PAT 498/598 (Fall 2024)

Special Topics: Generative AI for Music and Audio Creation

Lecture 16: Frequency-domain Audio Synthesis

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

Assignment 3: Open-ended Assignment

- **Two options!**
  - Unconditional **symbolic music generation**
  - Unconditional **music/audio synthesis**

Assignment 3: (Option 1) Symbolic Music Generation

- Use **any dataset of your choice**
  - **Nottingham** (ABC): github.com/jukedeck/nottingham-dataset
  - **JSB Chorales** (NPY): github.com/czhuang/JSB-Chorales-dataset/tree/master
  - **POP 909** (MIDI; melody or accompaniment): github.com/music-x-lab/POP909-Dataset
  - **MAESTRO** (MIDI): magenta.tensorflow.org/datasets/maestro
  - **Groove MIDI** (MIDI): magenta.tensorflow.org/datasets/groove
  - Your own corpus!?
- Useful libraries:
  - **mido**, **pretty_midi** and **MusPy** for processing MIDI files
  - **MidiTok** for tokenizing MIDI files

Assignment 3: (Option 2) Music/Audio Synthesis

- Use **any dataset of your choice**
  - **NSynth**: magenta.tensorflow.org/datasets/nsynth
  - **MAESTRO**: magenta.tensorflow.org/datasets/maestro
  - **Bach Violin**: hermandong.com/bach-violin-dataset/
  - **DCASE Foley**: dcase.community/challenge2023/task-foley-sound-synthesis
  - **ESC-50**: github.com/karolpiczak/ESC-50
  - Your own corpus!?
- Useful libraries:
  - **librosa** for loading and processing audio

Assignment 3: Open-ended Assignment

- Implement **any model of your choice**
- Using existing codebase is fine, but you need to provide proper references!
- Report the **training, validation and test losses**
- Show some **generated music/audio**

Assignment 3: (Option 1) Symbolic Music Generation

- A simple option
  - **Implement an n-gram-like model using MLPs or CNNs**
  - **Input**: $(\boldsymbol{x_{t-n}}, \boldsymbol{x_{t-n+1}}, \ldots, \boldsymbol{x_{t-1}})$
  - **Output**: $\boldsymbol{x_t}$
  - Compare the performance of the model for different n, say $n = 1, 10, 100, 1000$

Assignment 3: (Option 2) Music/Audio Synthesis

- A not-too-challenging (though not easy either) option
  - **Implement a diffusion model that generates audio spectrograms**
  - Using this nicely-written codebase: github.com/openai/improved-diffusion
  - **Input**: 64x64 noise
  - **Output**: 64x64 mel spectrograms
  - You might want to use the pretrained Hifi-GAN as your vocoder

Assignment 3: Open-ended Assignment

- Instructions will be released on Gradescope
- Due at **11:59pm ET** on **November 18**
- Late submissions: **3 point deducted per day**

Final Project

- Any topic of your choice that **involves some generative models**
- Group size: 1 or 2
- Can be an extension of your assignment 3 results
- Some suggestions
  - Think about "**data & model**" at the same time
  - An active GitHub codebase with **many open/closed issues** is usually a good sign
  - Always look for **backup codebase** so that you have a plan B

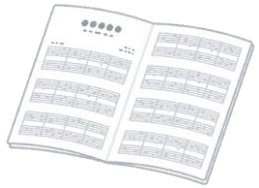
Final Project

- Milestones (all due at the specified date at **11:59 PM ET**)
  - **Pitch** November 6 Topic & high-level plans
  - **Proposal** November 18 Survey & plans (1 page)
  - **Presentation** December 9 Showcase & report
  - **Final report** December 15 Full report (3-5 pages)
- Instructions will be released on Gradescope
- Late submissions: **NOT accepted**

Updated Grading

- **Assignments** (40%)
  - Assignment 1 10%
  - Assignment 2 10%
  - Assignment 3 **20%**
- **Final Project** (60%)
  - Proposal 10%
  - Presentation **25%**
  - Final report **25%**

(Recap) Four Paradigms



Symbolic music generation

Text-based

Image-based

Audio-domain music generation

Time series-based

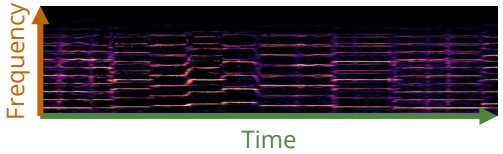
Image-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Piano roll

Waveform



Spectrogram

Today, we also have many latent-space based systems!

(Recap) Autoregressive Models (Mathematically)

- A class of machine learning models that learn the probability of the next value given previous values

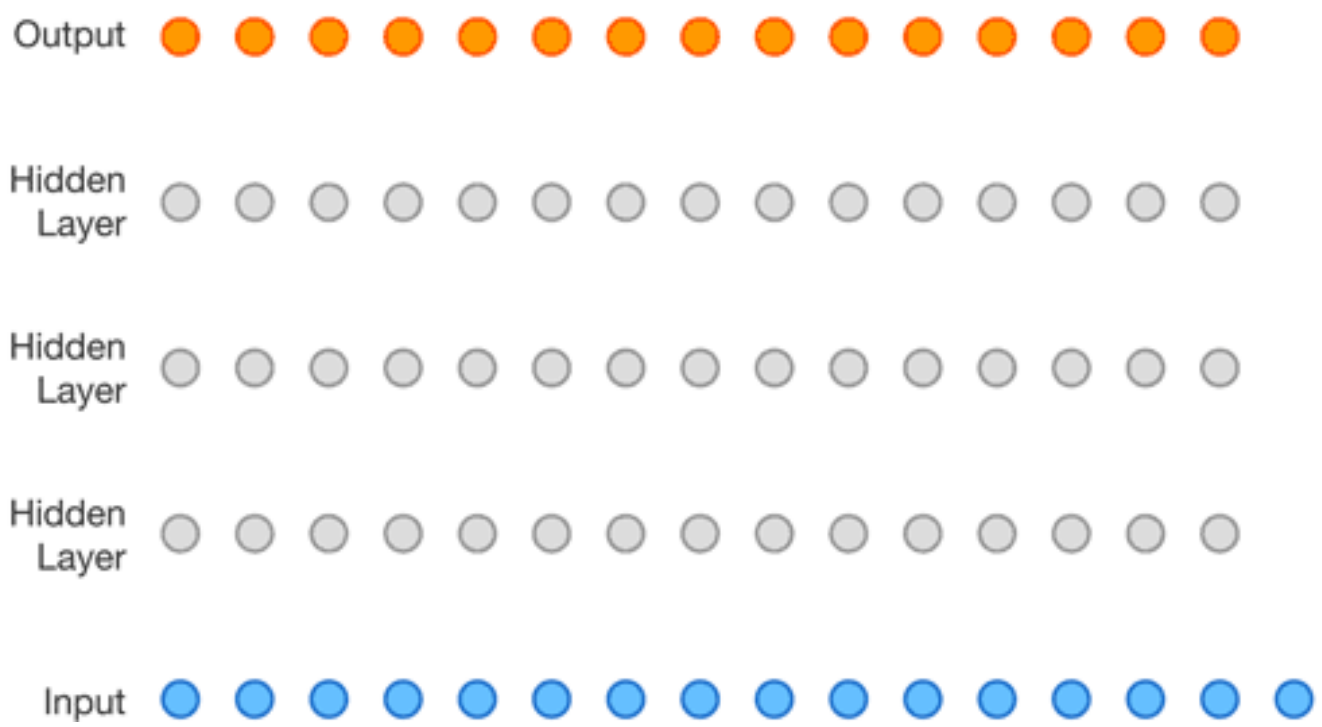
$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next number Previous numbers

$$P(\ 0.1 \mid 0.5, 0.4, 0.3, 0.2)$$
$$P(\ 0.09 \mid 0.5, 0.4, 0.3, 0.2)$$
$$P(\ 0.11 \mid 0.5, 0.4, 0.3, 0.2)$$
$$P(\ 99 \mid 0.5, 0.4, 0.3, 0.2)$$
$$P(\ -1 \mid 0.5, 0.4, 0.3, 0.2)$$

The term "autoregressive" has different definitions in machine learning and signal processing.
In signal processing, an autoregressive model needs to be a linear model.

(Recap) Example: WaveNet (van den Oord et al., 2016)

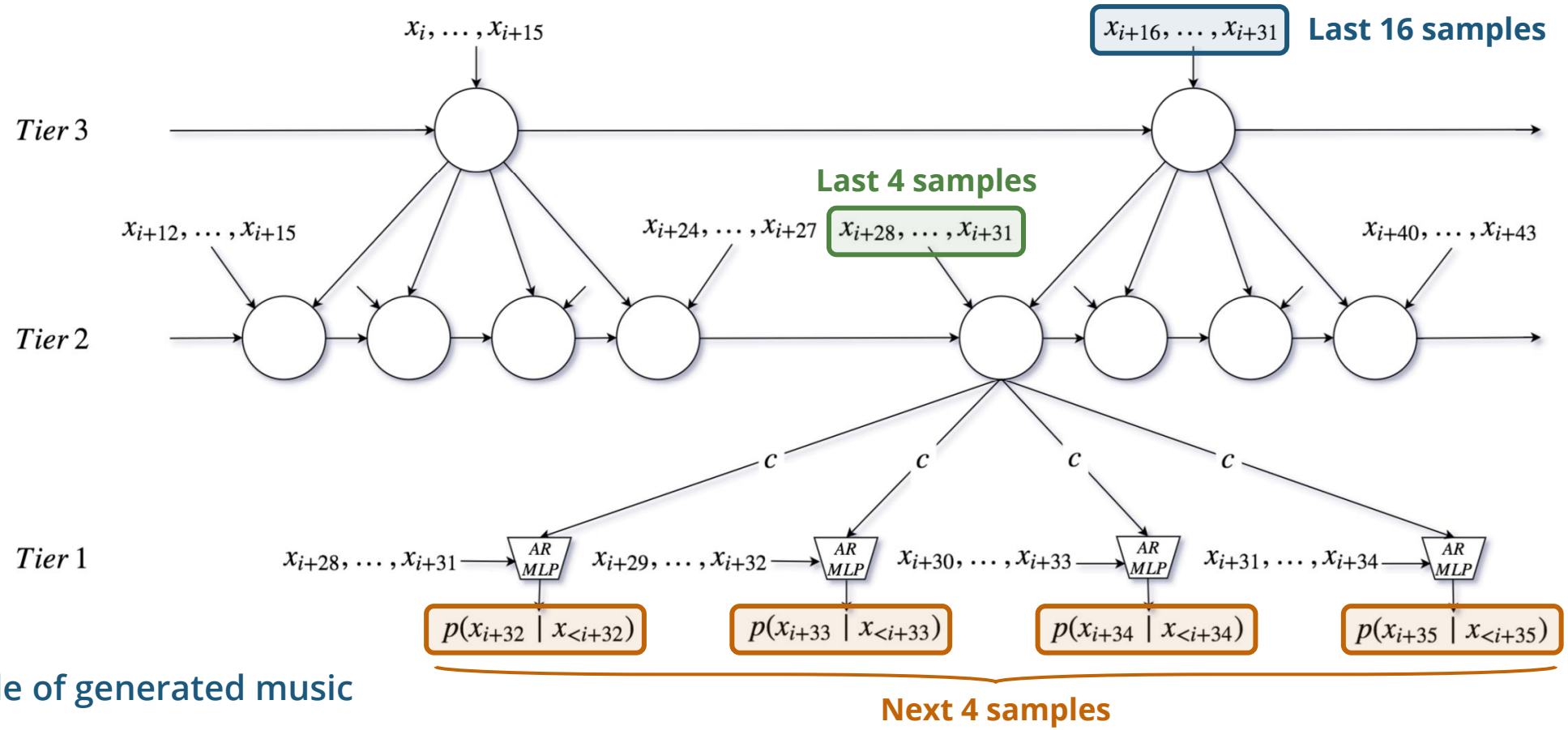


(Source: van den Oord et al., 2016)

A convolutional neural network for raw waveform generation

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

(Recap) Example: SampleRNN (Mehri et al., 2017)



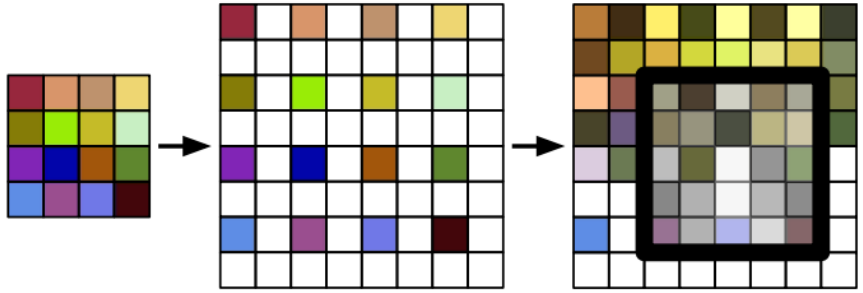
Example of generated music

(Source: Mehri et al., 2017)

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *ICLR*, 2017.

