Generating music in the waveform domain

https://benanne.github.io/2020/03/24/audio-generation.html

sanderdieleman@gmail.com

 @sedielem

Overview

Why audio? Why raw audio?

Generative models

Likelihood-based models of raw audio

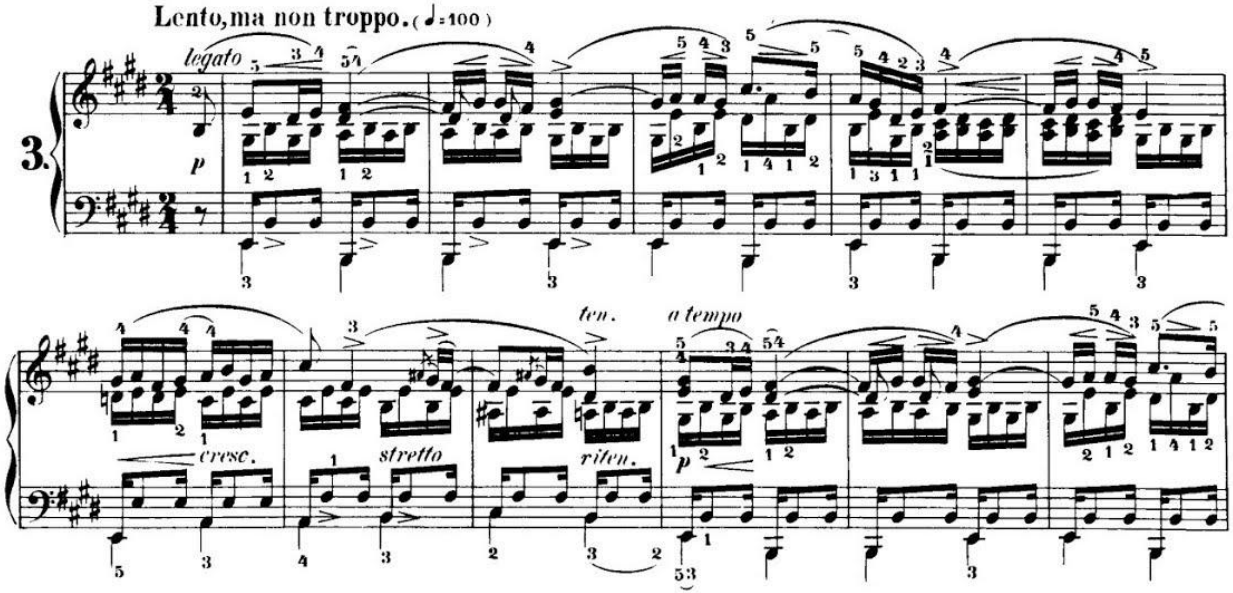
Adversarial models of raw audio

Summary

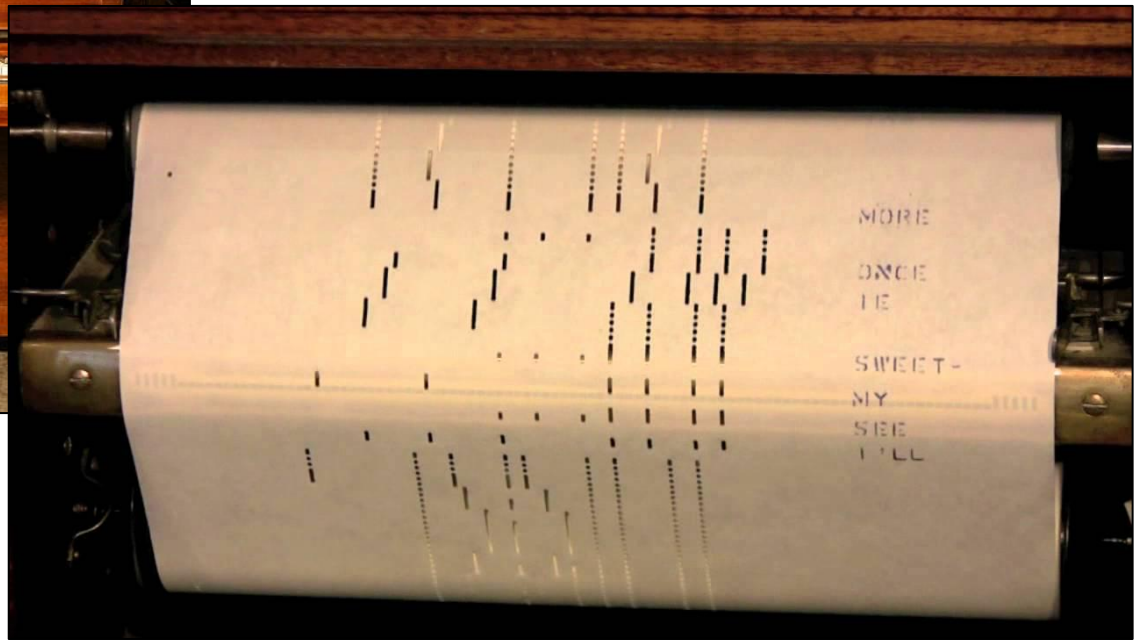
Why audio? Why raw audio?

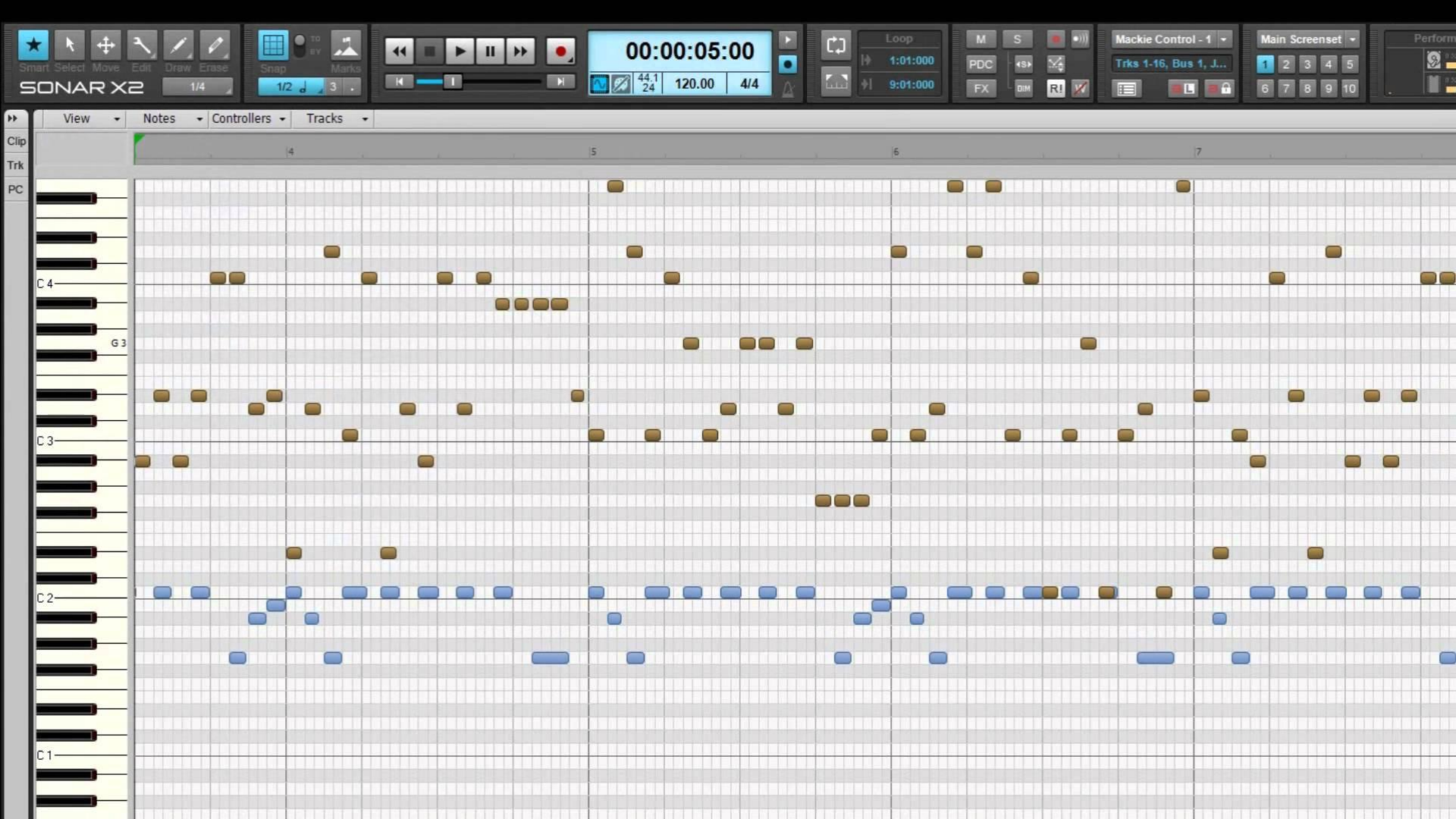
Why audio?

Music generation is typically studied in the symbolic domain




Steinway


MORE
ONCE
TE
SWEET-
MY
SEE


Smart
Select
Move
Edit
Draw
Erase
SONAR X2
1/4
Snap
Marks
1/2
3
00:00:05:00
44.1
24
120.00
4/4
Loop
1:01:000
9:01:000
M
S
PDC
FX
DIM
Mackie Control - 1
Trks 1-16, Bus 1, J...
Main Screenset
1
2
3
4
5
6
7
8
9
10
View
Notes
Controllers
Tracks
Clip
Trk
PC
4
5
6
7
C 4
G 3
C 3
C 2
C 1

Why audio?

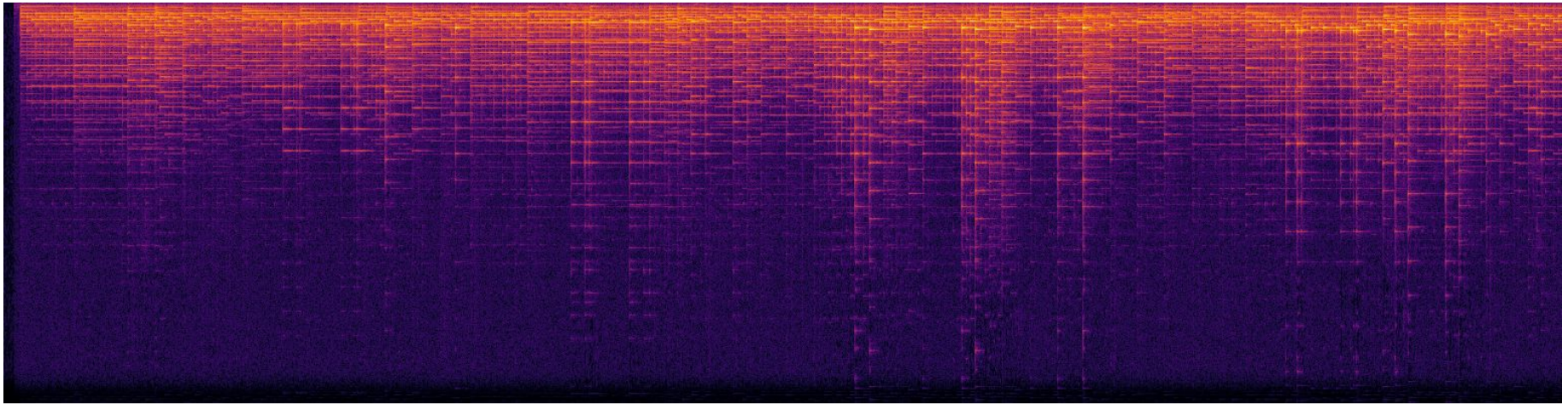
Many instruments have complex action spaces

Rich palette of sounds and timbral variations

Guitar

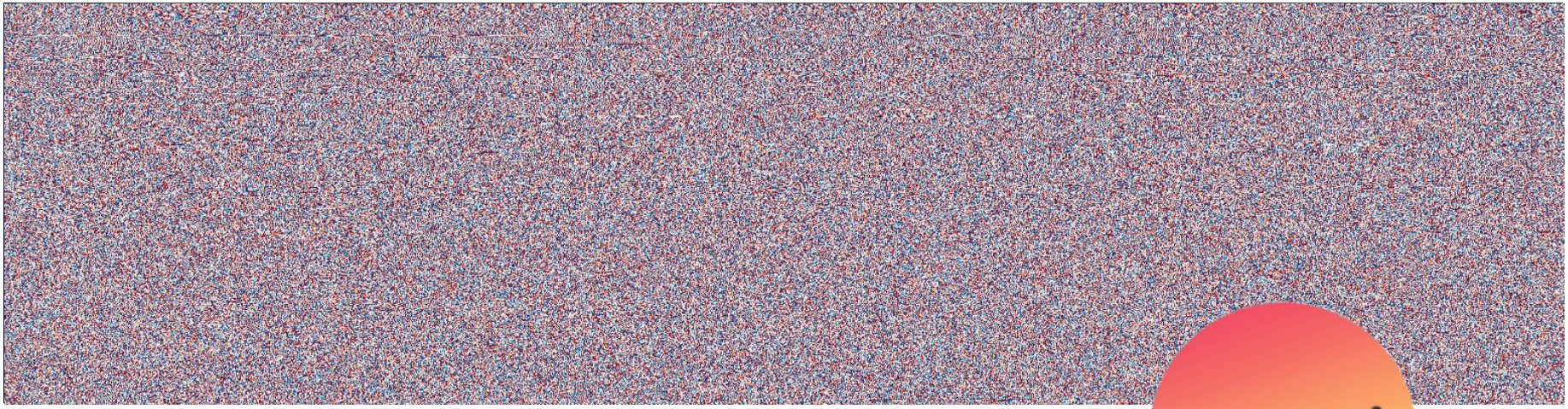
- pick vs. finger
- picking position
- frets
- harmonics
- ...

Why raw audio?



