MelGAN: Generative Adversarial Network for Conditional Waveform Synthesis

Kundan Kumar[1,2], **Rithesh Kumar**[1], Thibault de Boissiere[1],

Lucas Gestin[1], Wei Zhen Teoh[1], Jose Sotelo[1,2], Alexandre de Brebisson[1,2],

Yoshua Bengio[2], Aaron Courville[2]

1) descript Lyrebird

2) Mila

The Hard, Slow and Impossible

- Raw audio modelling with low capacity model is **hard** due to the high temporal resolution of the data (at least 16,000 samples per second) and the presence of structure at different timescales with short and long-term dependencies.
- High quality audio waveform generation is **slow** with most existing state-of-the-art deep neural net models, which operate in autoregressive manner.
- It has **yet to be shown possible** to train a raw audio model for speech in a GAN setup with only adversarial loss terms.

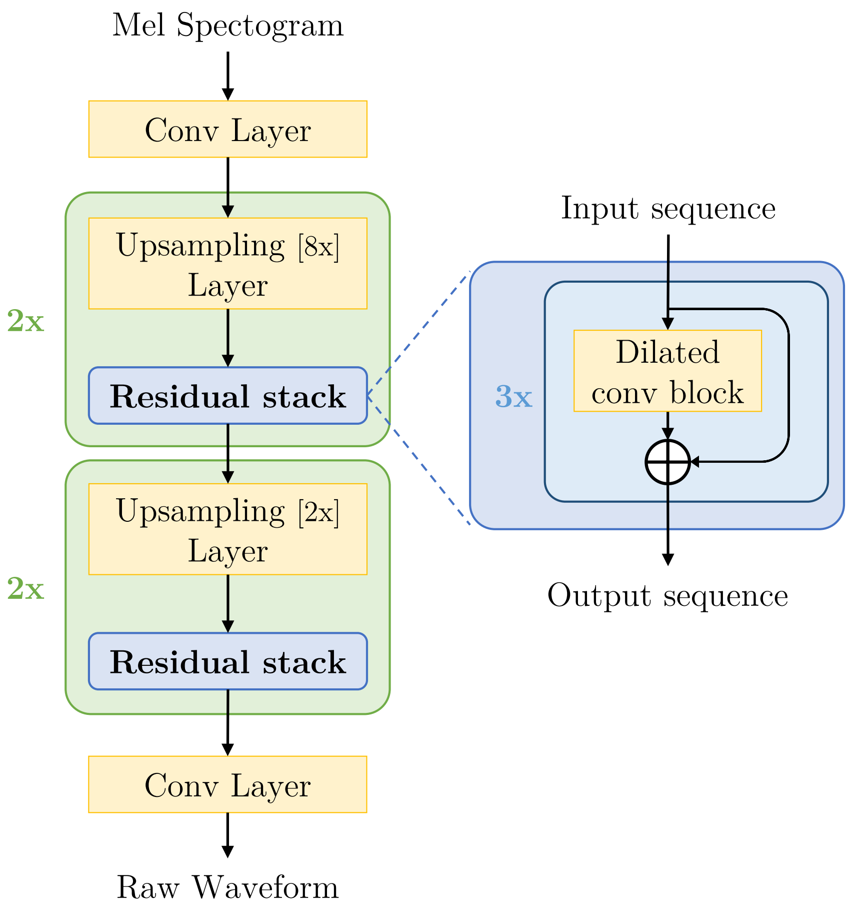
Our Contributions

MelGAN is a conditional waveform synthesis model that solves the hard, slow and impossible in raw audio modelling:

- considerably **smaller** than any existing deep neural net model for audio modelling with comparable output quality.
- non-autoregressive in generation, **10x faster** than the fastest available model to date with comparable output quality.
- trained in a GAN setup without additional perceptual loss term or probability distillation objective.

We show that existing high quality vocoder models can be readily replaced by MelGAN in text-to-speech.