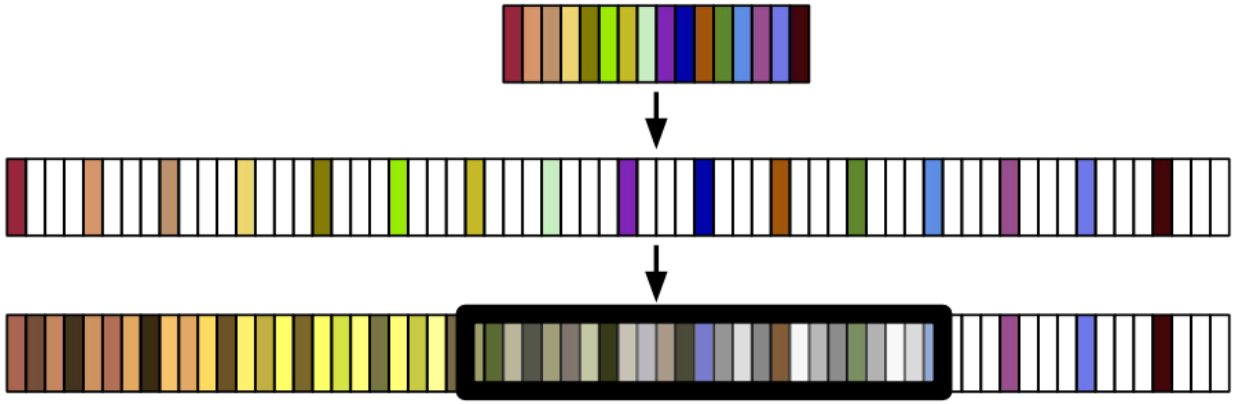
(Recap) Example: WaveGAN (Donahue et al., 2019)

DCGAN for images



5x5 kernels

WaveGAN for audio



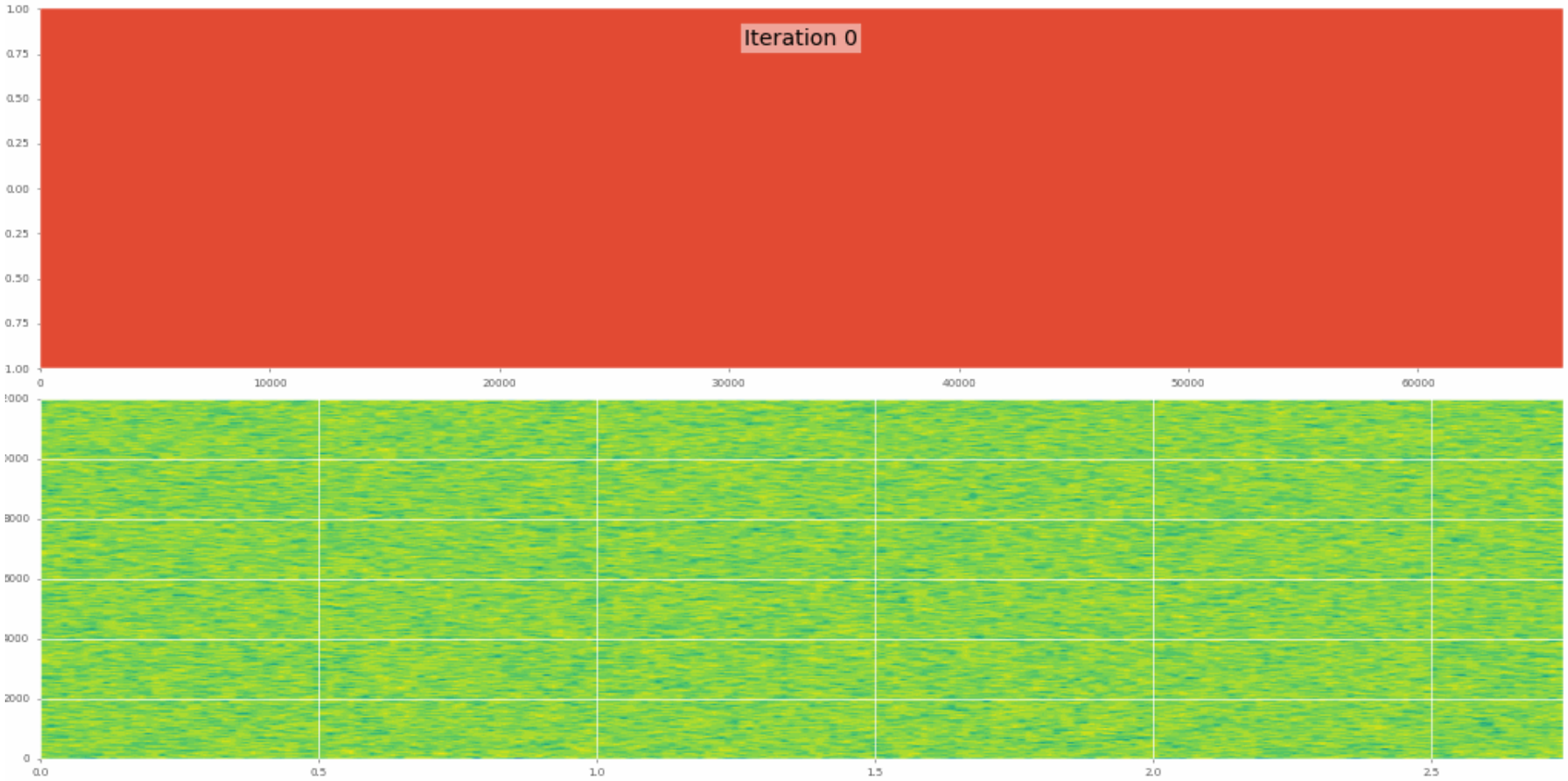
25x1 kernels

(Source: Donahue et al., 2019)

Example of generated music

Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

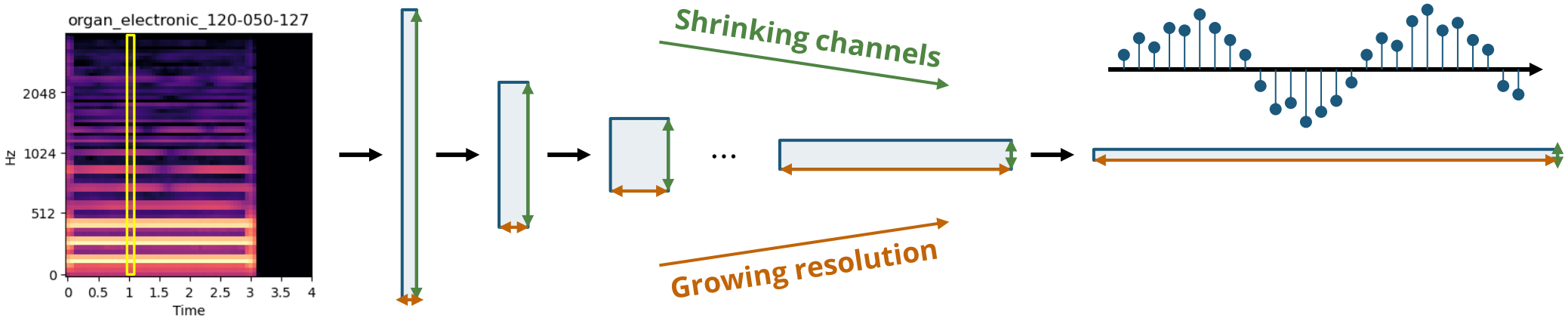
(Recap) Example: WaveGrad (Chen et al., 2021)



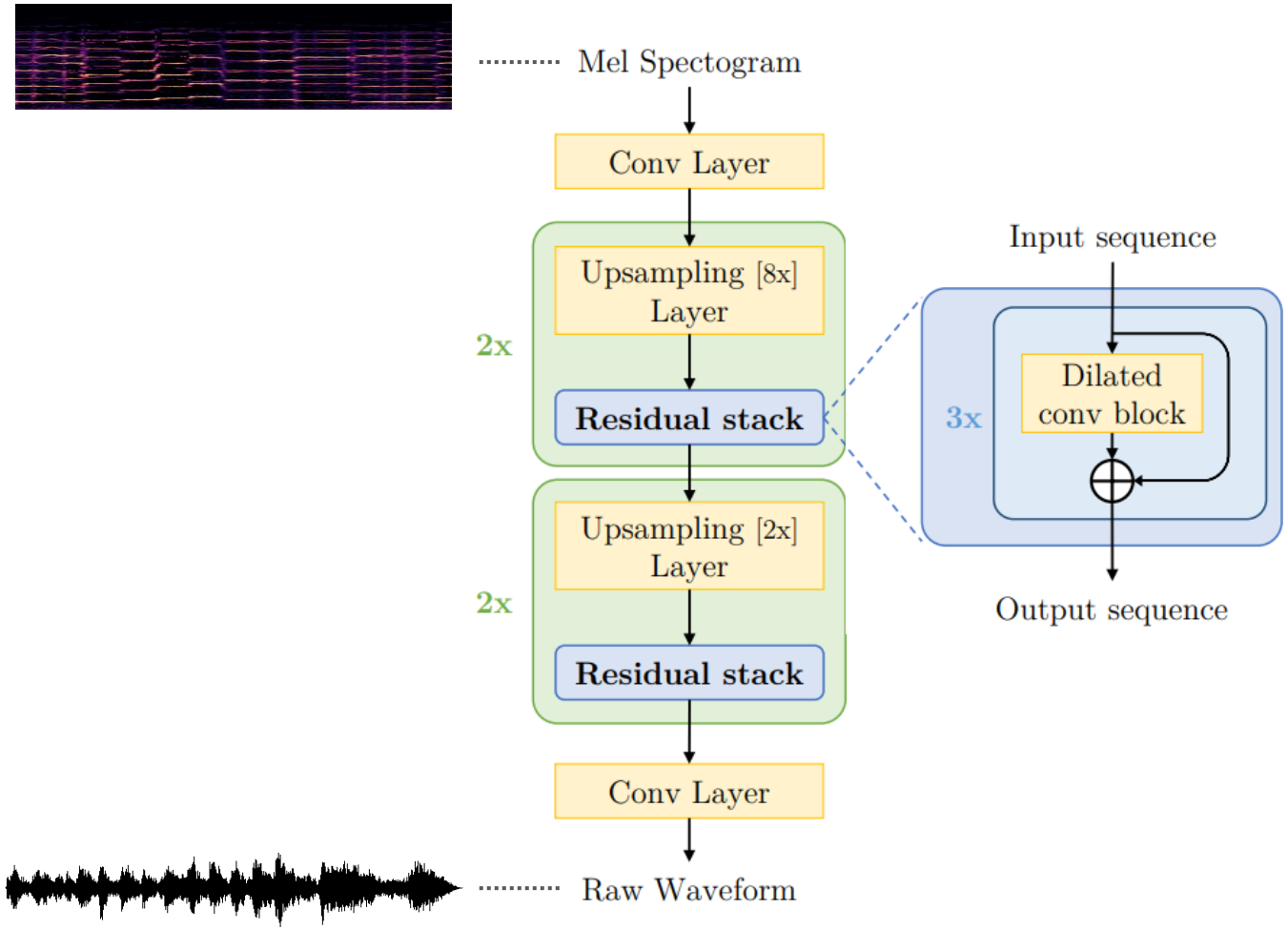
(Source: Chen et al., 2021)

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating Gradients for Waveform Generation," *ICLR*, 2021.