Magnitude spectrogram

Why raw audio?



Phase spectrogram

Why raw audio?

Phase is often unimportant in discriminative settings,
but is very important perceptually!

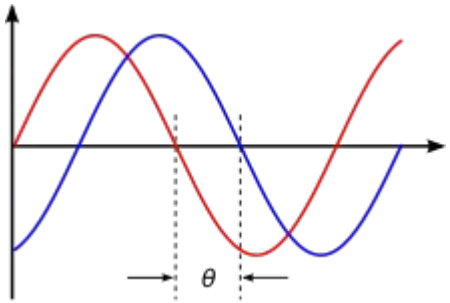
original phase

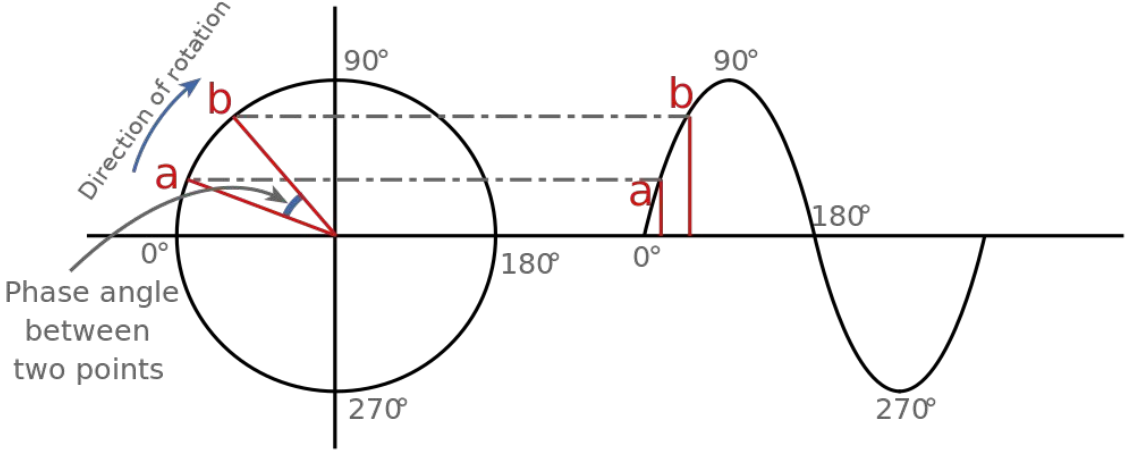
random phase

Why raw audio?

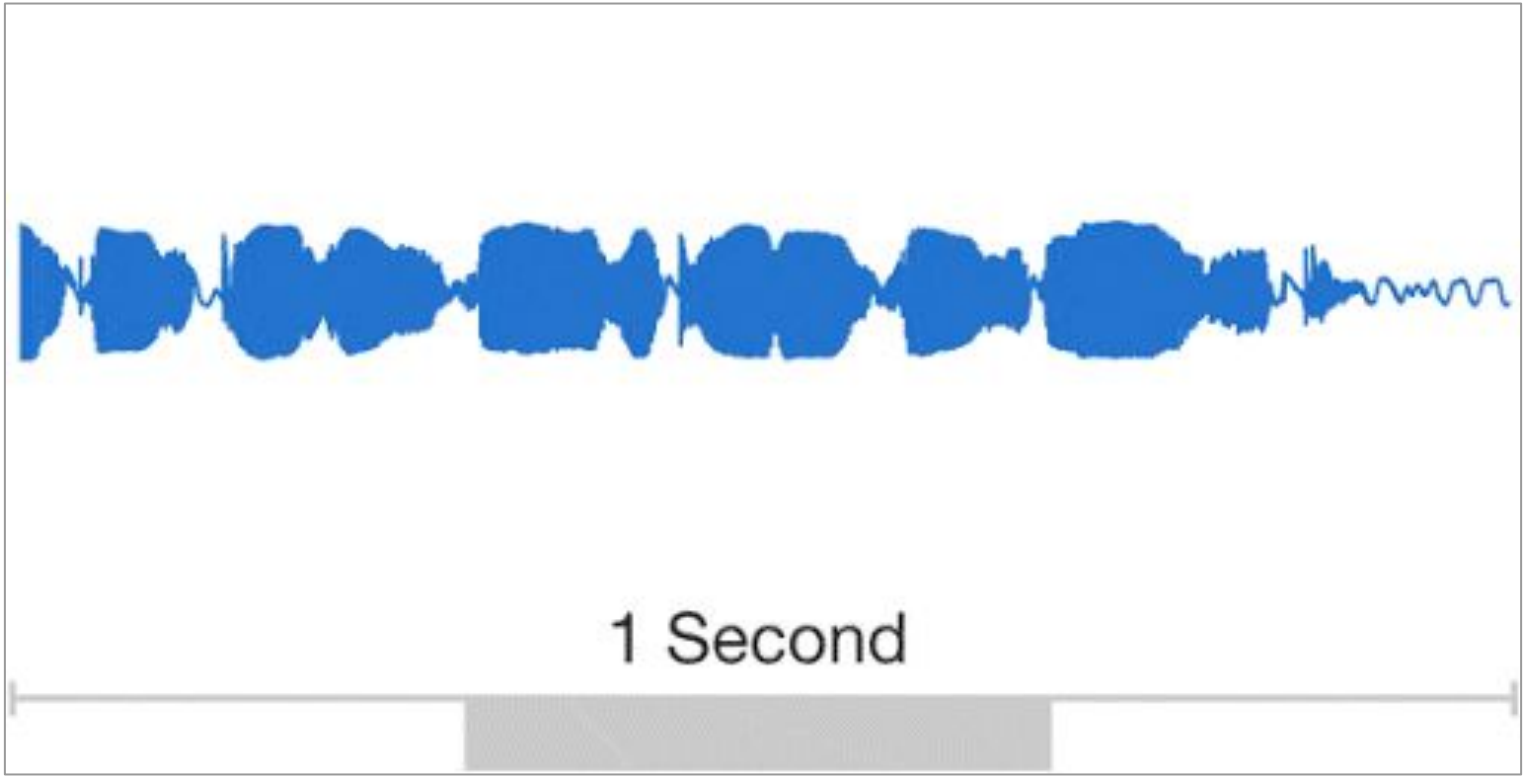
Phase is hard to model:

- it is an angle, so it wraps around
- it becomes random as the magnitude tends to 0
- absolute phase is less meaningful, but relative phase differences matter





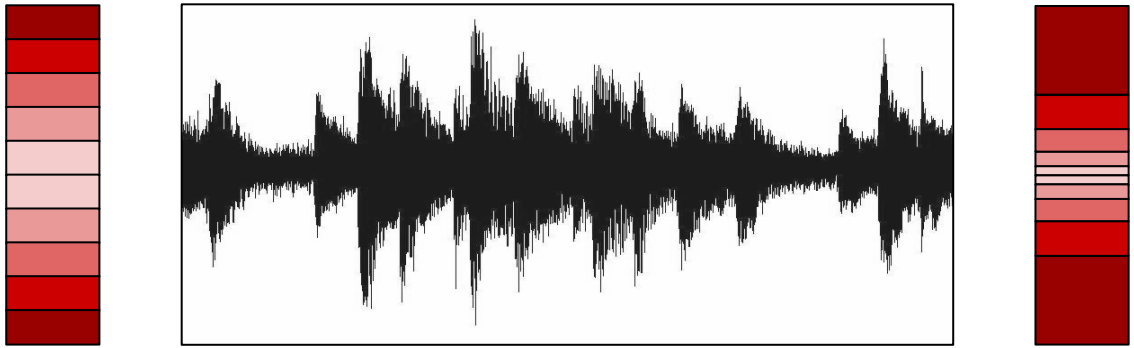
What is “raw audio” anyway?



Discretising audio

- Time
- Amplitude

uniform quantisation



μ-law quantisation

Generative models

Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**

Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**

Explicit: given **x ∈ X**, model can infer **p(x)**

Implicit: model can produce new samples **x ~ p(X)**

Generative models

Brock et al., 2019. “Large Scale GAN Training for High Fidelity Natural Image Synthesis”, ICLR.

Likelihood-based models

Likelihood-based models parameterise **p(X)** directly

Objective function: maximise $\sum_{x \in X} \log p(x)$

Autoregressive models

Autoregressive models factorise **p(X)**
into simpler (scalar) distributions

$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)$

$\mathbf{p(x)} = \prod_i \mathbf{p(x_i|x_{<i})}$ chain rule of probability

We can use the same model $\mathbf{p(x_i|x_{<i})}$ for all i!

Flow-based models

Flow-based models transform **p(X)** to a simple (factorised) distribution with an invertible mapping

$\mathbf{p(x) = p(z) \cdot |det\ J|^{-1}}$ change of variables theorem

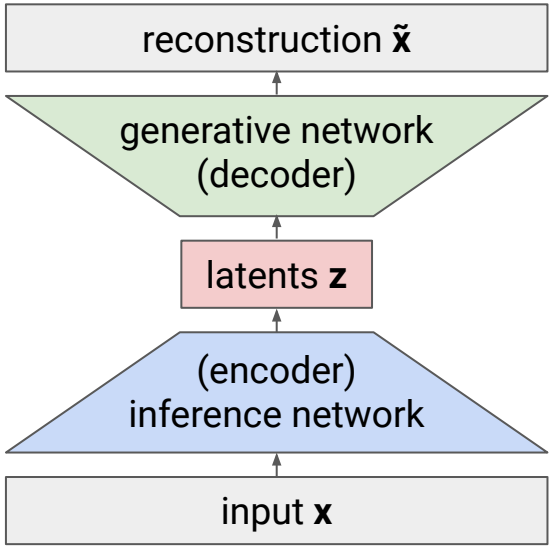
$\mathbf{J = d(g(z))/dz}$ $\mathbf{x = g(z)}$

Important constraints:

g(z) must be invertible **det J** must be tractable

Dinh et al., 2014. “NICE: Non-linear Independent Components Estimation”, arXiv.
Dinh et al., 2016. “Density estimation using Real NVP”, arXiv.

Variational autoencoders (VAEs)



$\mathbf{p_\varphi(x|z)}$

$\mathbf{z \sim \mathcal{N}(0,1)}$

$\mathbf{q_\theta(z|x)}$

VAE maps latents **z** from a simple distribution to **x** with a (non-invertible) generative network.

The inference network approximates the inverse operation.

p(x) cannot be computed exactly, the Evidence Lower BOund (ELBO) is maximised instead.

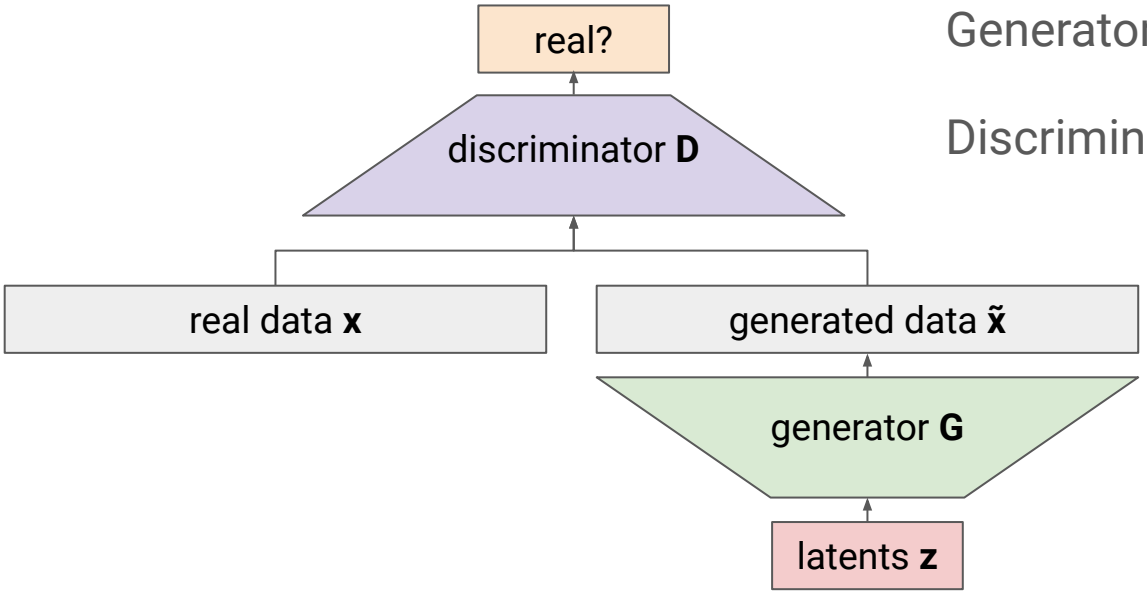
Kingma & Welling, 2014. “Auto-Encoding Variational Bayes”, ICLR 2014.
Rezende et al., 2014. “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”, ICML 2014.

Adversarial models

$$\mathcal{L} = \mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1 - D(G(z)))]$$

Generator minimises $\mathcal{L}$

Discriminator maximises $\mathcal{L}$



Goodfellow et al., 2014. “Generative Adversarial Nets”, NeurIPS.

More exotic flavours

- Implicit quantile networks
- Energy-based models
- Optimal transport (e.g. Wasserstein autoencoders)
- Score-based generative modelling
- Maximum mean discrepancy (energy distance)
- ...

