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

Discover the world at Leiden University
Index

1. Introduction

2. Background Research
   a. Hierarchically nested GANs
   b. Parallel WaveGAN
   c. MelGAN

3. Methodology
   a. Working
   b. Loss Formulas

4. Experiment and Results
   a. Datasets and Settings
   b. Ablation Study
   c. MOS Results

5. Conclusion
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

“A small sized bird that has a little grey belly and red organe eyes”
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
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. $\text{up}(u)$ denotes an up-sampling layer whose rate is $u$. $\text{conv}$ and $\text{pool}(v)$ are convolutional layer and down-sampling pooling layer with a stride of $v$, respectively. $\text{res stack}$ denotes a residual stack.
Methodology - Contd

• The Discriminator and Generator losses are given by

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

• The Joint Conditional Loss is given by

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

• The feature matching loss is given by

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

• The total generator loss is given by

$$L_G^{total}(G, D) = L_G^{JC}(G, D) + \alpha L_{FM}(G, D) + \beta L_{STFT}(G)$$
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, β1 = 0.5 and β2 = 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>MOS</th>
<th>Inference Speed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GPU</td>
<td>CPU</td>
</tr>
<tr>
<td>MelGAN</td>
<td>3.898±0.091</td>
<td><strong>574.7x</strong></td>
<td><strong>3.73x</strong></td>
</tr>
<tr>
<td>Parallel WaveGAN</td>
<td>4.098±0.085</td>
<td>125.0x</td>
<td>0.47x</td>
</tr>
<tr>
<td>VocGAN (proposed)</td>
<td><strong>4.202±0.081</strong></td>
<td>416.7x</td>
<td><strong>3.24x</strong></td>
</tr>
<tr>
<td>Ground Truth</td>
<td>4.721±0.052</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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


Discover the world at Leiden University