(Recap) Transposed Convolution for Vocoders



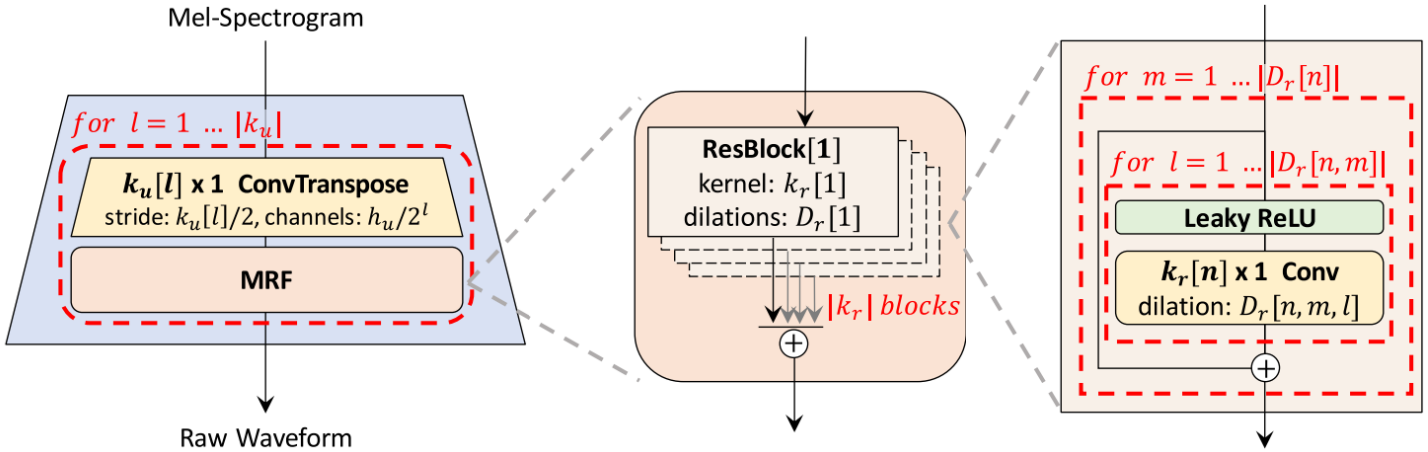
(Recap) Example: MelGAN (Kumar et al., 2019)

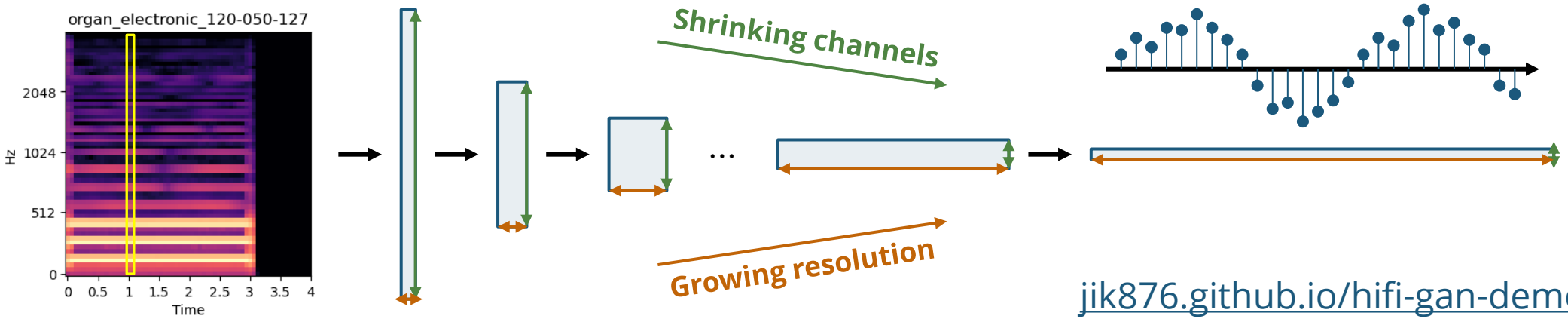


(Source: Kumar et al., 2019)

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.

(Recap) Example: Hifi-GAN (Kong et al., 2020)

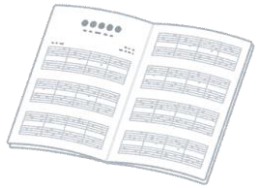




jik876.github.io/hifi-gan-demo

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, 2020.

(Recap) Four Paradigms



Symbolic music generation

Text-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Image-based

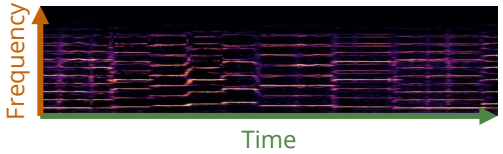
Piano roll

Audio-domain music generation

Time series-based

Waveform

Image-based

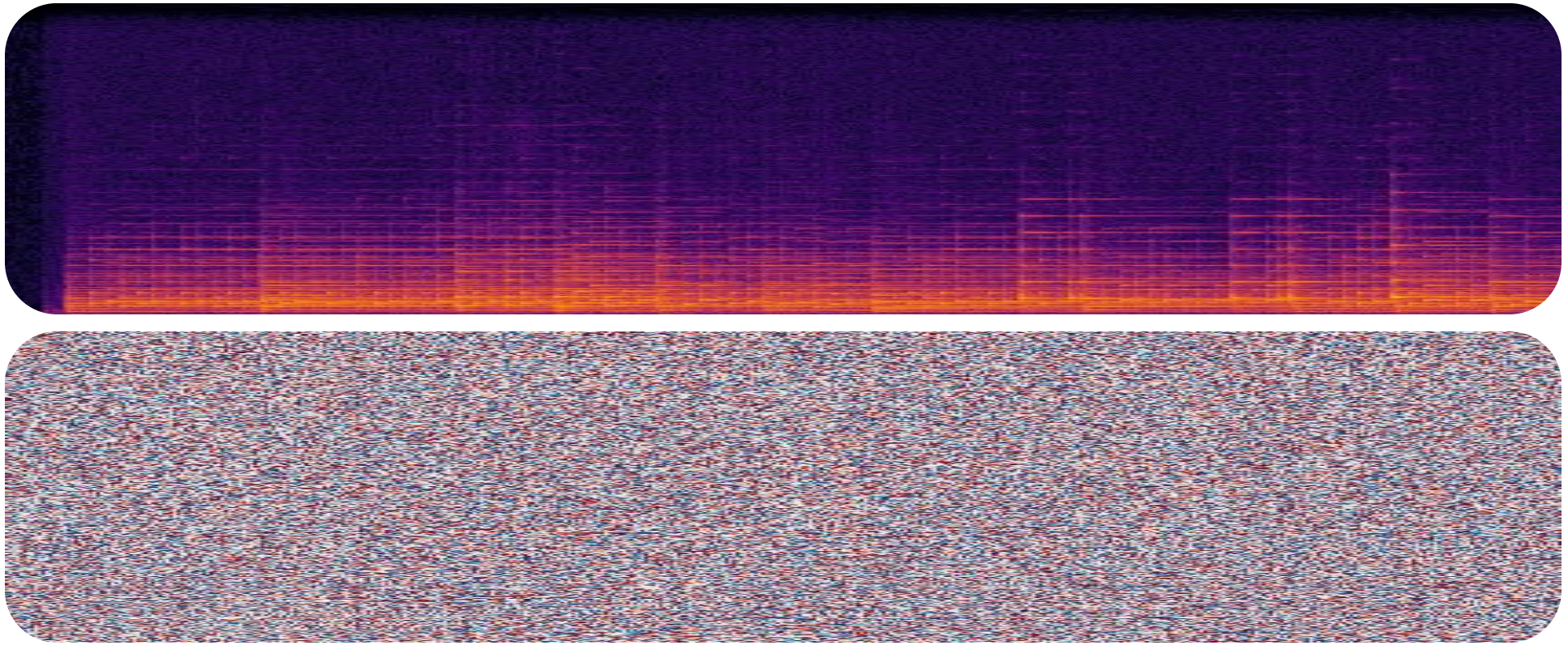


Spectrogram

Today, we also have many latent-space based systems!

Frequency-domain Audio Synthesis

Importance of the Phase Information

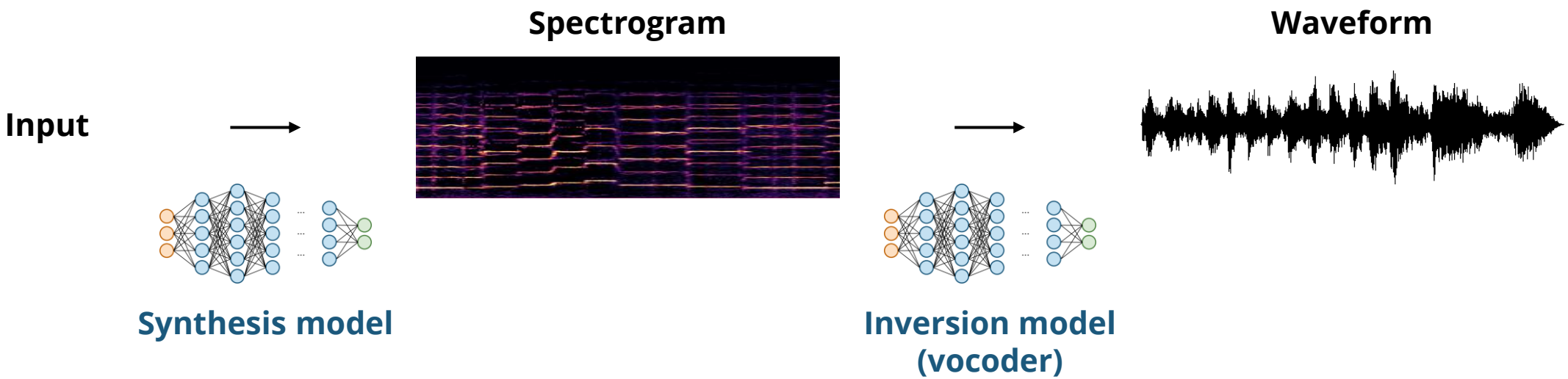


(Source: Dieleman et al., 2020)

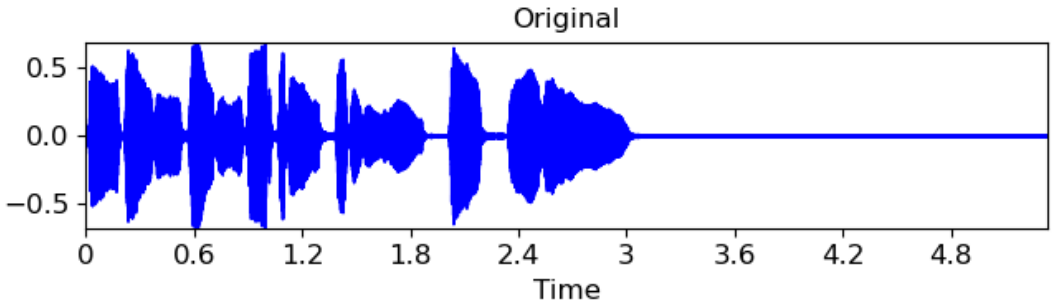
Real phase **Random phase**

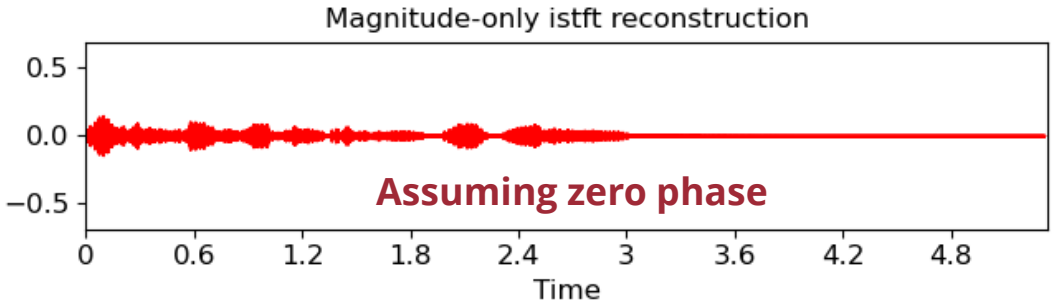
Sander Dieleman, "Generating music in the waveform domain," *Sander Dieleman's Blog*, March 24, 2020.