Conditional generative models

Conditioning is “side information” which allows for control over the model output

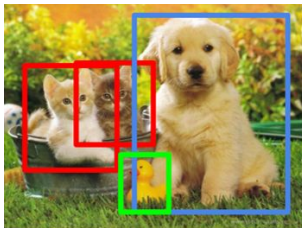
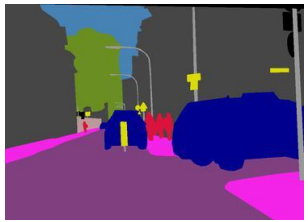
p(x|c) vs. **p(x)**

Conditional generative models

Conditioning is “side information” which allows for control over the model output

p(x|c) vs. **p(x)**

sparsely conditioned → densely conditioned

image models	class labels	bounding boxes	segmentation	grayscale image (colorisation)
	y = “cat”			

Conditional generative models

Conditioning is “side information” which allows for control over the model output

p(x|c) vs. **p(x)**

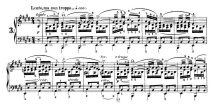
sparsely conditioned → densely conditioned

music audio models

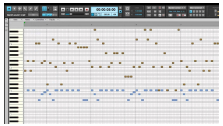
composer
instrument(s)
tempo
timbre
...

note density
musical form
…

score

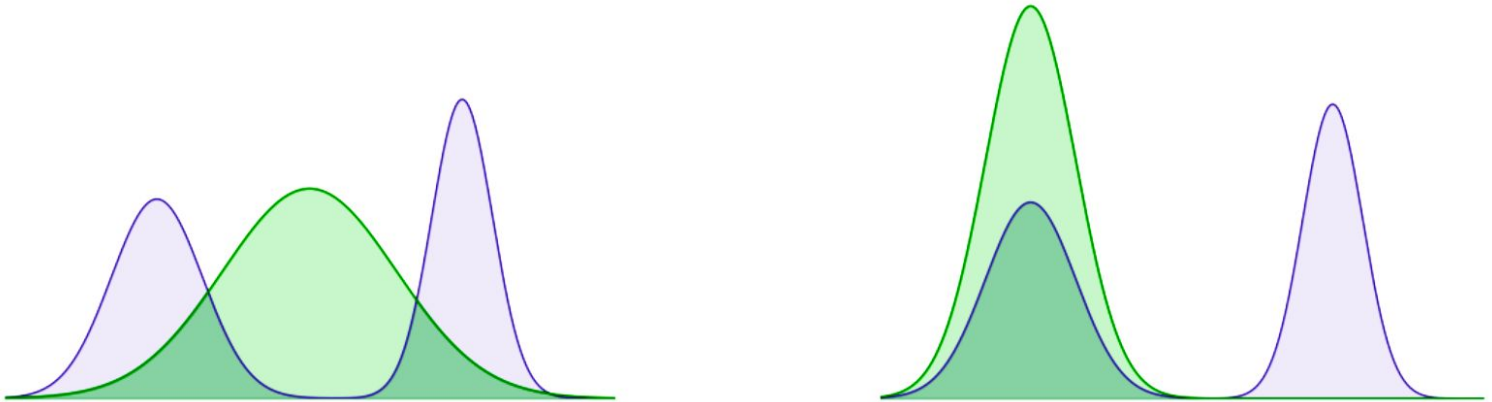


MIDI



other audio signals

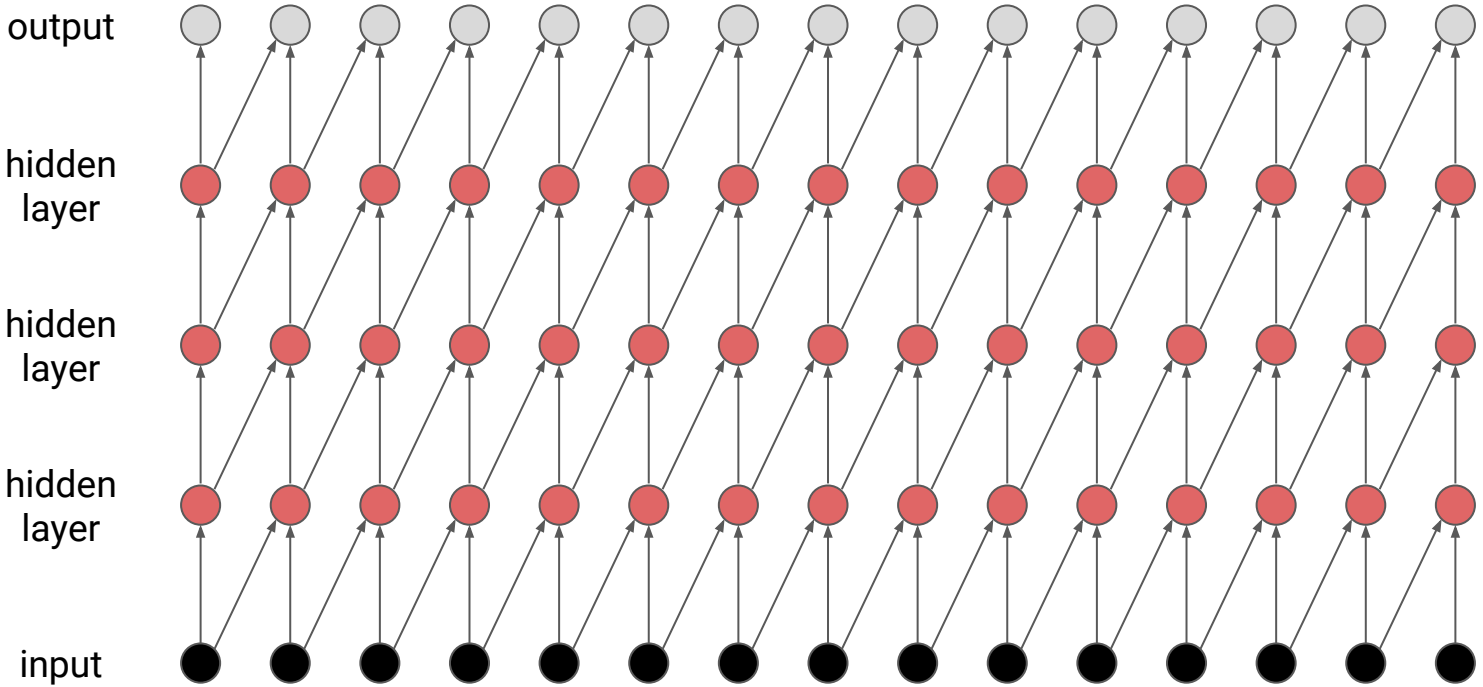
Mode-covering vs. mode-seeking behaviour



- Likelihood-based models are mode-covering
- Adversarial models are (typically) mode-seeking
- In more densely conditioned settings, we tend to care less about covering all the modes

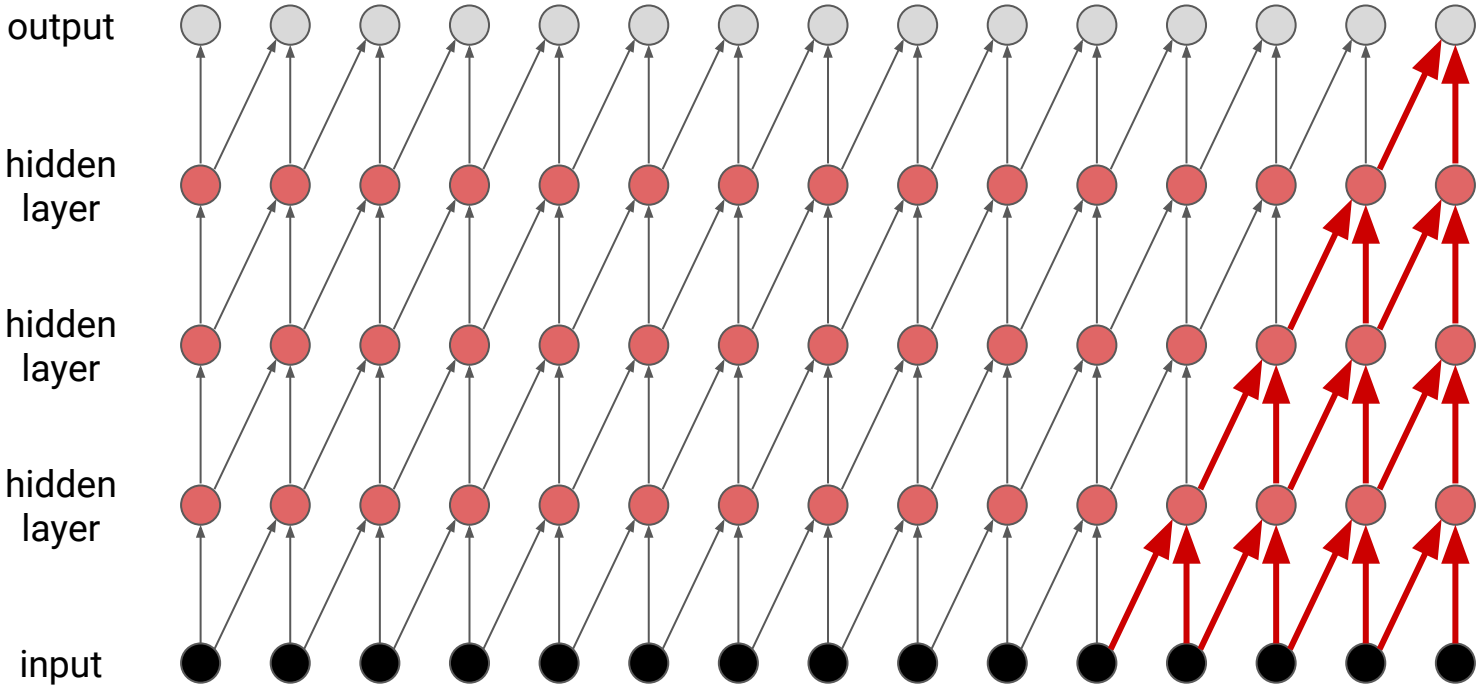
Likelihood-based models

WaveNet



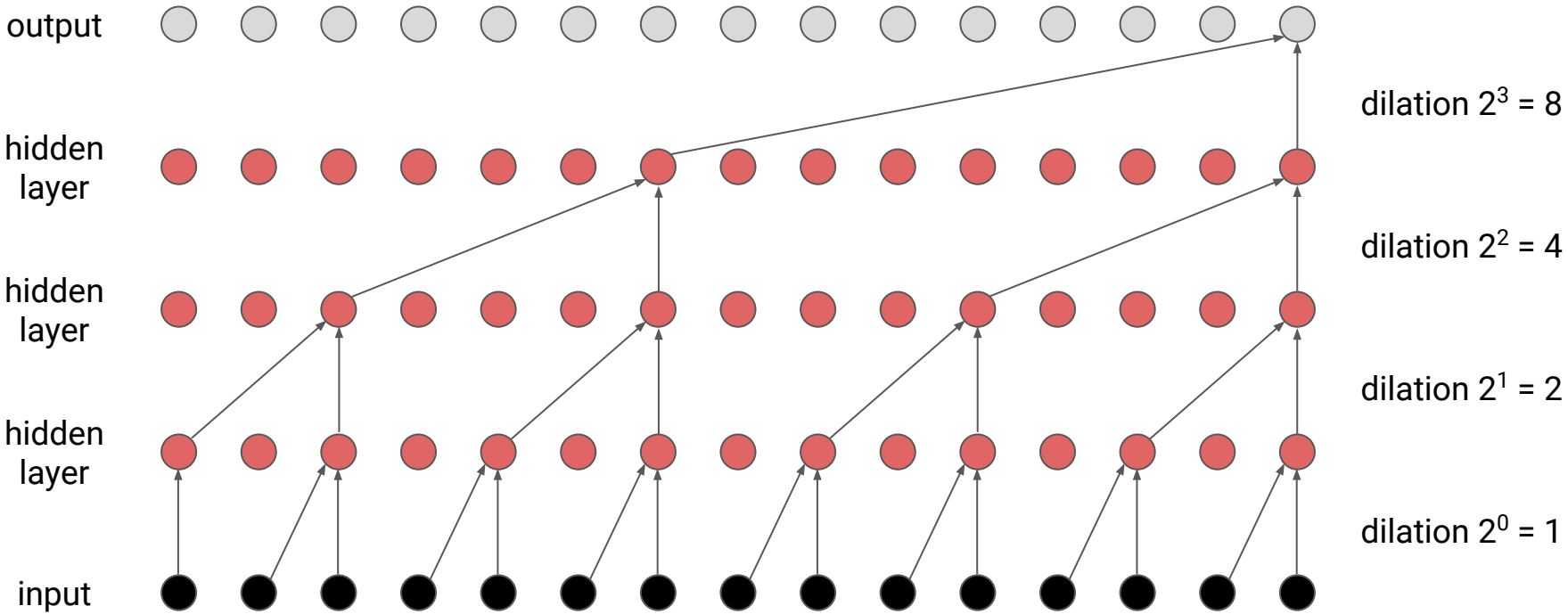
van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