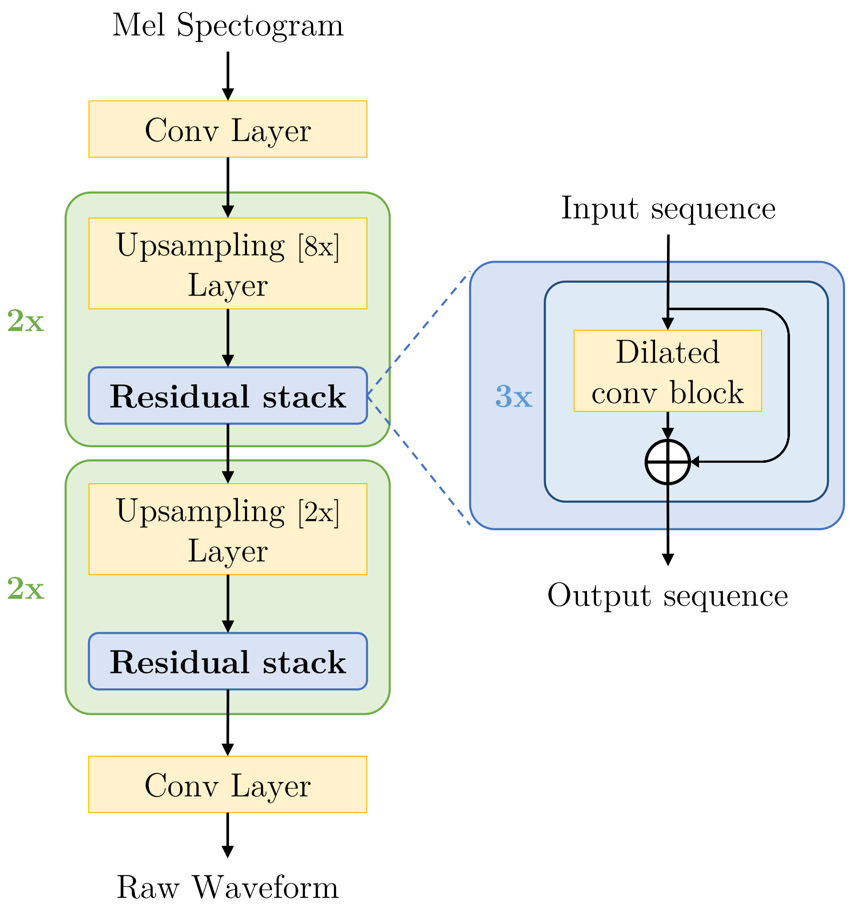
Key Designs - Generator



Architecture

- Stack of transposed convolutional layers to upsample the input sequence.
- Each transposed convolutional layer followed by a stack of residual blocks.

Key Designs - Generator



Induced Receptive Field

- **Residual blocks with dilations** so temporally far output activations of each layer has significant overlapping inputs.
- Receptive field of a stack of dilated convolution layers **increases exponentially** with the number of layers.

Key Designs - Generator

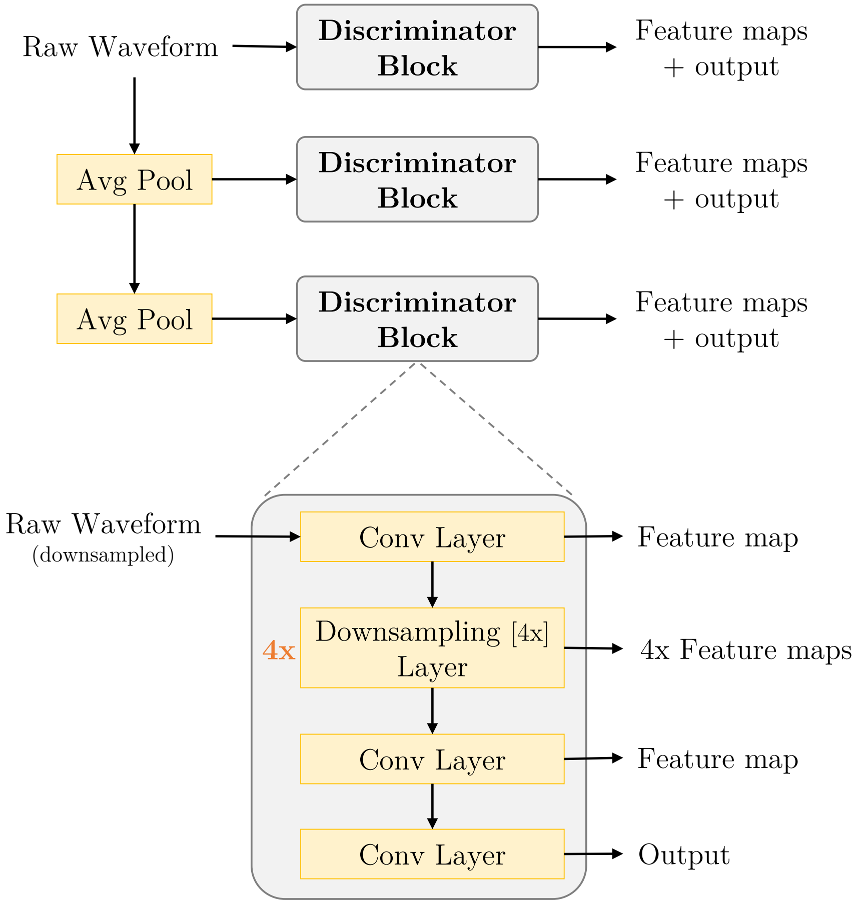
Checkerboard Artifacts

- Kernel-size as a **multiple** of stride
- Dilation grows as a **power** of the kernel-size
- Receptive field of the stack looks like a fully balanced (seeing input uniformly) and symmetric tree