Frequency-domain Audio Synthesis



Inverse STFT without Phase Information





(Source: librosa documentation)

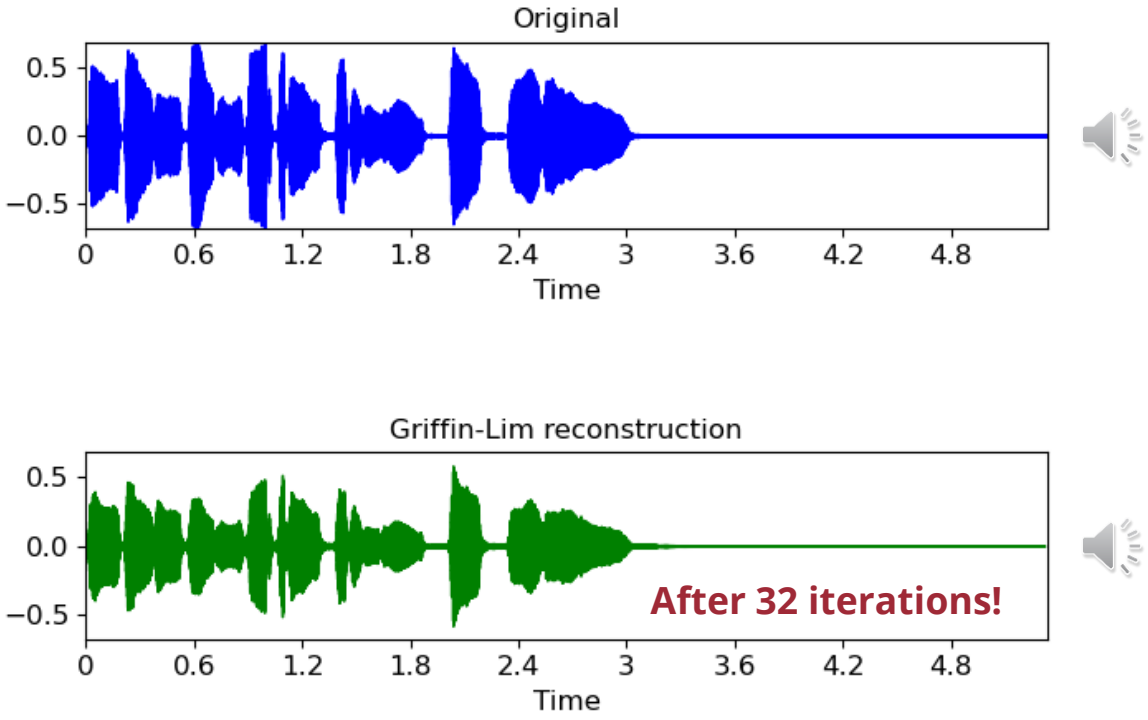
Complex-valued STFT matrix

$$\text{ISTFT}(M) = \arg\min_{y} (M - \text{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

librosa.org/doc/main/generated/librosa.griffinlim.html

Griffin-Lim Algorithm (Griffin & Lim, 1984)



(Source: librosa documentation)

Given a magnitude-only STFT matrix

↓

Randomly initialize the phase

↓

$$y' = \arg\min_{y} (M - \mathrm{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

↓

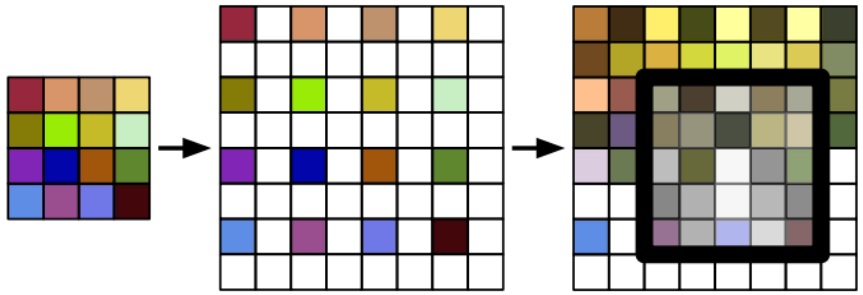
$$M' = \mathrm{STFT}(y')$$

Find the STFT of the signal y

Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, 1984.
librosa.org/doc/main/generated/librosa.griffinlim.html

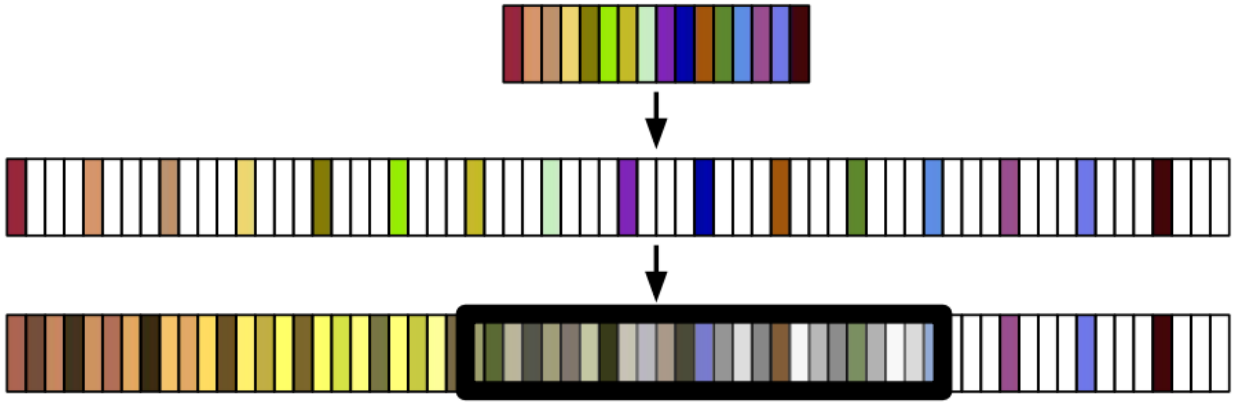
(Recap) Example: WaveGAN (Donahue et al., 2019)

DCGAN for images



5x5 kernels

WaveGAN for audio



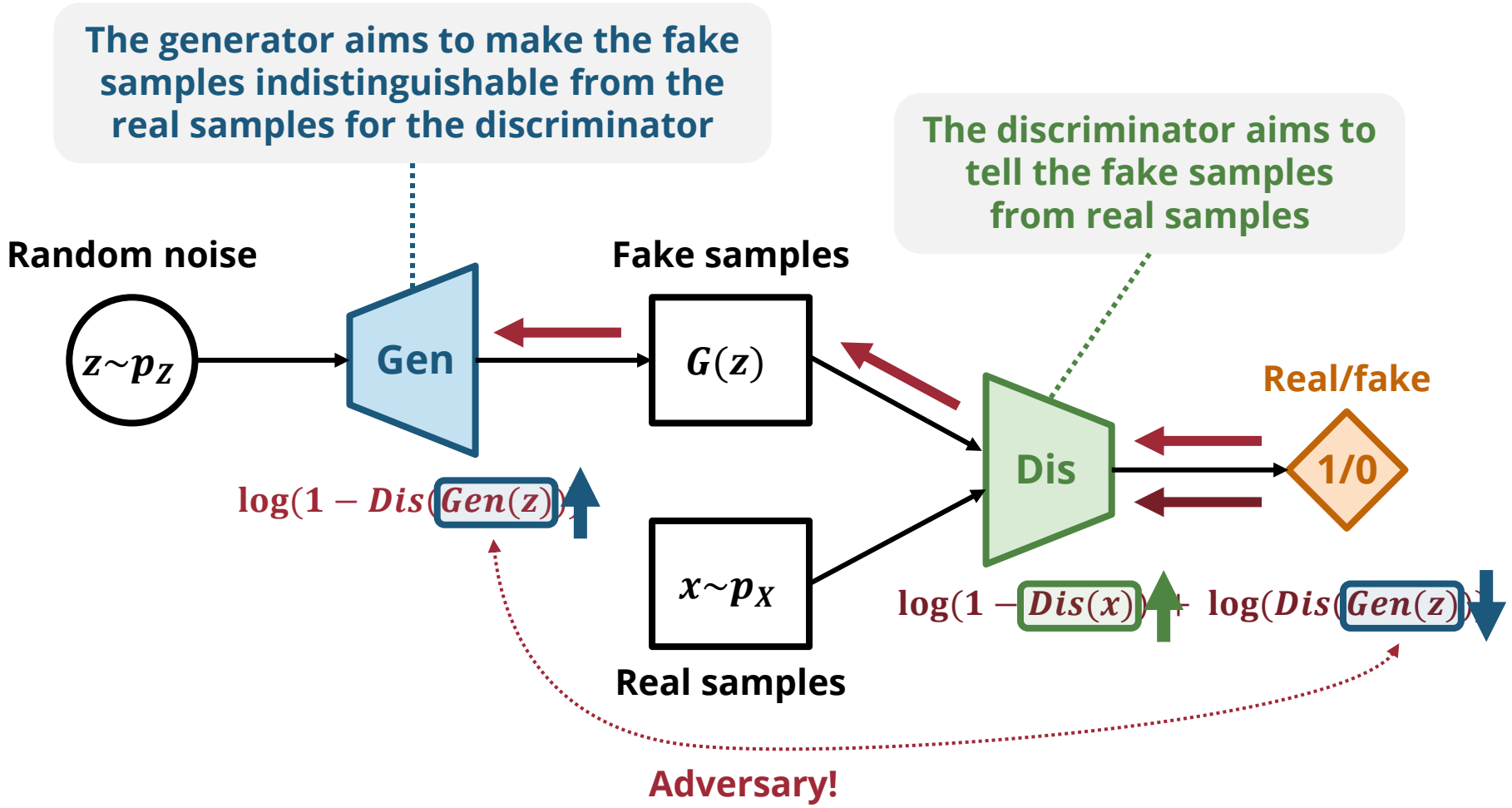
25x1 kernels

(Source: Donahue et al., 2019)

Example of generated music

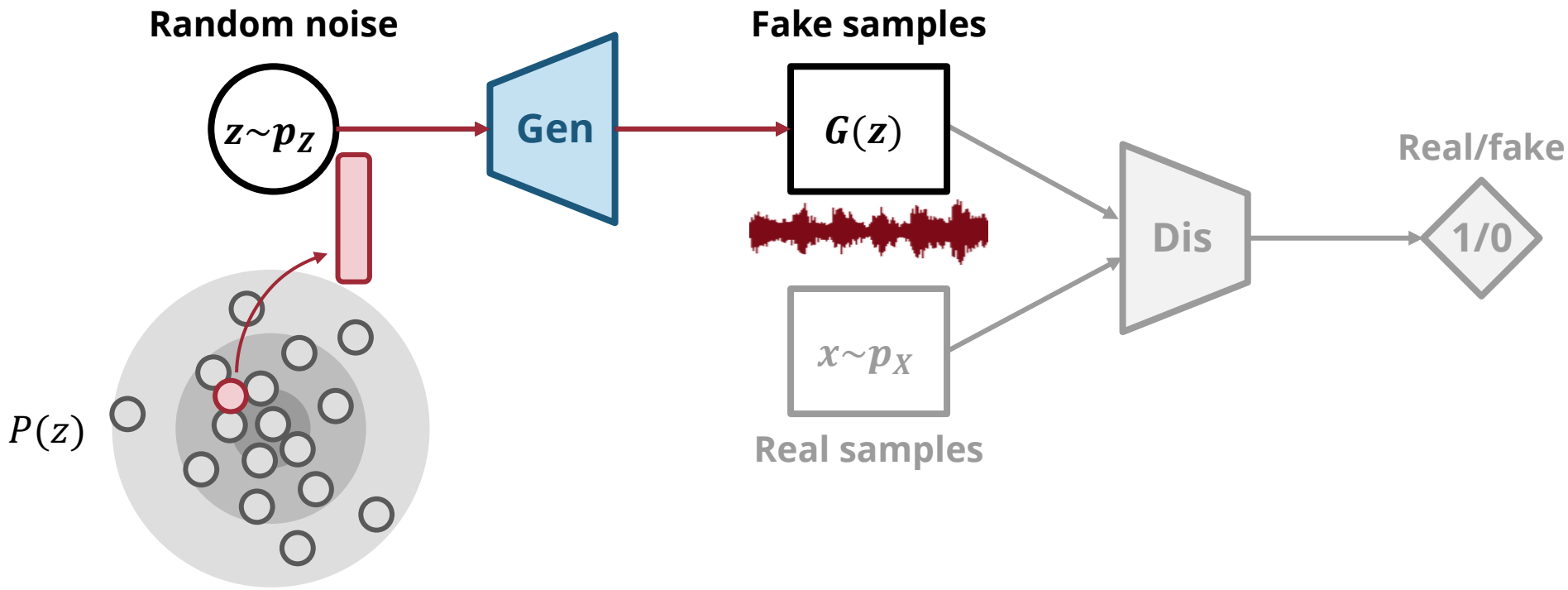
Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