WaveNet



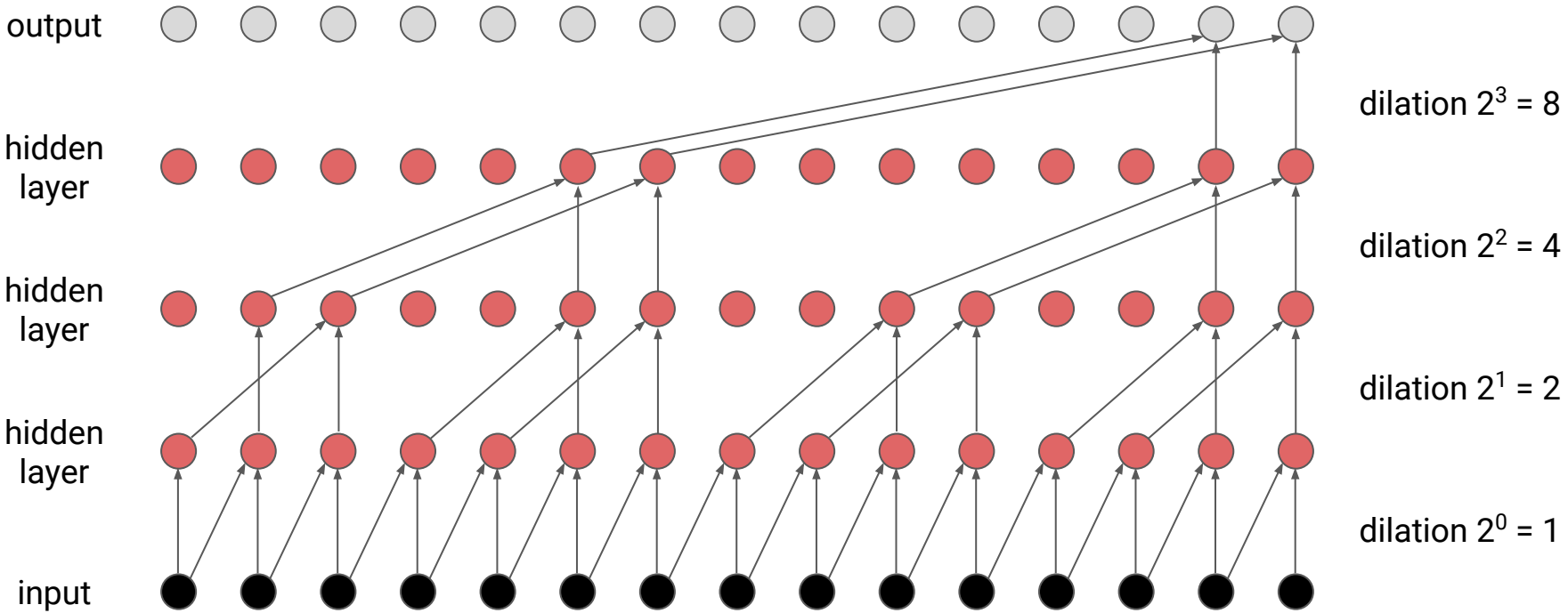
van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

WaveNet: dilated convolutions

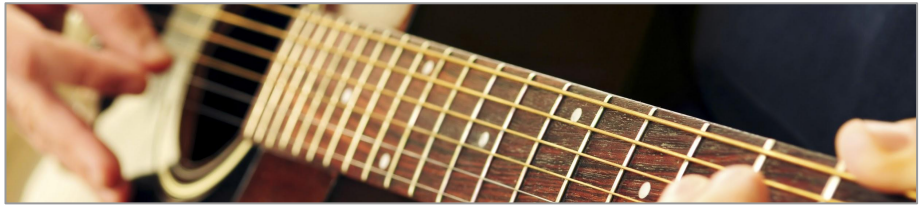


van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

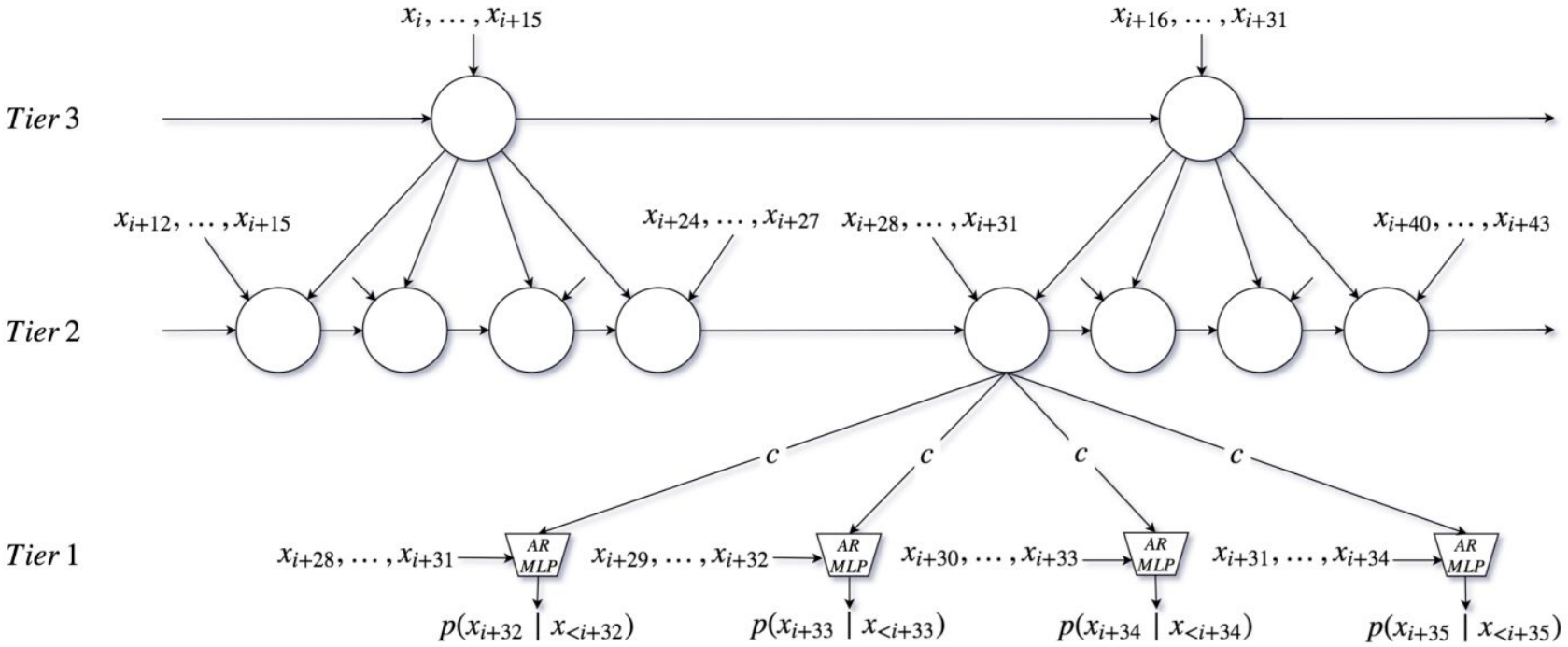
WaveNet: dilated convolutions



van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

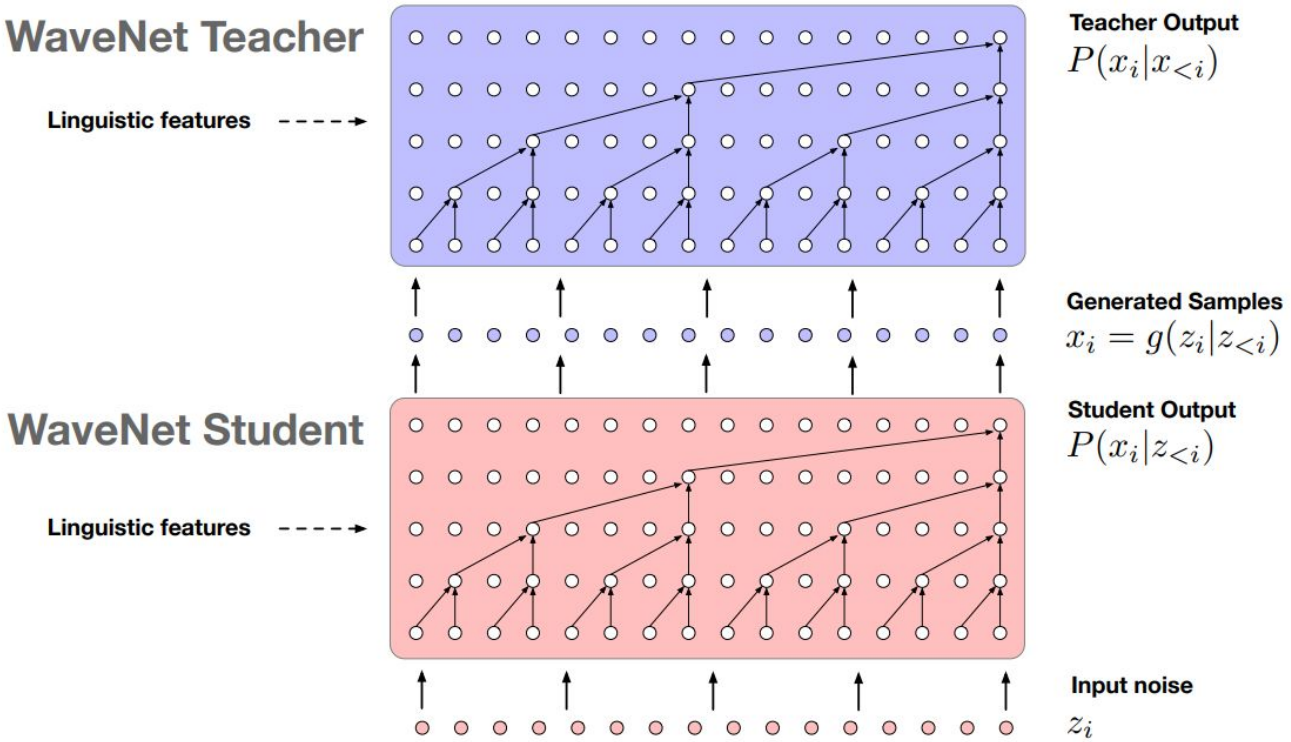


SampleRNN



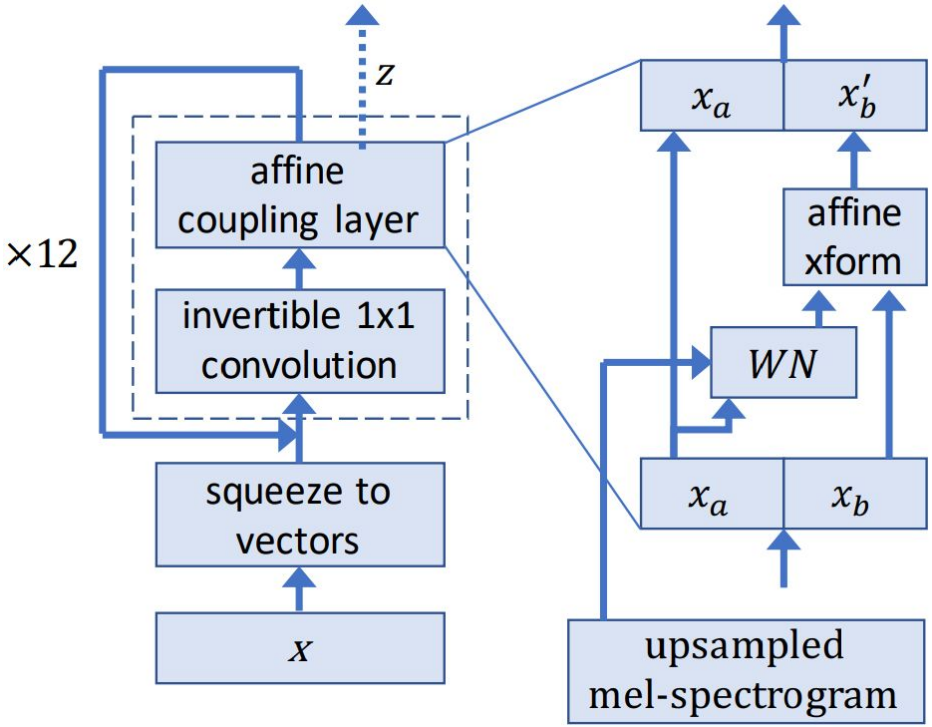
Mehri et al., 2017. “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model”, ICLR.

Parallel WaveNet, ClariNet



van den Oord et al., 2019. “Parallel WaveNet: Fast High-Fidelity Speech Synthesis”, ICML.

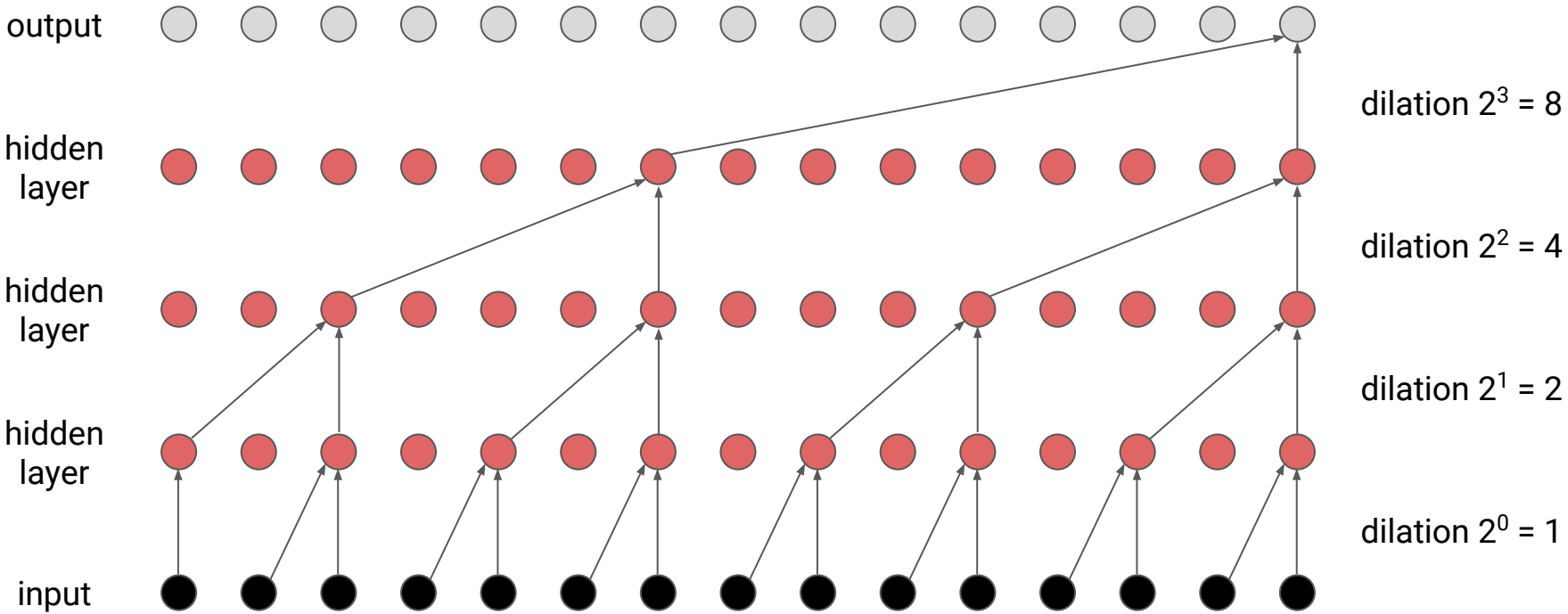
WaveGlow, FloWaveNet



Prenger et al., 2019. “Waveglow: A flow-based generative network for speech synthesis”, ICASSP.

