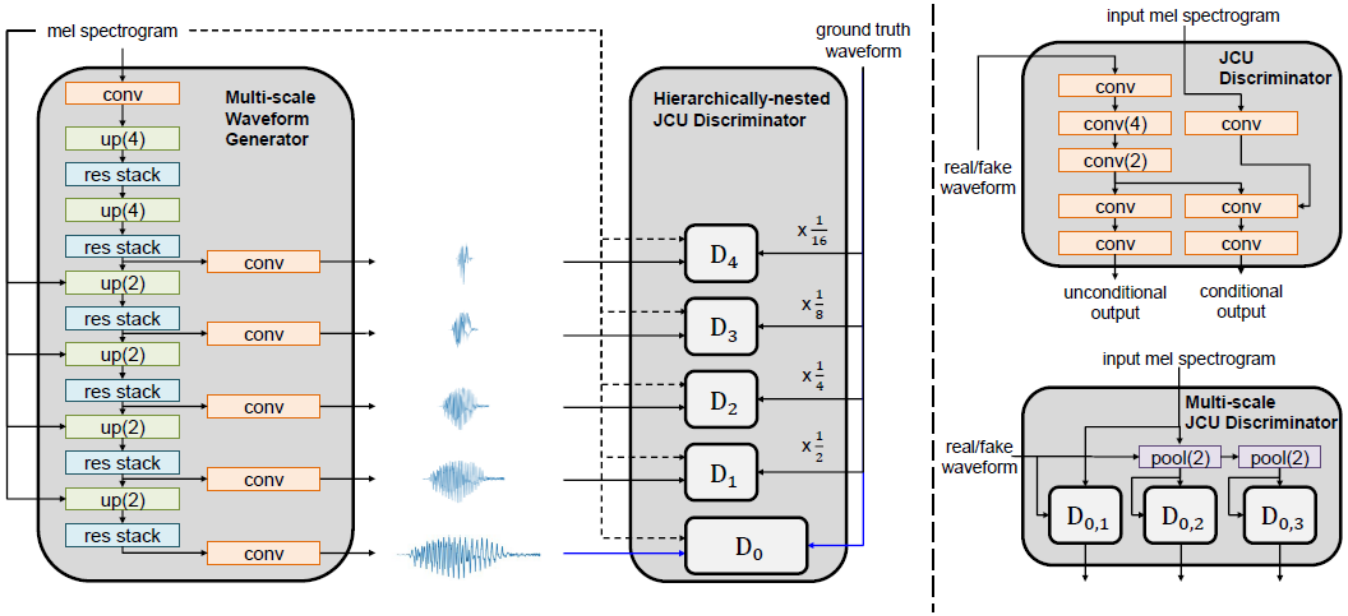


Figure 2: *Model architecture.* $\times\frac{1}{2^k}$ *denotes a down-sampling rate.* $up(u)$ *denotes an up-sampling layer whose rate is* u*.* $conv$ *and* $pool(v)$ *are convolutional layer and down-sampling pooling layer with a stride of* v*, respectively.* $res\ stack$ *denotes a residual stack.*

Methodology - Contd

- The Discriminator and Generator losses are given by

$$L_D(G,D) = \sum_{k=0}^{K} V_k(G,D_k) \quad and \quad L_G(G,D) = \sum_{k=0}^{K} \frac{1}{2}\mathbb{E}_s[(D_k(\hat{x}_k) - 1)^2]$$

- The Joint Conditional Loss is given by

$$L_G^{JCU}(G,D) = \sum_{k=0}^{K} \frac{1}{2}\mathbb{E}_s\left[(D_k(\hat{x}_k) - 1)^2 + (D_k(\hat{x}_k, s) - 1)^2\right]$$

- The feature matching loss is given by

$$L_{FM}(G,D) = \mathbb{E}_{(s,x)}\left[\sum_{k=0}^{K}\sum_{t=1}^{T_k}\frac{1}{N_t}||D_k^{(t)}(x_k) - D_k^{(t)}(\hat{x}_k)||\right]$$

- The total generator loss is given by

$$L_G^{total}(G,D) = L_G^{JCU}(G,D) + \alpha L_{FM}(G,D) + \beta L_{STFT}(G)$$

Experiments and Results

Datasets and Settings

- Korean Single Speaker Speech dataset - 12,853 samples
- LJ Speech dataset - 13,100 samples
- The total lengths are 12 and 24 hours
- Trained for 3000 epochs
- Adam optimizer(learning rate = 0.0001, $\beta 1 = 0.5$ and $\beta 2 = 0.9$) for both generator and discriminator
- For multi-resolution STFT loss, three STFT losses with frame sizes of 512, 1024 and 2048, window sizes of 240, 600 and 1200 and frame shifts of 50, 120 and 240, respectively were applied

Experiments and Results - Contd

Table 1: *The result of ablation study. MCD(dB) and F_0 RMSE(Hz): the lower, the better. PESQ: the higher, the better.*

Method	KSS			LJ		
	MCD	F_0 RMSE	PESQ	MCD	F_0 RMSE	PESQ
Baseline (MelGAN)	4.478	38.80	2.51	4.614	50.04	2.74
+ Hierarchically-nested structure and loss	3.986	37.84	2.66	3.827	49.39	2.91
+ JCU loss	3.441	35.39	2.93	3.551	45.87	3.06
+ Hierarchically-nested structure and loss + JCU loss	3.229	**32.36**	3.37	**3.144**	44.19	3.32
+ Hierarchically-nested structure and loss + STFT loss	3.438	34.99	3.03	3.707	48.68	3.03
VocGAN	**2.974**	32.85	**3.48**	3.199	**43.10**	**3.44**
Ground Truth	0.0	0.0	4.5	0.0	0.0	4.5

Experiments and Results - Contd

Table 2: *MOS with 95% confidence intervals. The unit of inference speed is real-time factor that measures how many times faster than real-time.*

Method	MOS	Inference Speed	
		GPU	CPU
MelGAN	3.898±0.091	**574.7x**	**3.73x**
Parallel WaveGAN	4.098±0.085	125.0x	0.47x
VocGAN (proposed)	**4.202±0.081**	416.7x	3.24x
Ground Truth	4.721±0.052	-	-

Conclusion

- Improved stability and efficiency of learning over MelGAN
- Synthesis speed 416.6x times and 3.24x times better on GPU and CPU

Future Work

- Increase speed of inferencing on CPU

References

1. Yang, J., Lee, J., Kim, Y., Cho, H., & Kim, I. (2020). VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network. *arXiv preprint arXiv:2007.15256*.
2. Zhang, Z., Xie, Y., & Yang, L. (2018). Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6199-6208).
3. R. Yamamoto, E. Song and J. -M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6199-6203, doi: 10.1109/ICASSP40776.2020.9053795.
4. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., ... & Courville, A. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*.