(Recap) Generative Adversarial Nets (GANs) – Training



Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

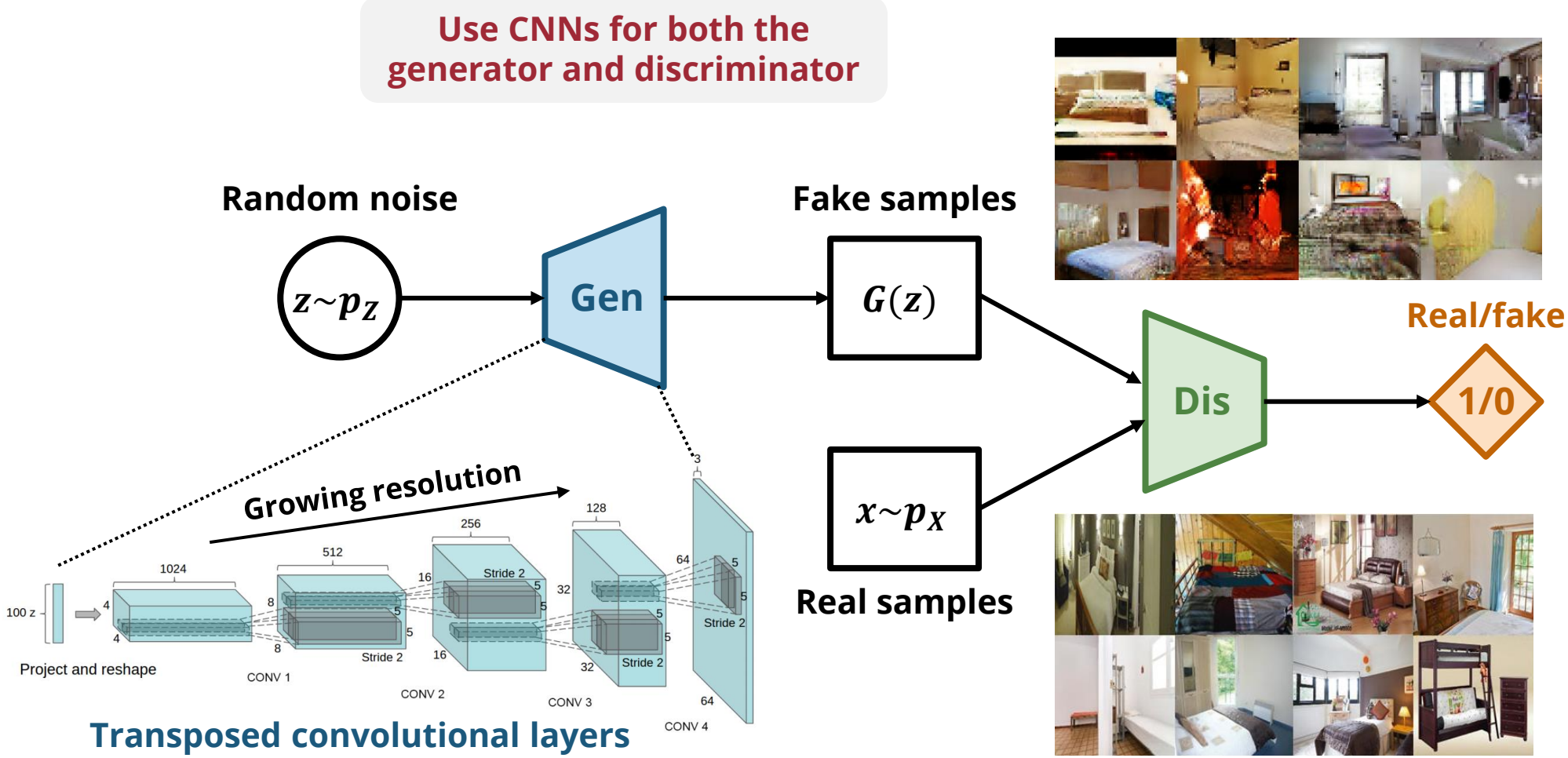
(Recap) Generative Adversarial Nets (GANs) – Generation



Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

(Recap) Deep Convolutional GANs (DCGANs)

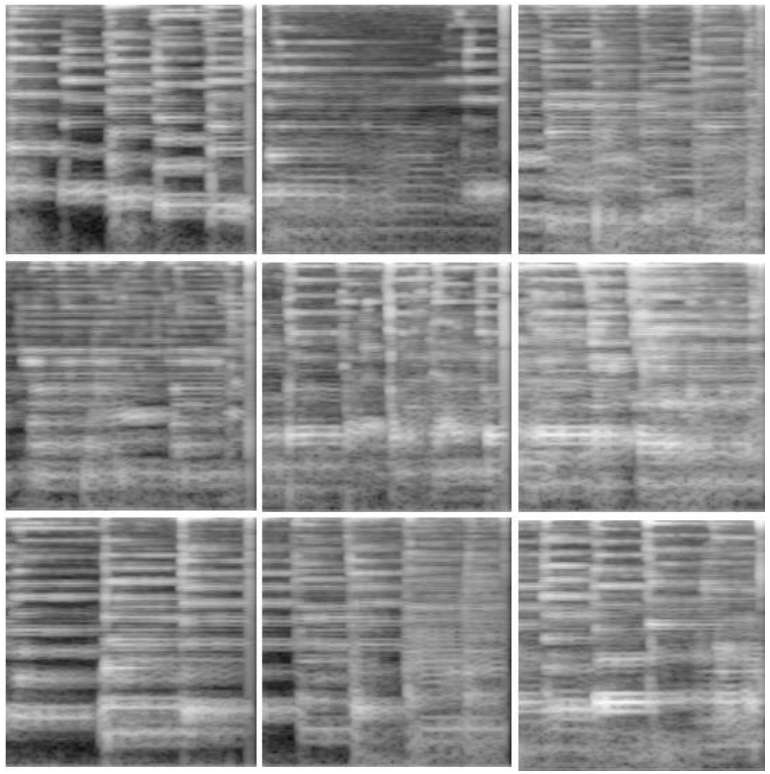
Use CNNs for both the generator and discriminator



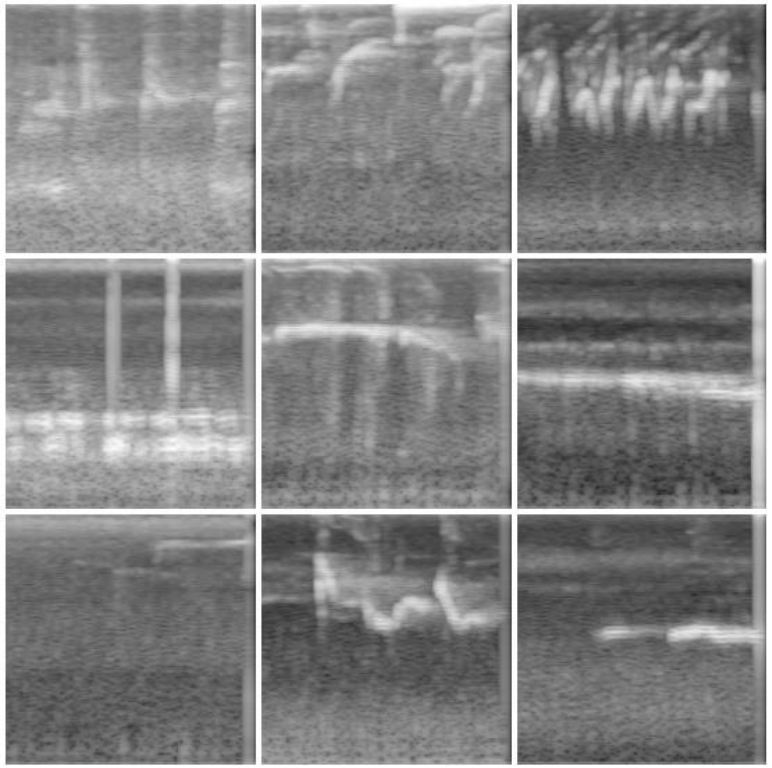
Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

Example: SpecGAN (Donahue et al., 2019)

Bird or piano sounds?



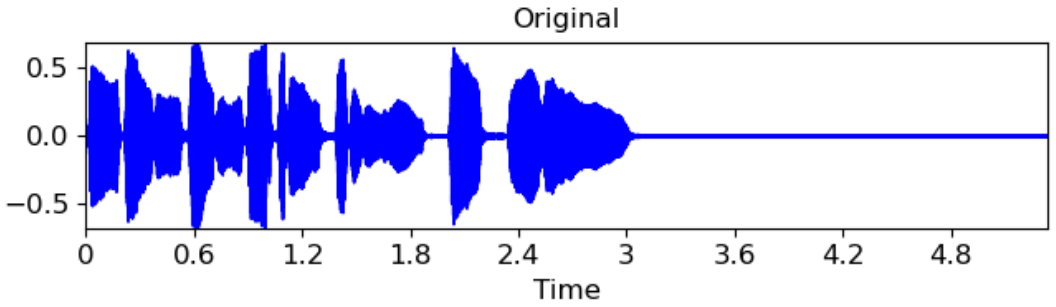
Bird or piano sounds?

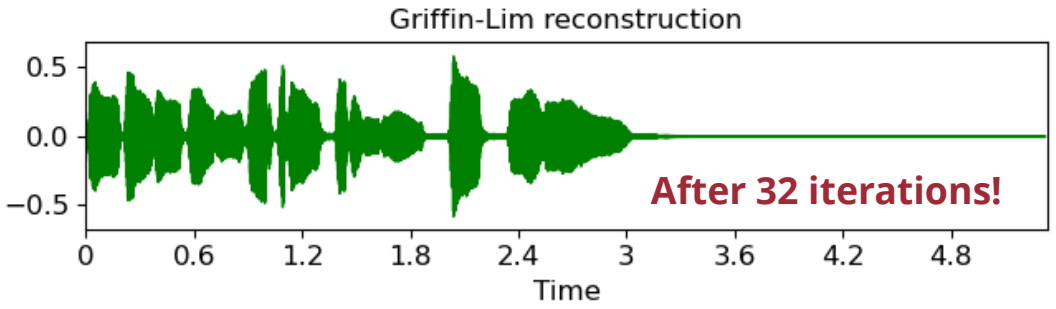


(Source: Donahue et al., 2019)

Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

Griffin-Lim Algorithm (Griffin & Lim, 1984)





(Source: librosa documentation)

Given a magnitude-only STFT matrix

↓

Randomly initialize the phase

↓

$$y' = \arg\min_{y} (M - \mathrm{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