Kim et al., 2019. “FloWaveNet: A generative flow for raw audio”, ICML.

WaveNet: #layers ~ log(receptive field length)



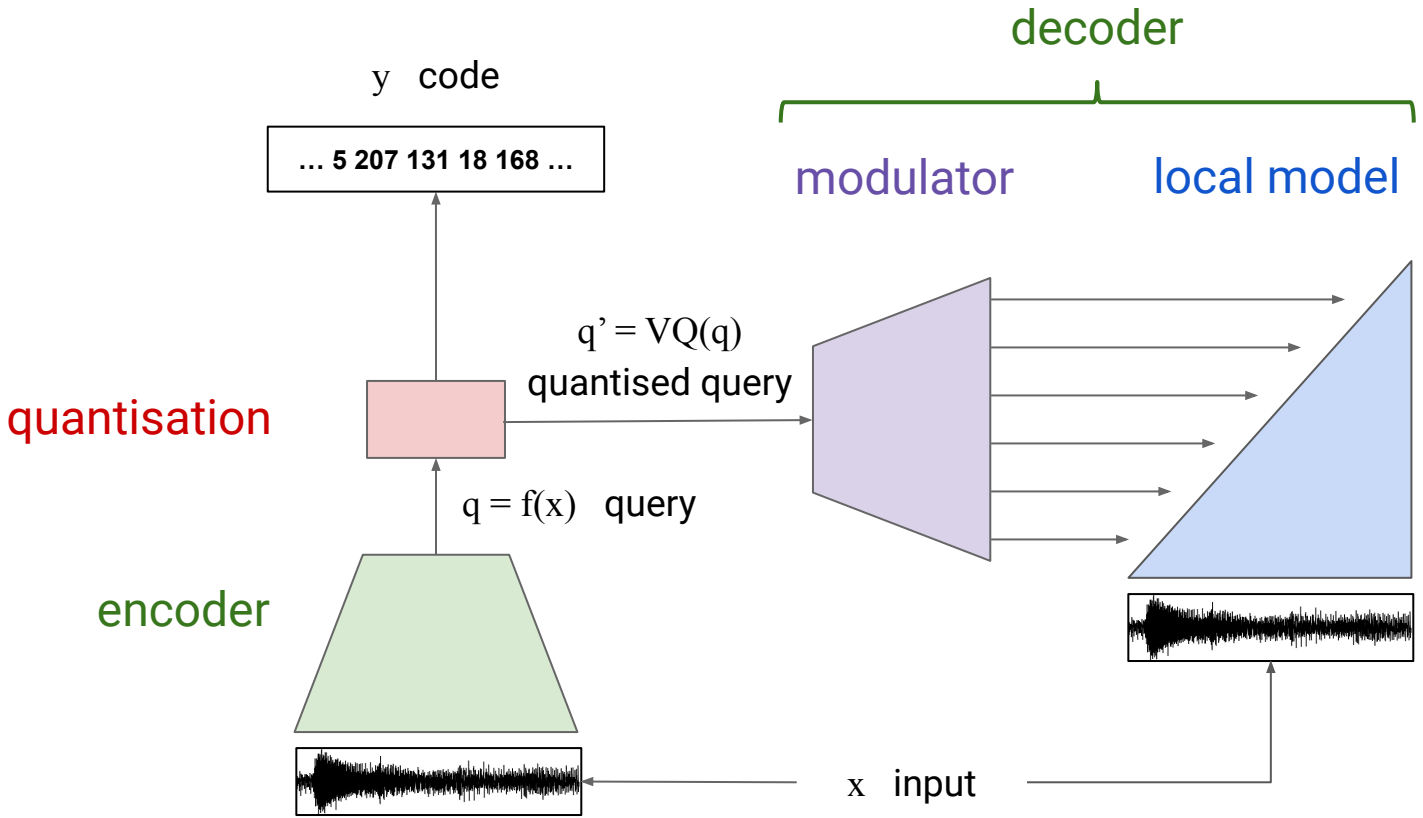
… but memory usage ~ receptive field length

Required model depth is **logarithmic** in the desired receptive field length

Required memory usage during training is still **linear** in the desired receptive field length!

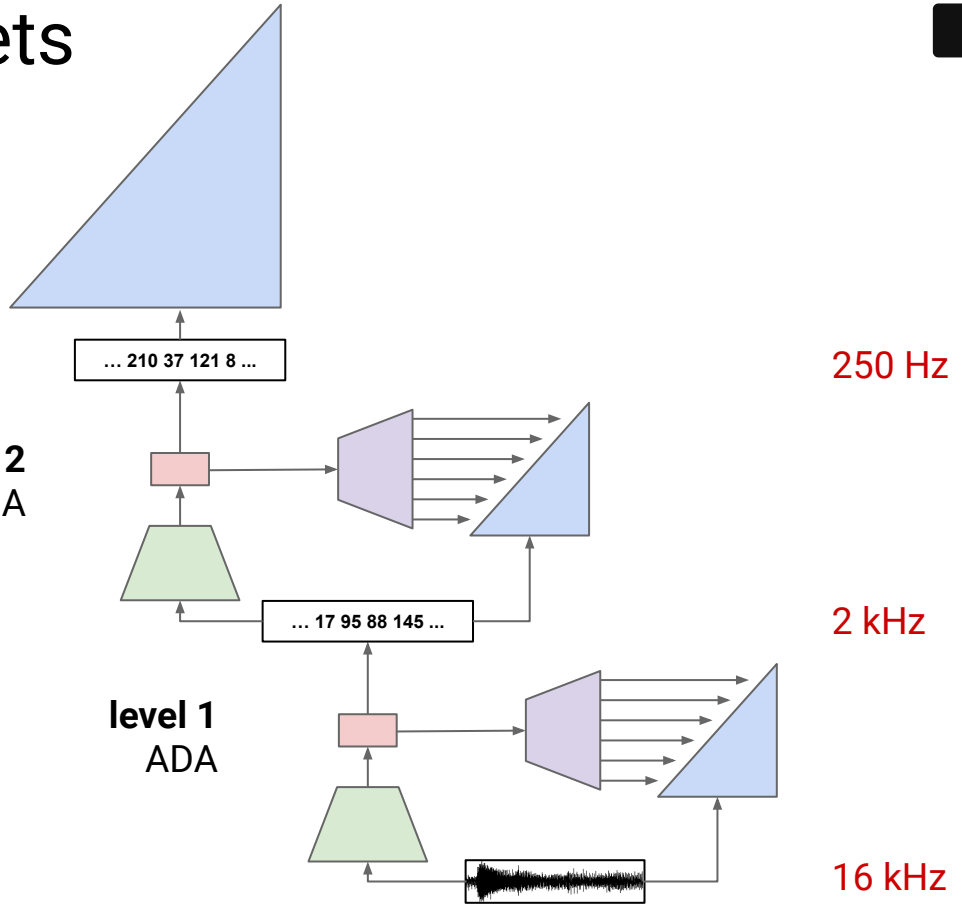
⇒ We cannot scale indefinitely using dilation

Autoregressive discrete autoencoders



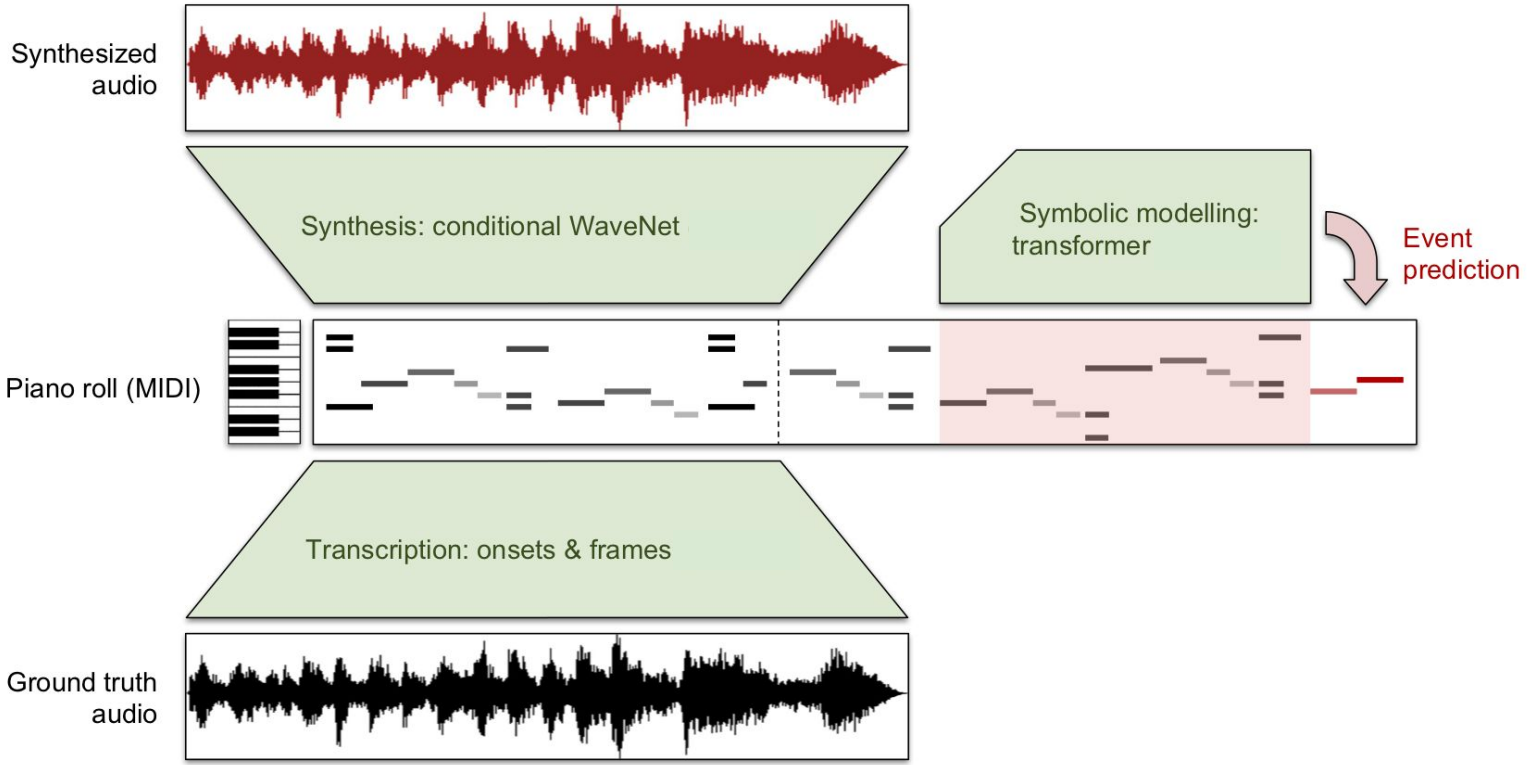
van den Oord et al., 2017. "Neural discrete representation learning", NeurIPS.

Hierarchical WaveNets



Dieleman et al., 2018. “The challenge of realistic music generation: modelling raw audio at scale”, NeurIPS.

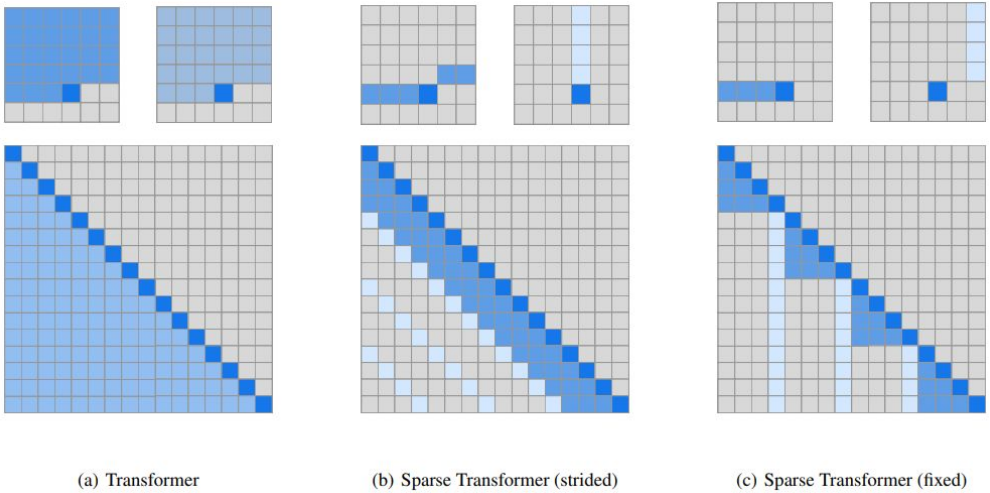
Wave2Midi2Wave and the MAESTRO dataset



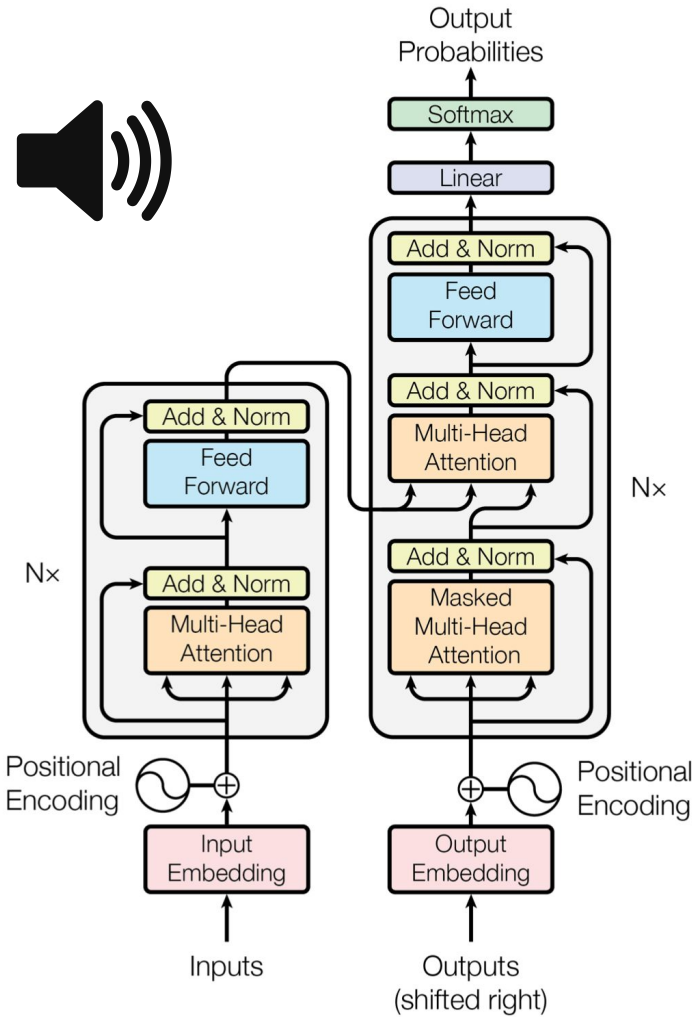
Hawthorne et al., 2019. “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset”, ICLR.