Normalization

- Instance-norm washes away pitch information, making audio sound metallic
- Spectral-norm's strong Lipshitz constraint badly impacts feature matching objective
- **Weight-norm** works best

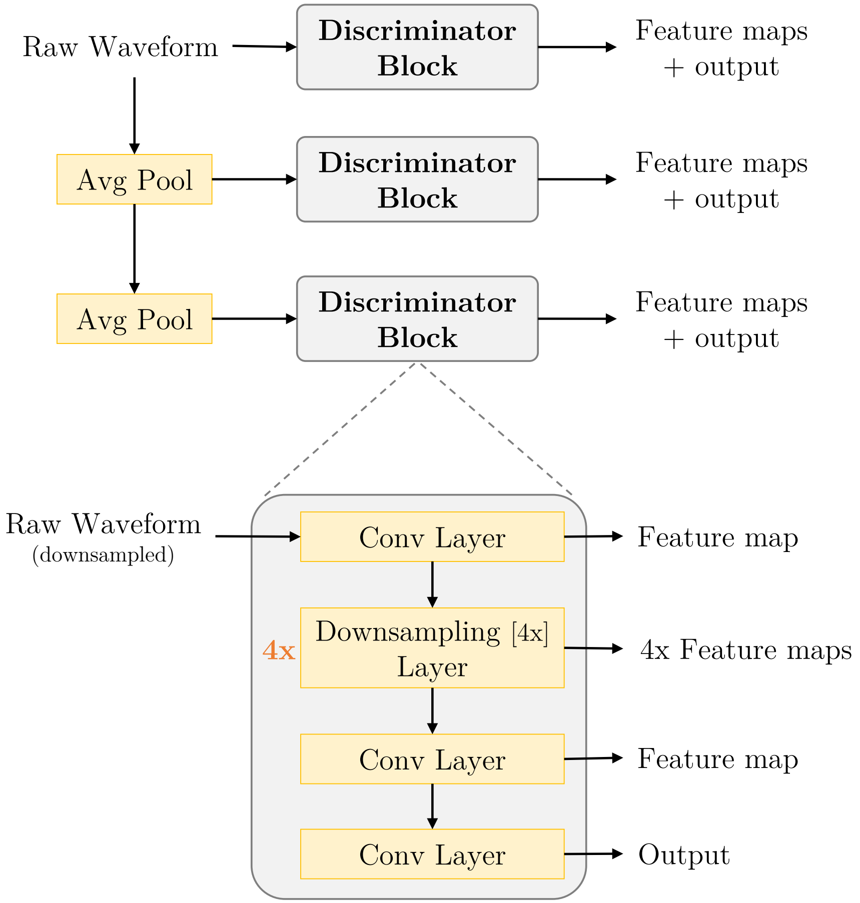
Key Designs - Discriminator



Multiscale Architecture

- 3 discriminators (identical structure) operate on different audio scales -- original scale, 2x and 4x downsampled.
- Each discriminator biased to learn features for **different frequency range** of the audio.

Key Designs - Discriminator



Window-based objective

- Each individual discriminator is a Markovian window-based discriminator (analogues to image patches, Isola et al. (2017))
- Discriminator learns to classify between distributions of **small audio chunks**.
- Overlapping large windows maintain coherence across patches