↓

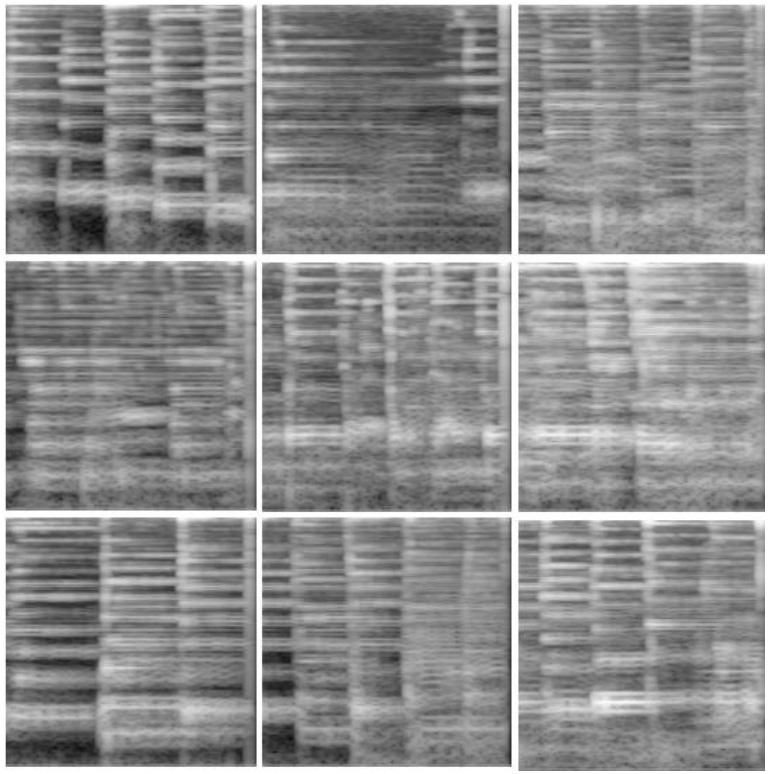
$$M' = \mathrm{STFT}(y')$$

Find the STFT of the signal y

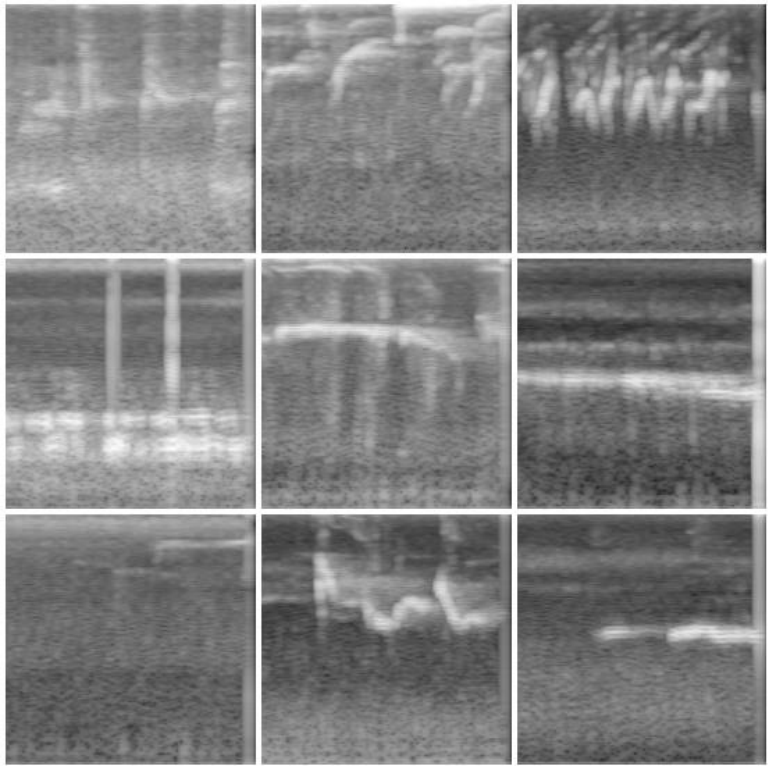
Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, 1984.
librosa.org/doc/main/generated/librosa.griffinlim.html

Example: SpecGAN (Donahue et al., 2019)

Piano sounds



Bird sounds

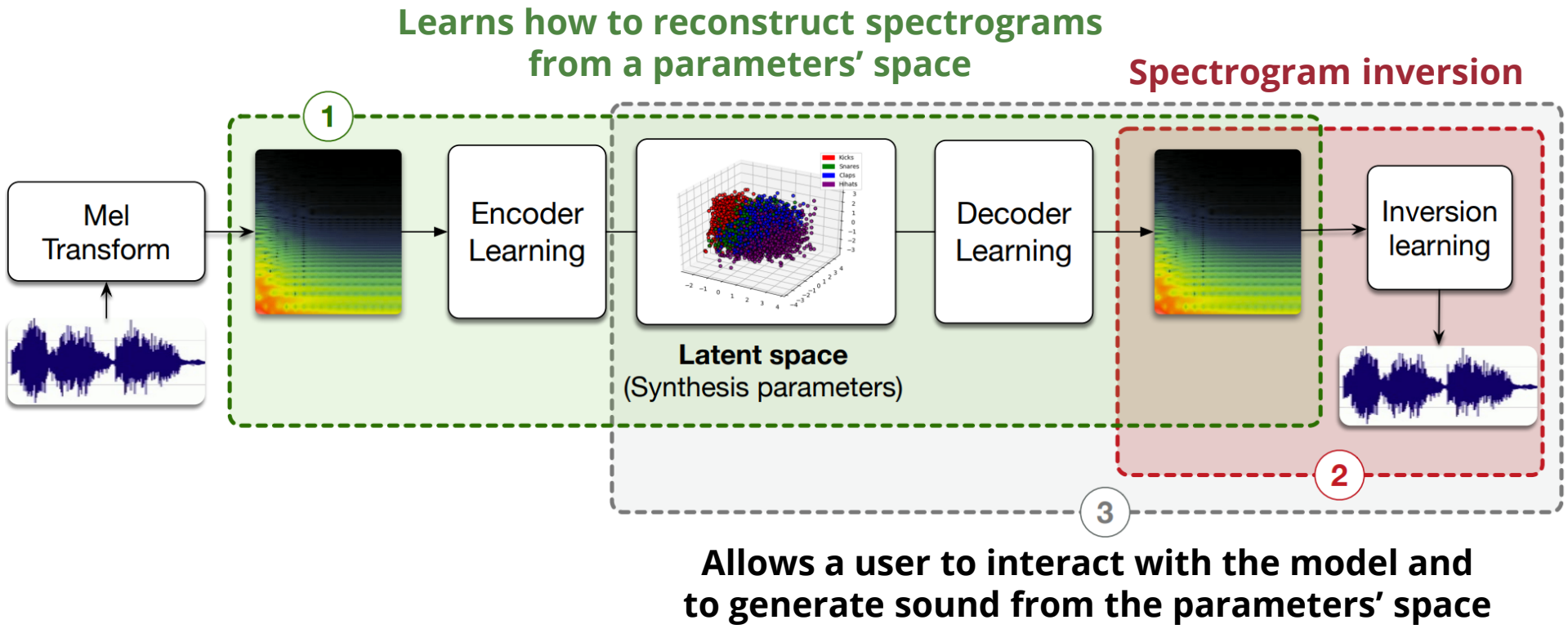


Example of generated music

(Source: Donahue et al., 2019)

Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

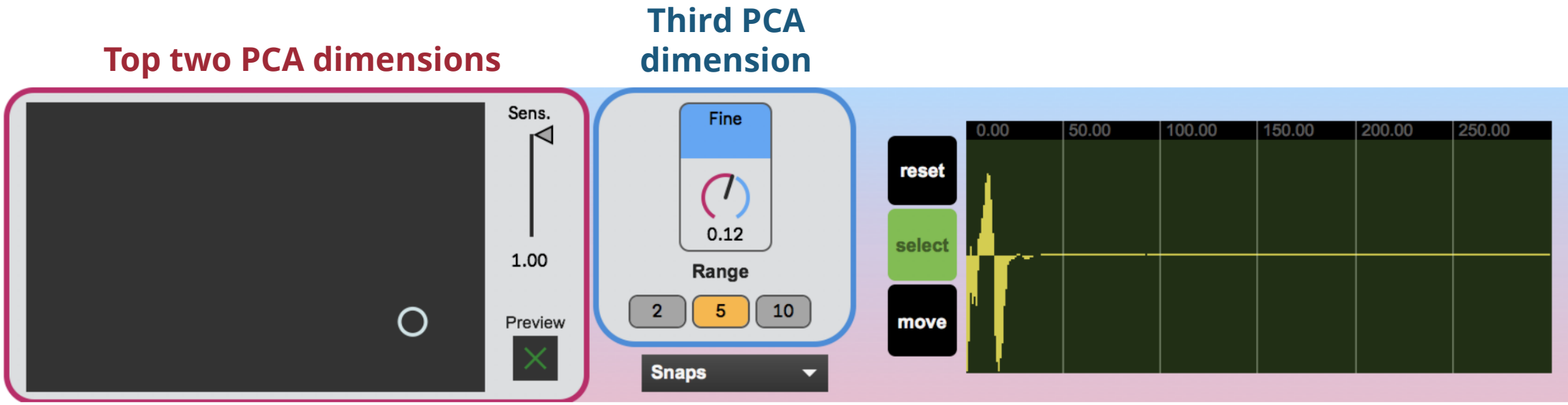
Example: Neural Drum Machine (Aouameur et al., 2019)



(Source: Aouameur et al., 2019)

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

Example: Neural Drum Machine (Aouameur et al., 2019)



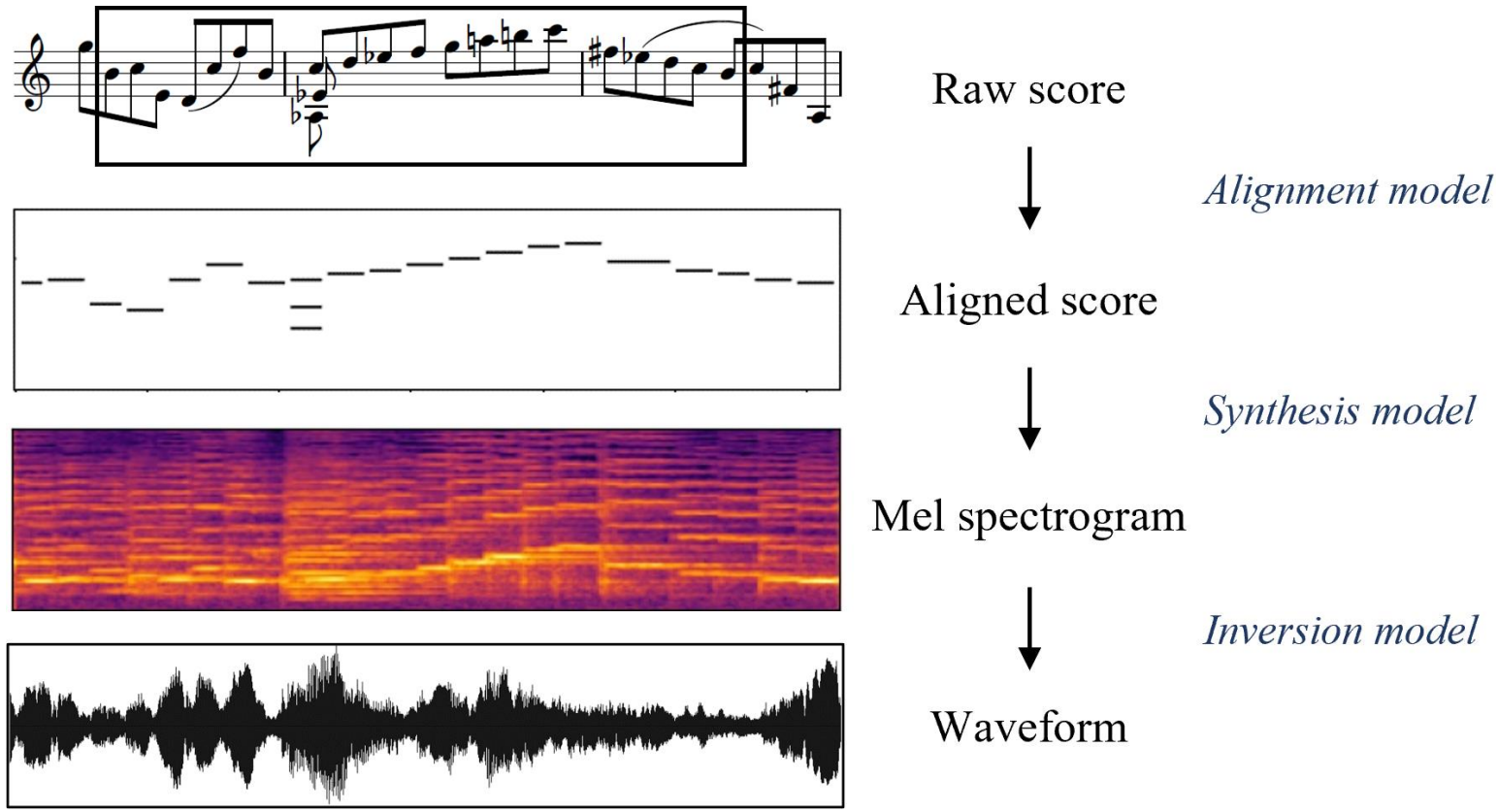
(Source: Aouameur et al., 2019)

drive.google.com/file/d/1DDo0_KnwkWirCM4t0PT8cp6uotsfuufj/view

Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," *ICCC*, 2019.