Sparse transformers

Attention (with sparse masks) instead of recurrence / convolutions.



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

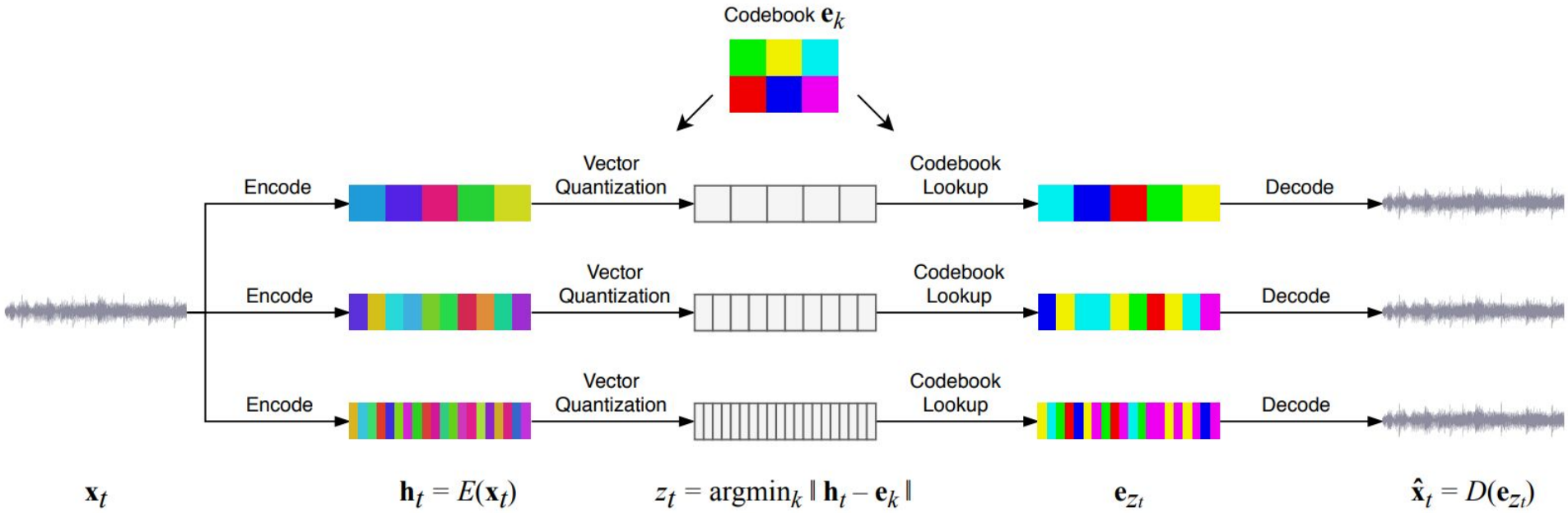


Child et al., 2019. “Generating long sequences with sparse transformers”, arXiv.

https://openai.com/blog/jukebox/

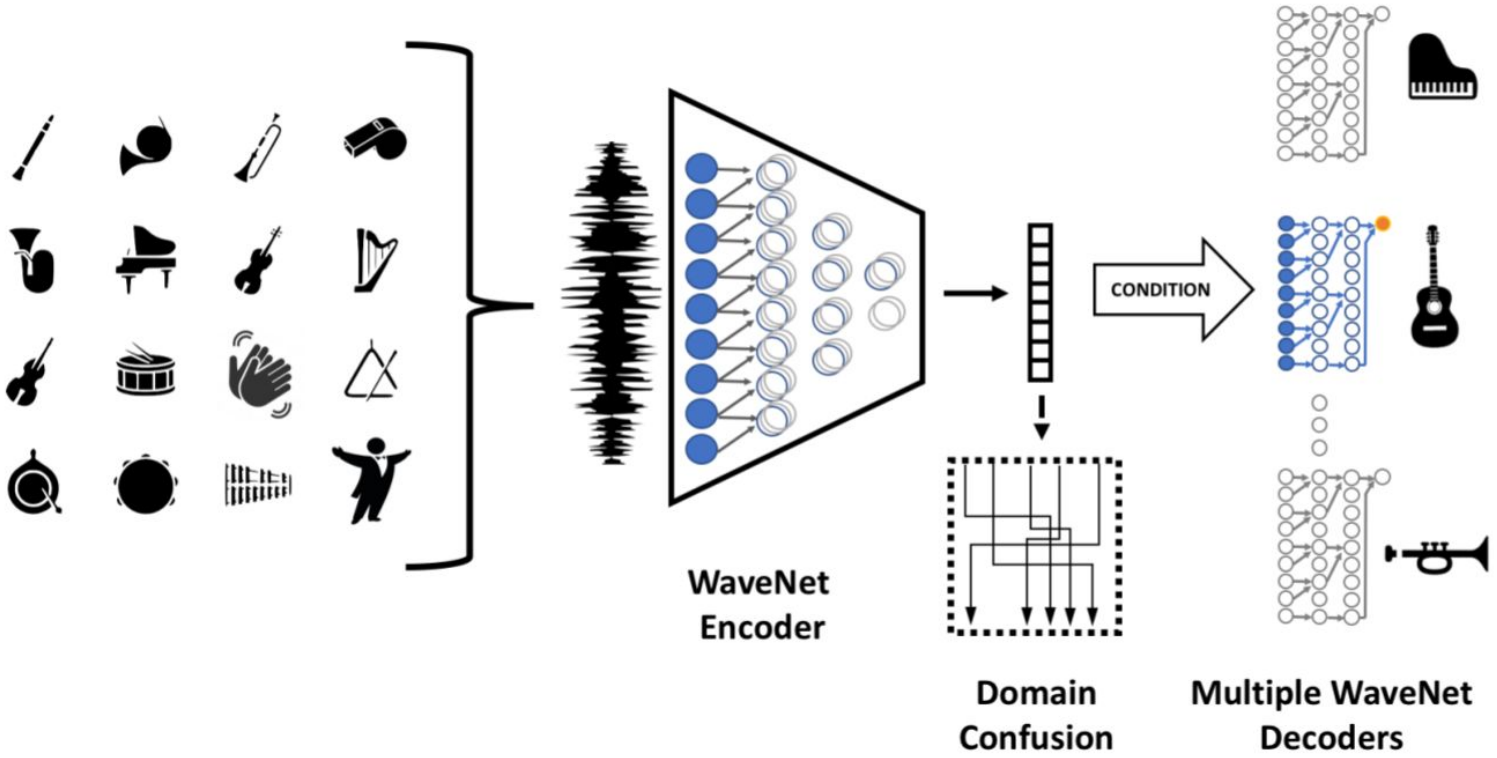
Jukebox

VQ-VAE + Transformer at scale



Dhariwal et al., 2020. “Jukebox: A Generative Model for Music”, arXiv.

Universal music translation network



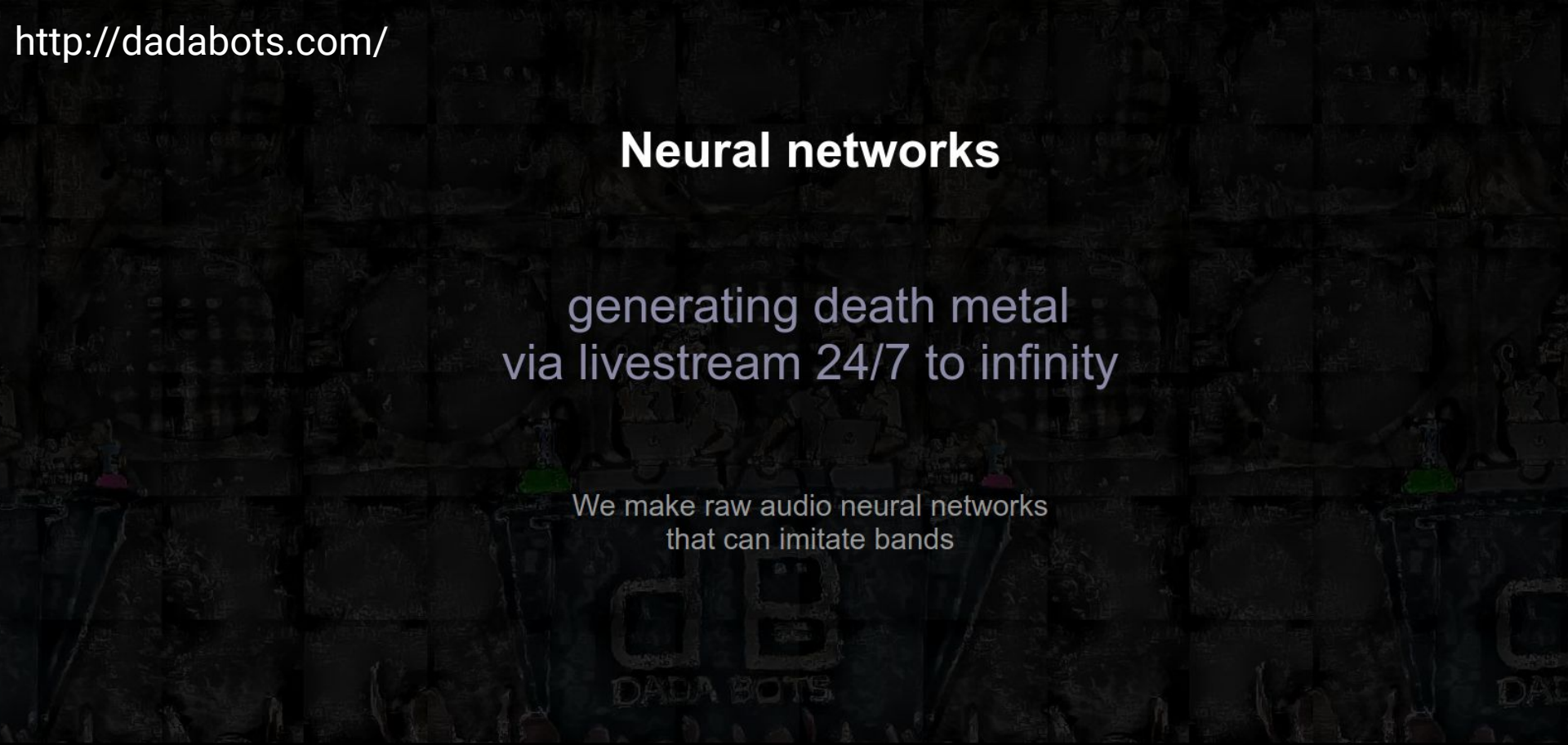
Mor et al., 2019. “A Universal Music Translation Network”, ICLR.