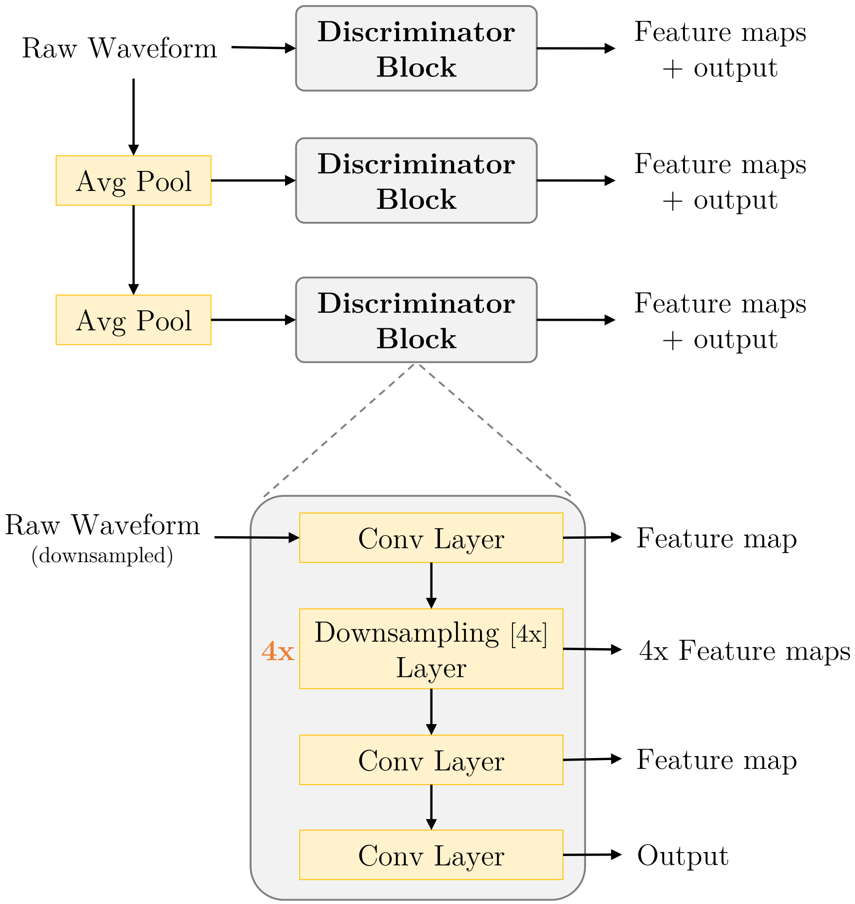
Training

We use the hinge loss formulation (Lim & Ye, 2017; Miyato et al., 2018):

$$\min_{D_k} \mathbb{E}_x \Big[\min(0, 1 - D_k(x)) \Big] + \mathbb{E}_{s,z} \Big[\min(0, 1 + D_k(G(s, z))) \Big], \ \forall k = 1, 2, 3$$

$$\min_{G} \mathbb{E}_{s,z} \left[\sum_{k=1,2,3} -D_k(G(s, z)) \right]$$

Training



We additionally use a feature matching objective (Larsen et al., 2015) to train the generator:

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{x,s\sim p_{\text{data}}}\left[\sum_{i=1}^{T}\frac{1}{N_i}||D_k^{(i)}(x) - D_k^{(i)}(G(s))||_1\right]$$

Overall Generator Objective:

$$\min_G\left(\mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] + \lambda\sum_{k=1}^{3}\mathcal{L}_{\text{FM}}(G, D_k)\right)$$

Smaller, Faster

Model	Number of parameters (in millions)	Speed on CPU (in kHz)	Speed on GPU (in kHz)
Wavenet (Shen et al., 2018)	24.7	0.0627	0.0787
Clarinet (Ping et al., 2018)	10.0	1.96	221
WaveGlow (Prenger et al., 2019)	87.9	1.58	223
MelGAN (ours)	**4.26**	**51.9**	**2500**

Spectrogram Inversion

MelGAN vs existing methods for inverting ground truth mel-spectrograms to raw audio on LJ Speech Dataset (Ito, 2017):

Model	MOS	95% CI
Griffin Lim	1.57	±0.04
WaveGlow	4.09	±0.06
WaveNet	4.03	±0.06
MelGAN	3.60	±0.06
Original	**4.50**	± **0.05**

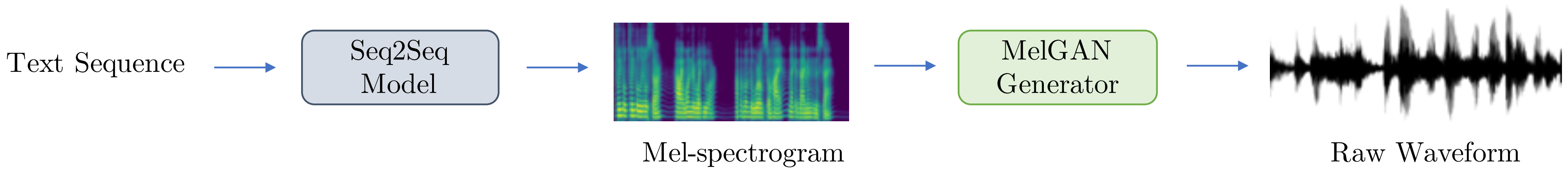
MelGAN trained on an internal 6-speaker dataset generalizes to unseen VCTK dataset (Veaux et al., 2017) speakers:

Model	MOS	95% CI
Griffin Lim	1.72	±0.07
MelGAN	3.49	±0.09
Original	**4.19**	± **0.08**

End-to-end Speech Synthesis

MelGAN vs existing methods as vocoder component of TTS pipeline:

Model	MOS	95% CI
Tacotron2 + WaveGlow	3.49	±0.04
Text2mel + WaveGlow	4.07	±0.04
Text2mel + MelGAN	3.70	±0.04
Text2mel + Griffin-Lim	1.42	±0.05
Original	4.45	± 0.04



Conclusion

- MelGAN is lightweight, very fast at inference time and is capable of producing high quality audio output.
- MelGAN generator can be a high quality plug-and-play replacement to compute-heavy alternatives for audio related tasks, particularly in speech domain.