Score-to-Audio Synthesis

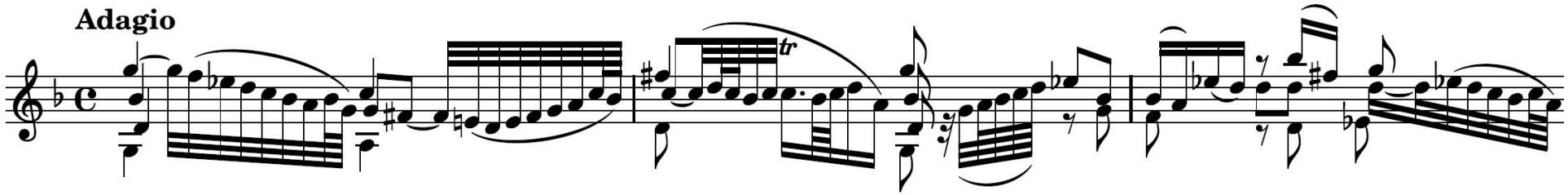
Example: DeepPerformer (Dong et al., 2022)

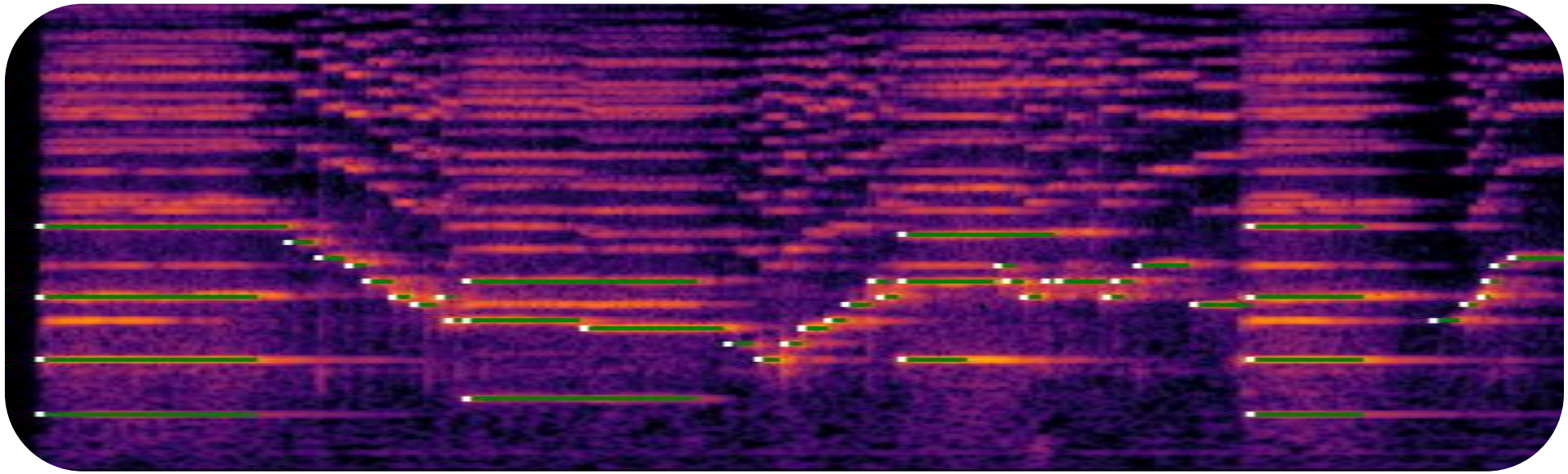


(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

Example: DeepPerformer (Dong et al., 2022)





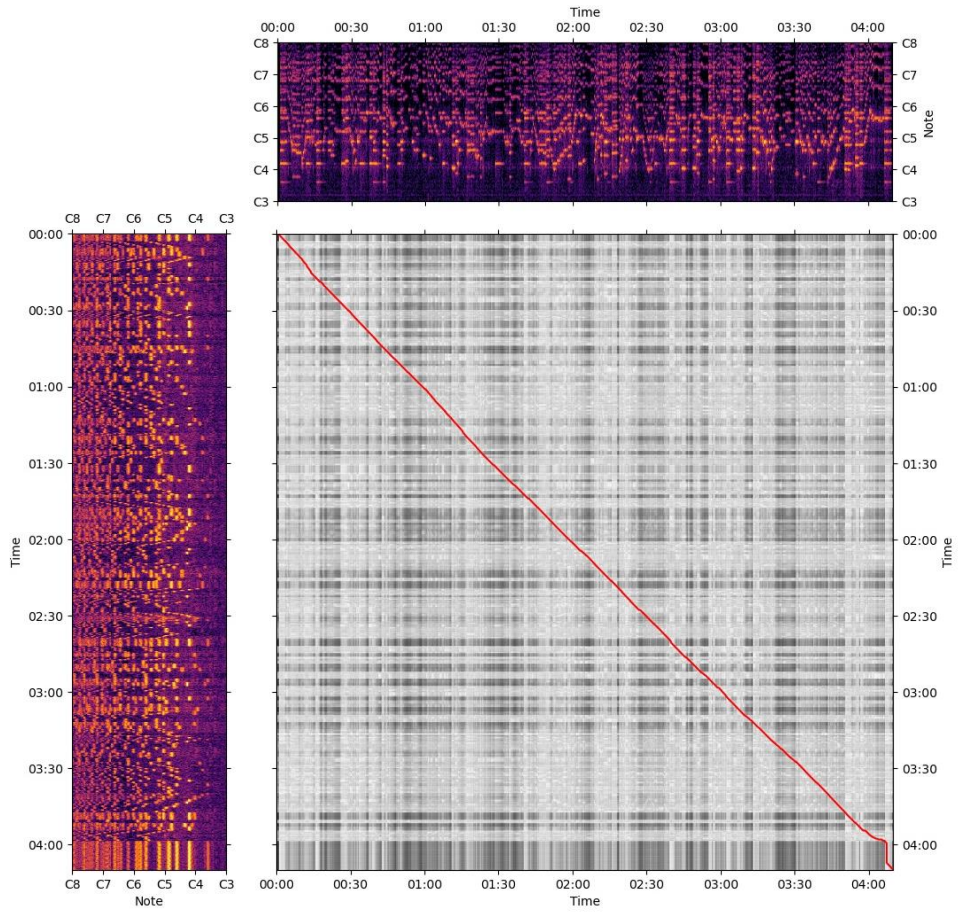
(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

Example: DeepPerformer (Dong et al., 2022)

Performance recording

Synthesized audio

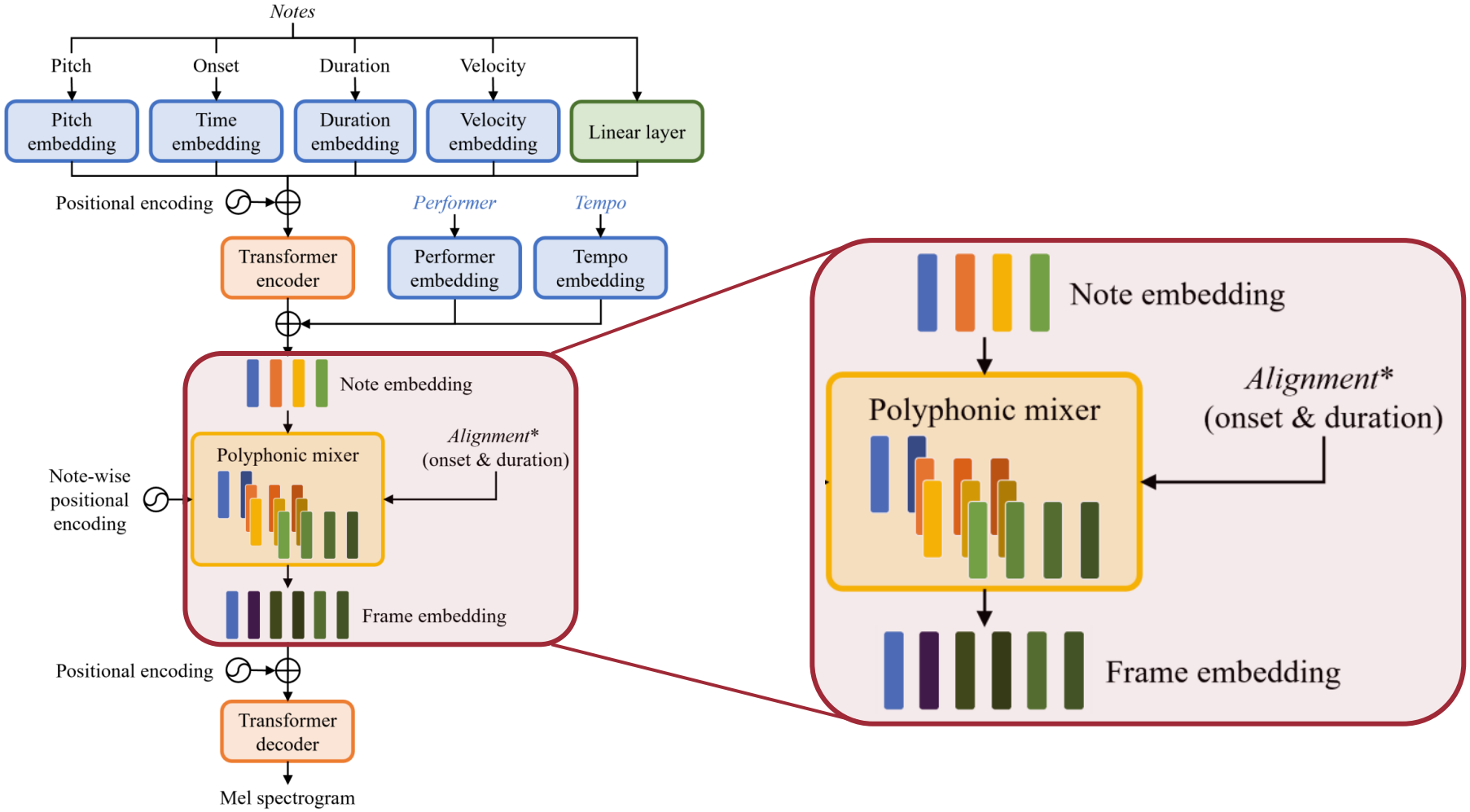


Dynamic time warping!

(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

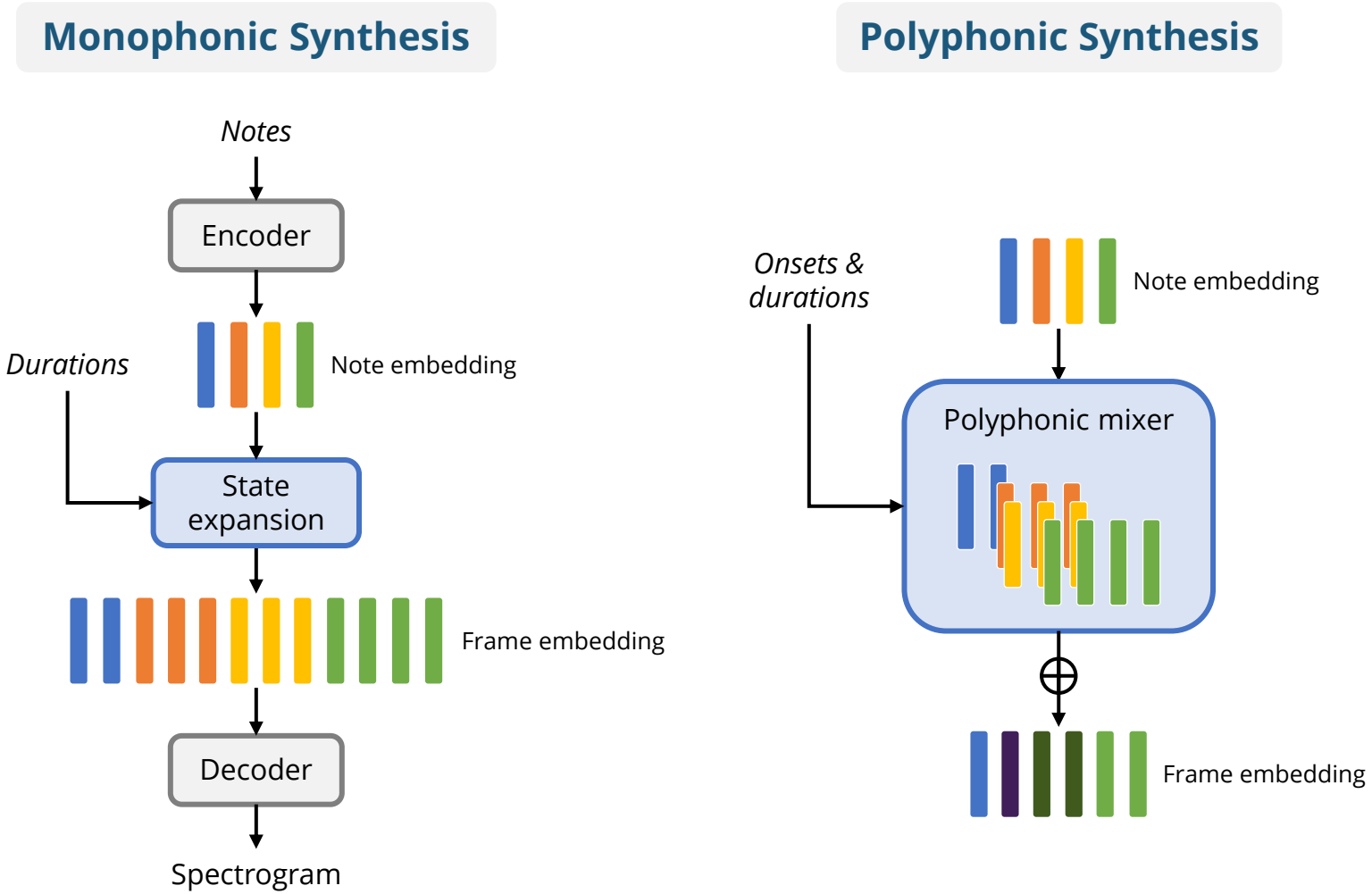
Example: DeepPerformer (Dong et al., 2022)