http://dadabots.com/

Neural networks

generating death metal
via livestream 24/7 to infinity

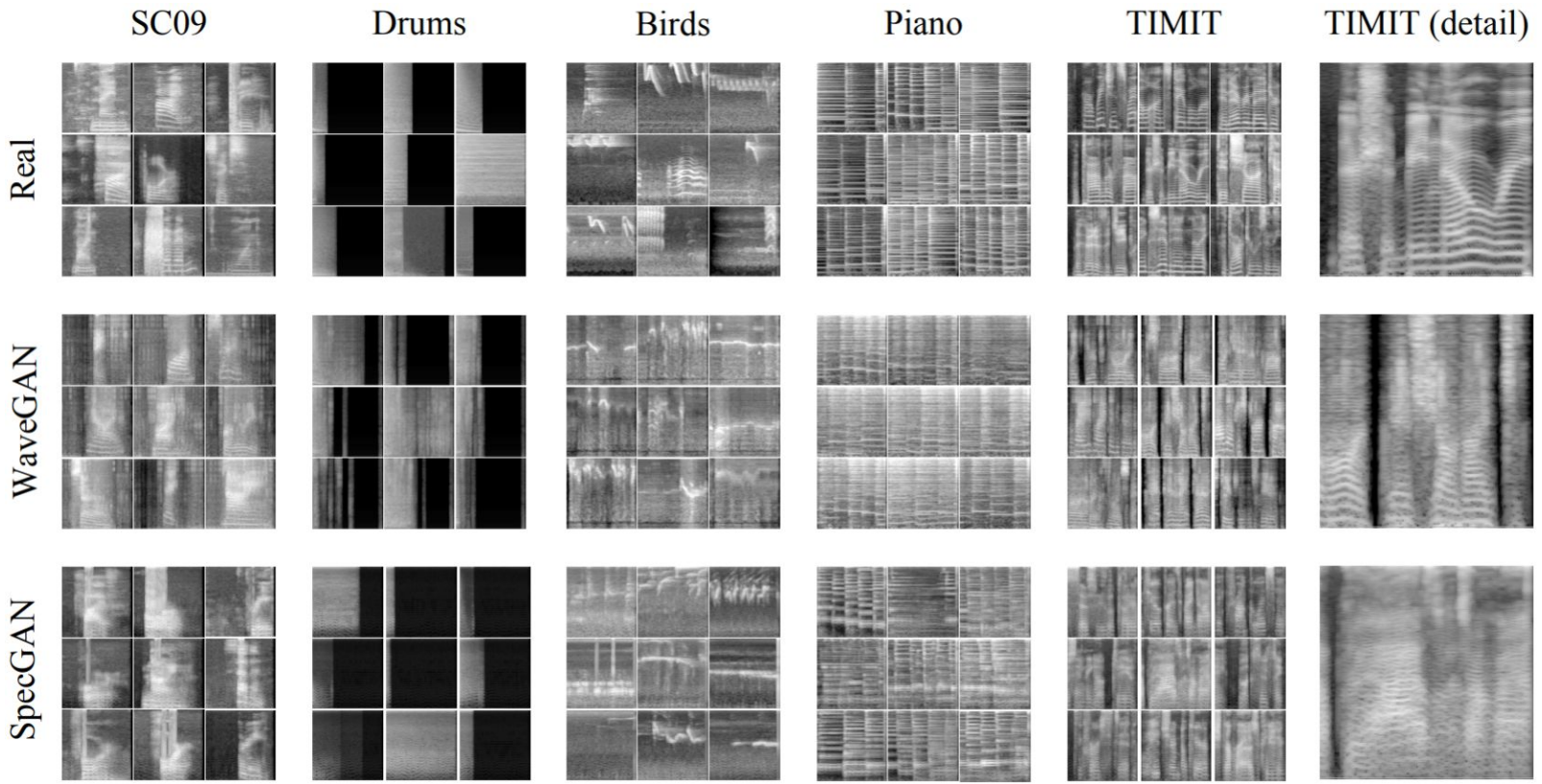
We make raw audio neural networks
that can imitate bands



Carr & Zukowski, 2018. “Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands”, arXiv.

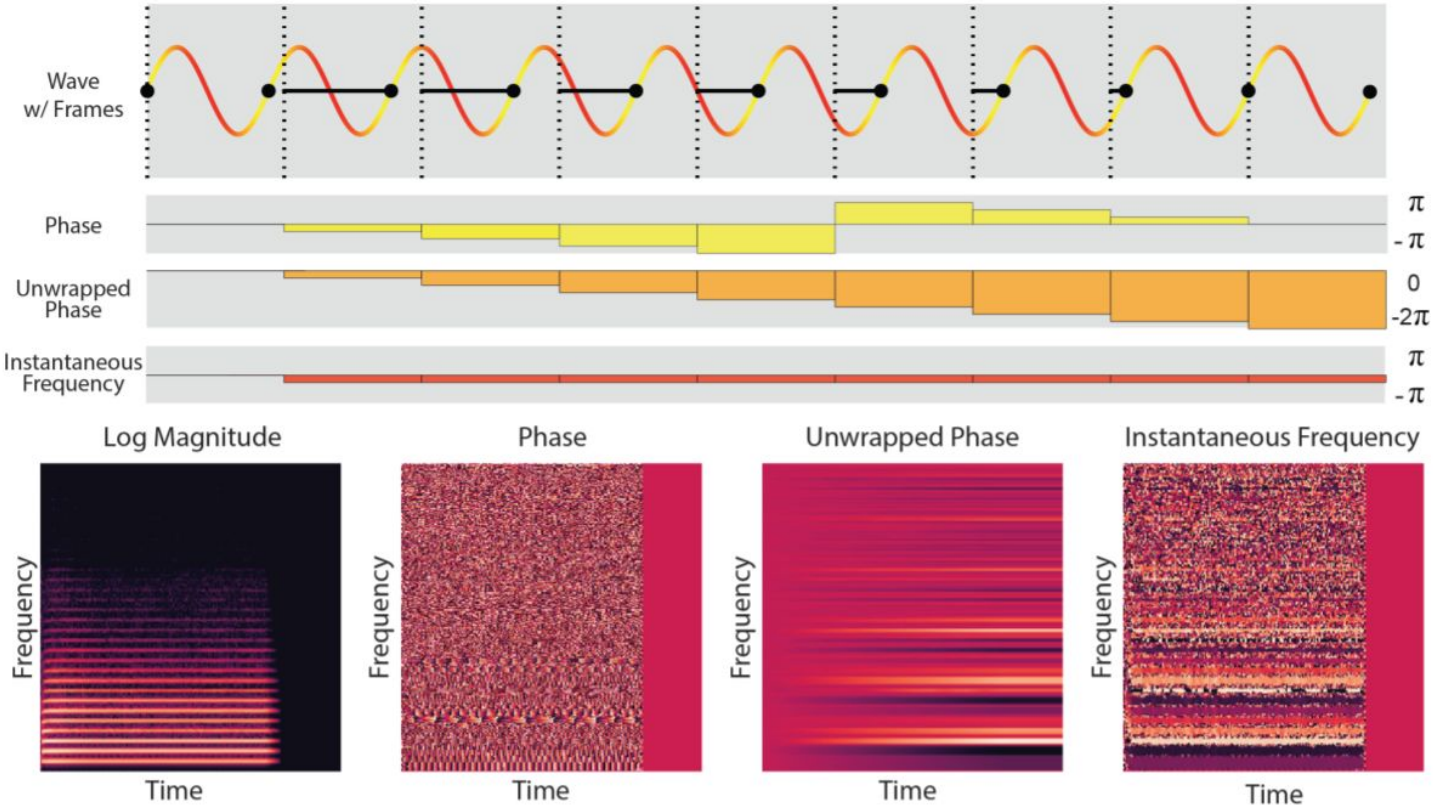
Adversarial models

WaveGAN



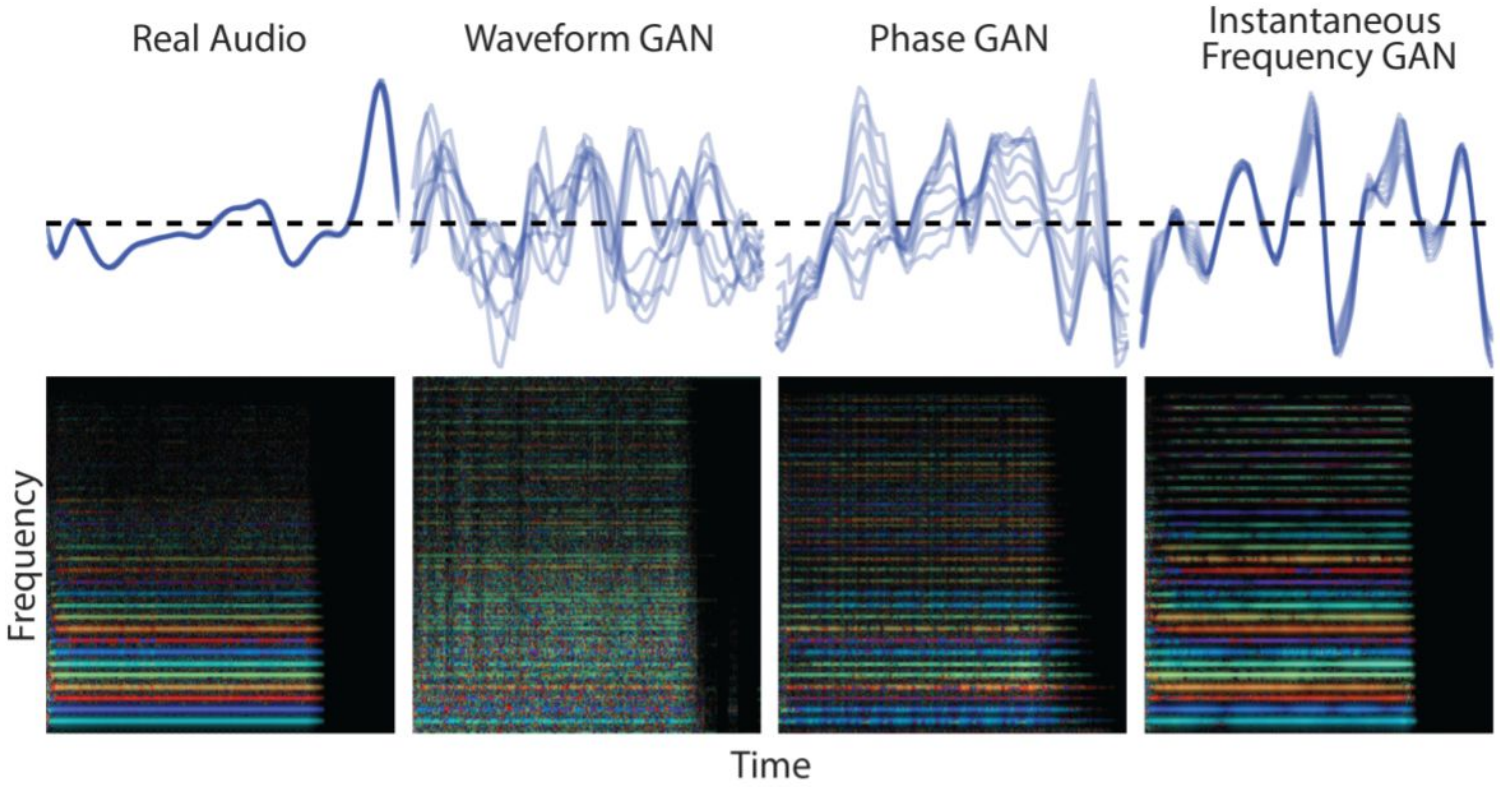
Donahue et al., 2019. “Adversarial Audio Synthesis”, ICLR.

GANSynth



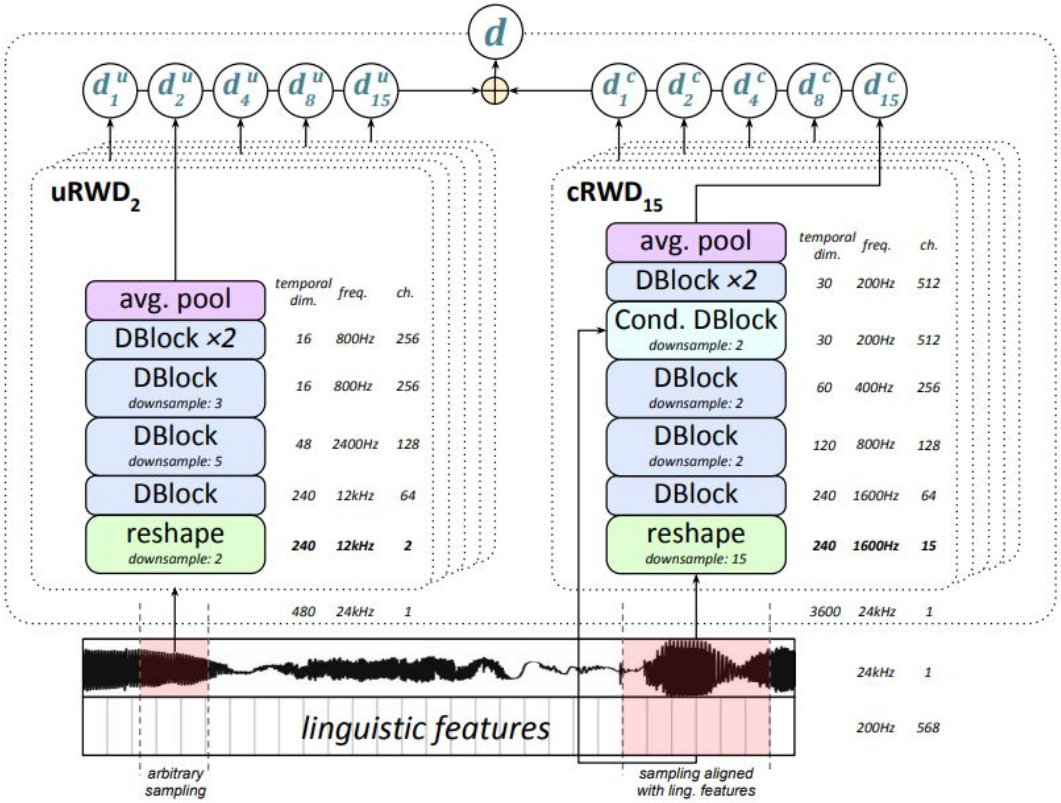
Engel et al., 2019. “GANSynth: Adversarial Neural Audio Synthesis”, ICLR.

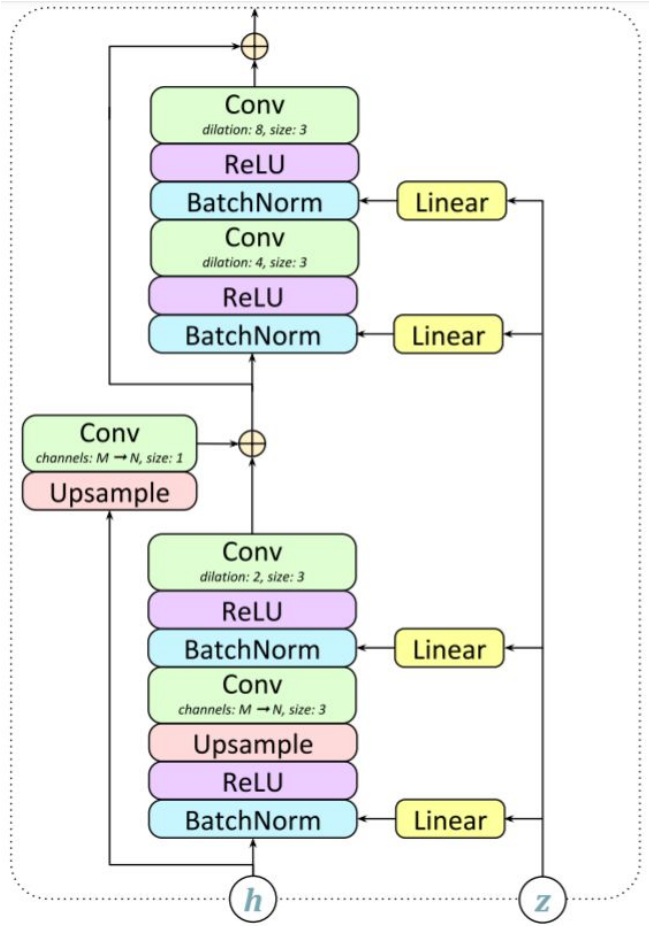
GANSynth



Engel et al., 2019. “GANSynth: Adversarial Neural Audio Synthesis”, ICLR.

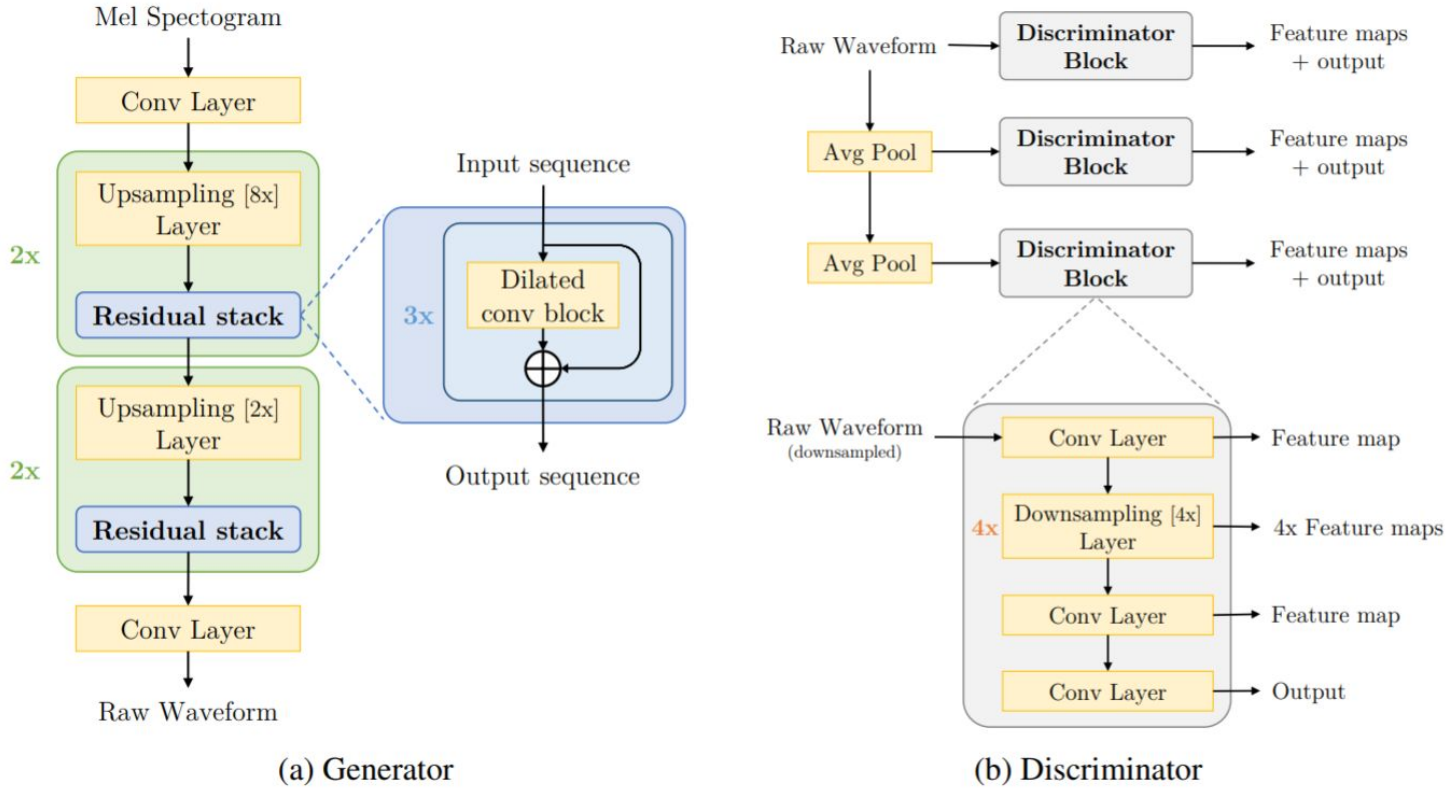
GAN-TTS





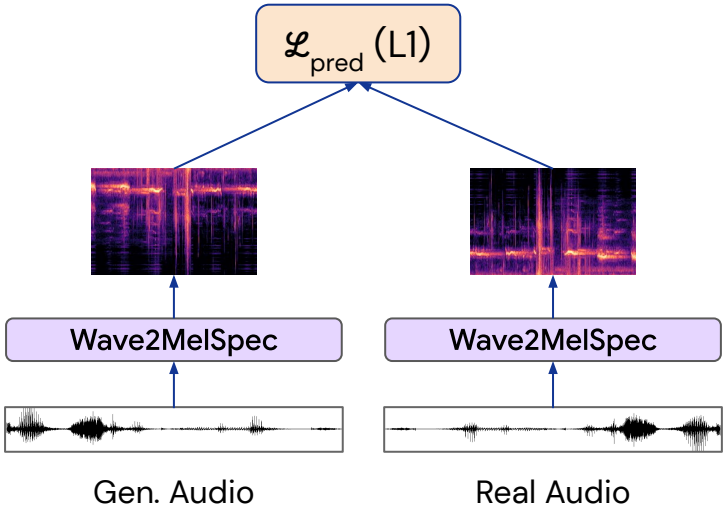
Binkowski et al., 2019. “High Fidelity Speech Synthesis with Adversarial Networks”, arXiv.

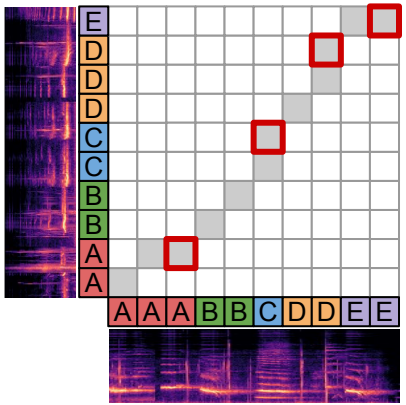
MelGAN

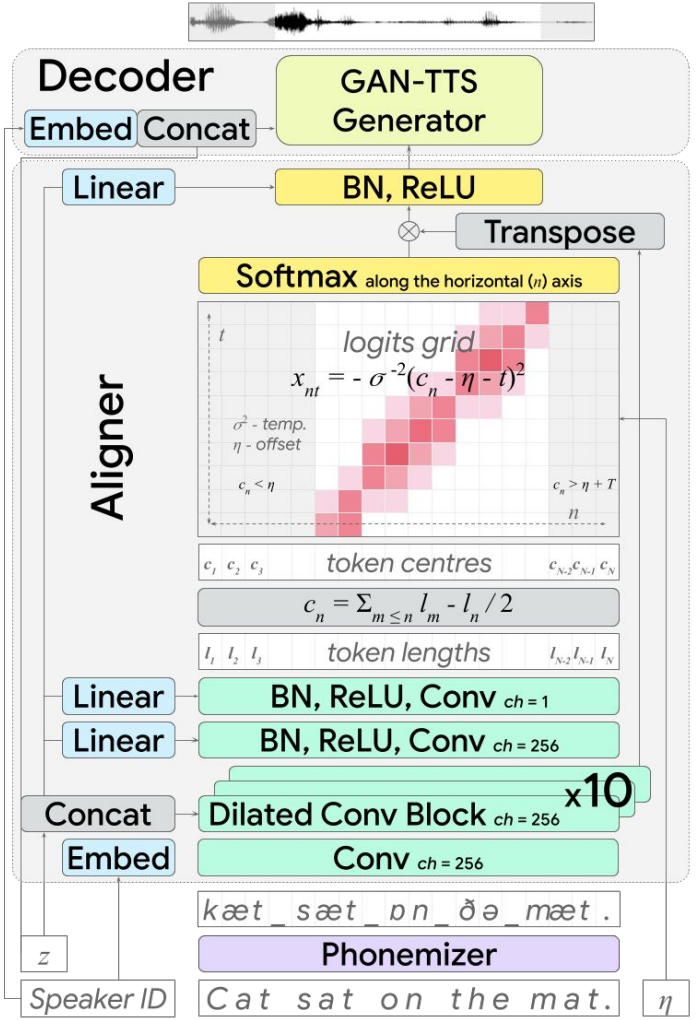


Kumar et al., 2019. “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis”, arXiv.

EATS







Donahue et al., 2020. “End-to-End Adversarial Text-to-Speech”, arXiv.

Why the emphasis on likelihood in music modelling?

Most popular generative modelling paradigm:

GANs

Why the emphasis on likelihood in music modelling?

Most popular generative modelling paradigm:

GANs

Most popular generative modelling paradigm for music:

likelihood-based (autoregressive)

Why the emphasis on likelihood in music modelling?

- We are still figuring out the right architectural priors for audio discriminators
  - For images, a stack of convolutions is all you need
  - What do we need for audio? Multiresolution? Dilation? Something phase shift invariant?

Ulyanov et al., 2018. “Deep Image Prior”, CVPR.

Why the emphasis on likelihood in music modelling?

- We are still figuring out the right architectural priors for audio discriminators
  - For images, a stack of convolutions is all you need
  - What do we need for audio? Multiresolution? Dilation? Something phase shift invariant?

- The sparsely-conditioned setting is dominant
  - We care about “creativity” and capturing diversity
  - GANs are worse at this than likelihood-based models

Ulyanov et al., 2018. “Deep Image Prior”, CVPR.

Alternatives to modelling raw audio directly

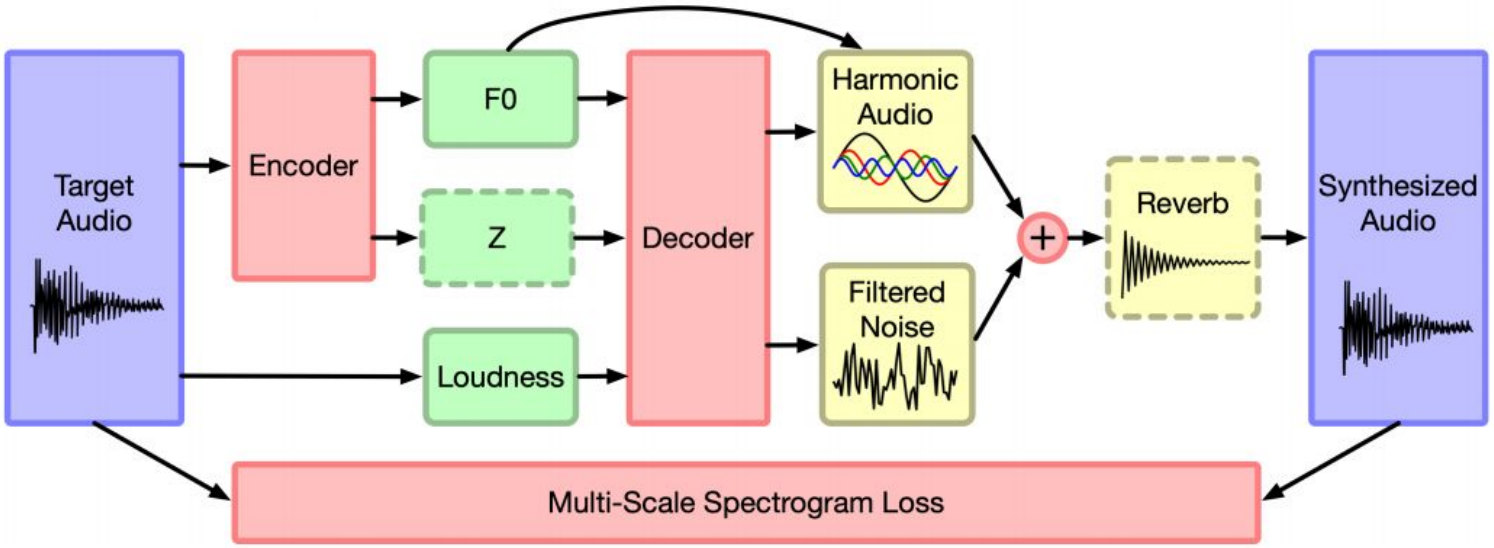
- Model complex-valued spectrograms and “deal” with phase (GANSynth)

- Model magnitude spectrograms
  Use a vocoder or Griffin-Lim to invert
  (Tacotron 1 & 2, MelGAN, MelNet, …)

Wang et al., 2017. “Tacotron: Towards end-to-end speech synthesis”, ISCA.
Shen et al., 2018. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, ICASSP.
Vasquez & Lewis, 2019. “MelNet: A Generative Model for Audio in the Frequency Domain”, arXiv.

Alternatives to modelling raw audio directly

- Differentiable Digital Signal Processing
  Use raw audio input, but put DSP components in the model



Engel et al., 2020. “Differentiable Digital Signal Processing”, ICLR 2020.

Summary

- Generative modelling of raw audio is feasible, even in the sparsely conditioned setting
- Likelihood-based models dominate, but GANs are making in-roads in the densely conditioned setting
- Modelling large-scale structure from raw audio is an unsolved problem

Thank you

https://benanne.github.io/2020/03/24/audio-generation.html

sanderdieleman@gmail.com

 @sedielem