(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

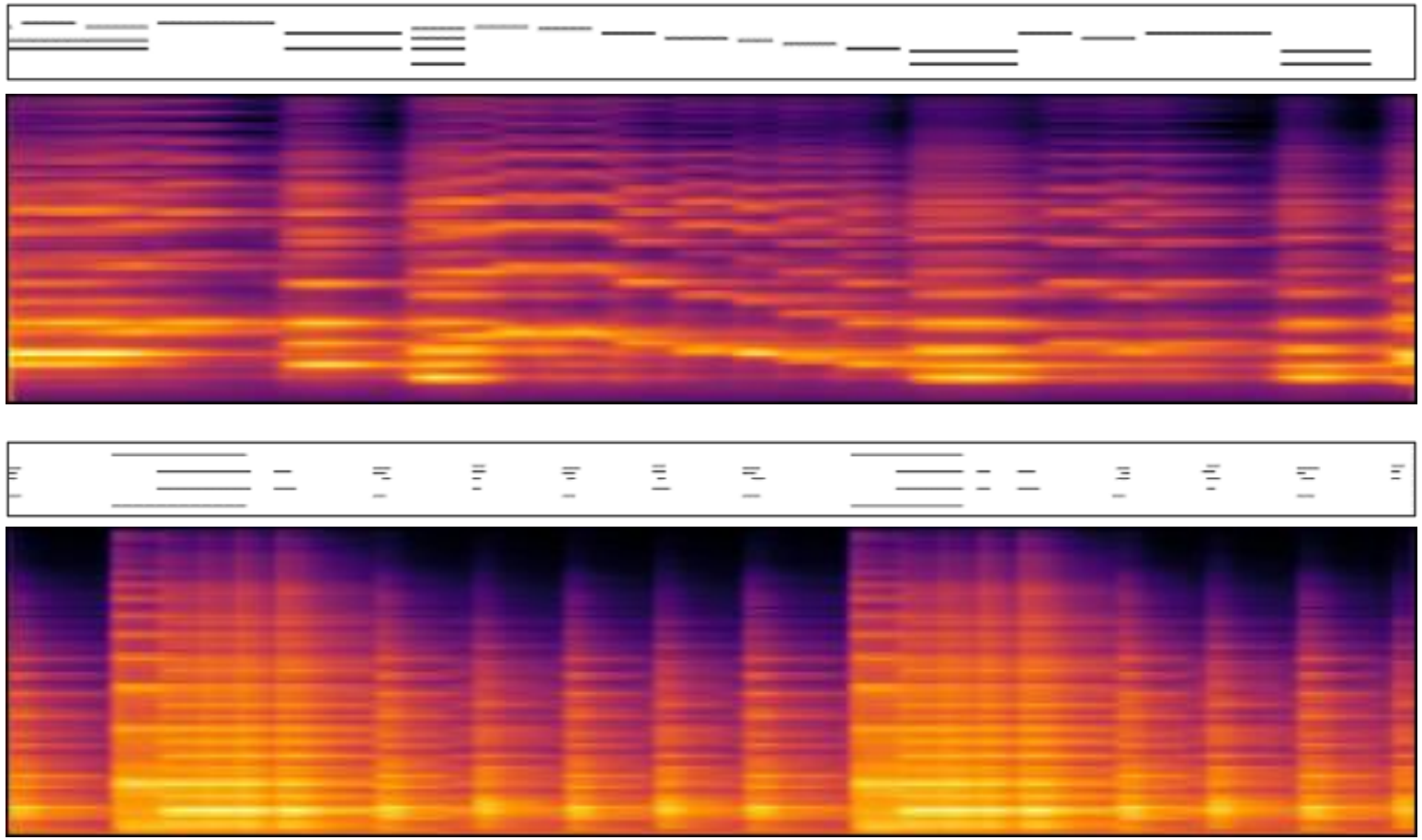
Example: DeepPerformer (Dong et al., 2022)



(Source: Dong et al., 2022)

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

Example: DeepPerformer (Dong et al., 2022)



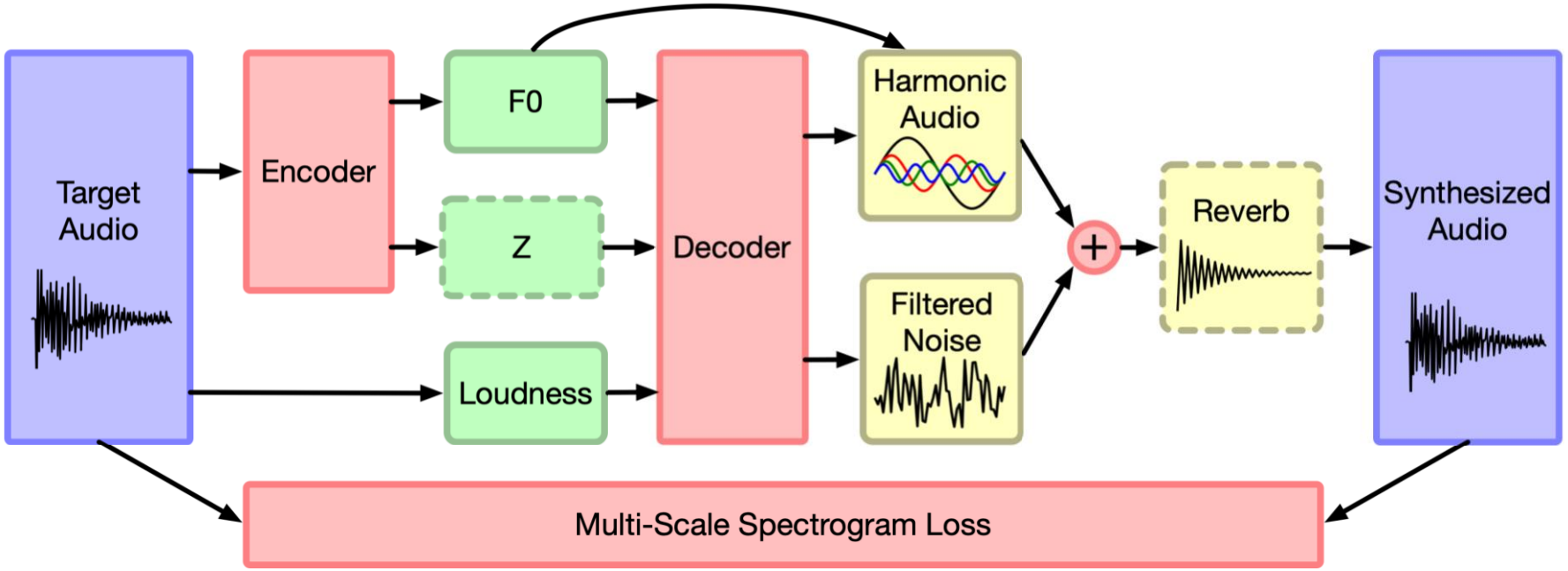
(Source: Dong et al., 2022)

hermandong.com/deepperformer

Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

Differentiable DSP

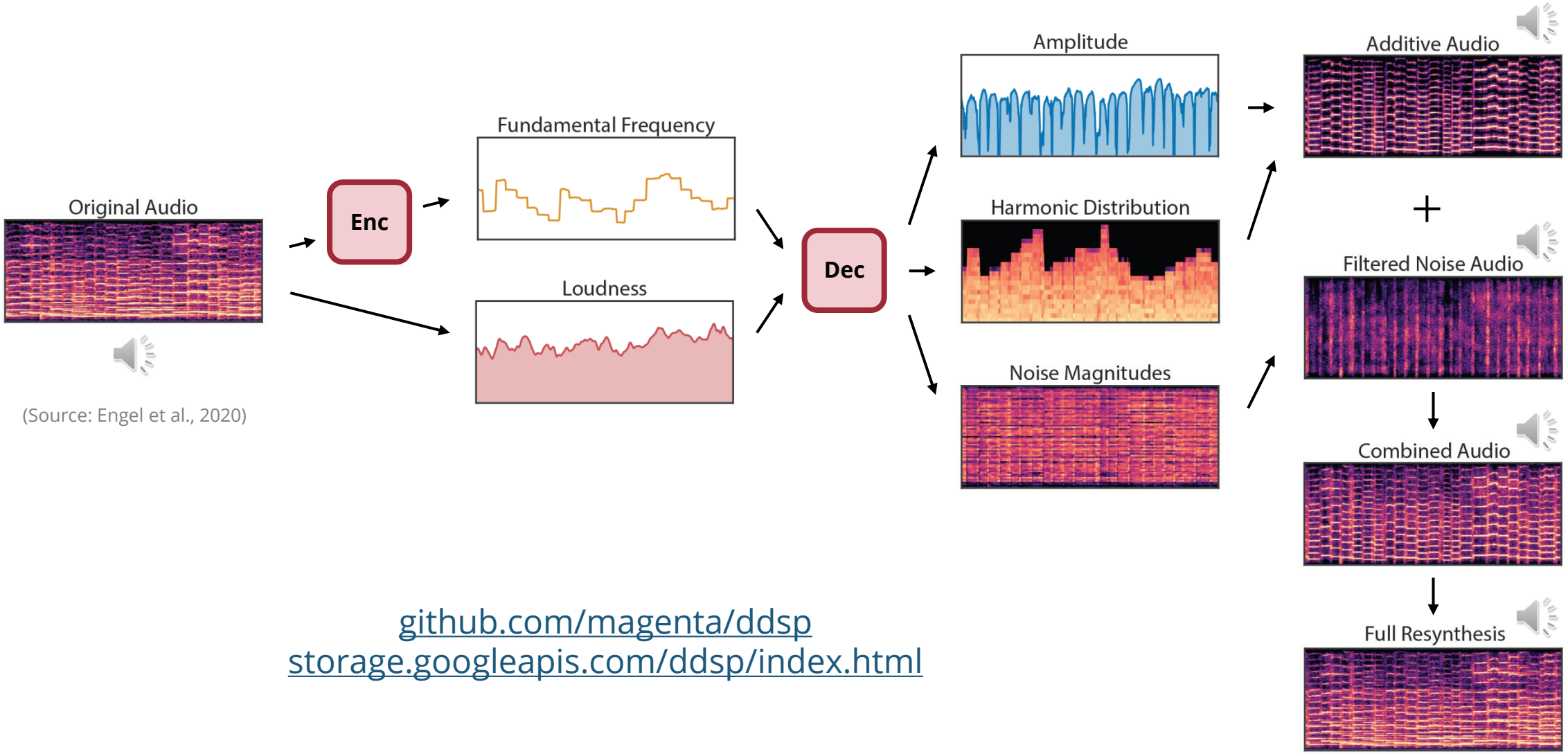
Example: Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Example: Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

Yaboi Hanoi – Entering Demons & Gods (2022)

youtu.be/PbrRoR3nEVw

soundcloud.com/yaboihanoi/enter-demons-and-gods