PAT 463/563 (Fall 2025)

Music & AI

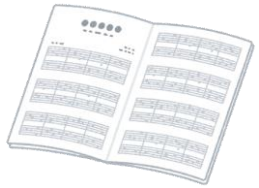
Lecture 14: Audio-domain Music Generation

Instructor: Hao-Wen Dong

UNIVERSITY OF MICHIGAN

Autoregressive Waveform Synthesis

Four Paradigms of Music Generation



Symbolic music generation

Audio-domain music generation

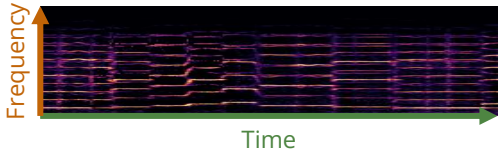
Text-based

Image-based

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

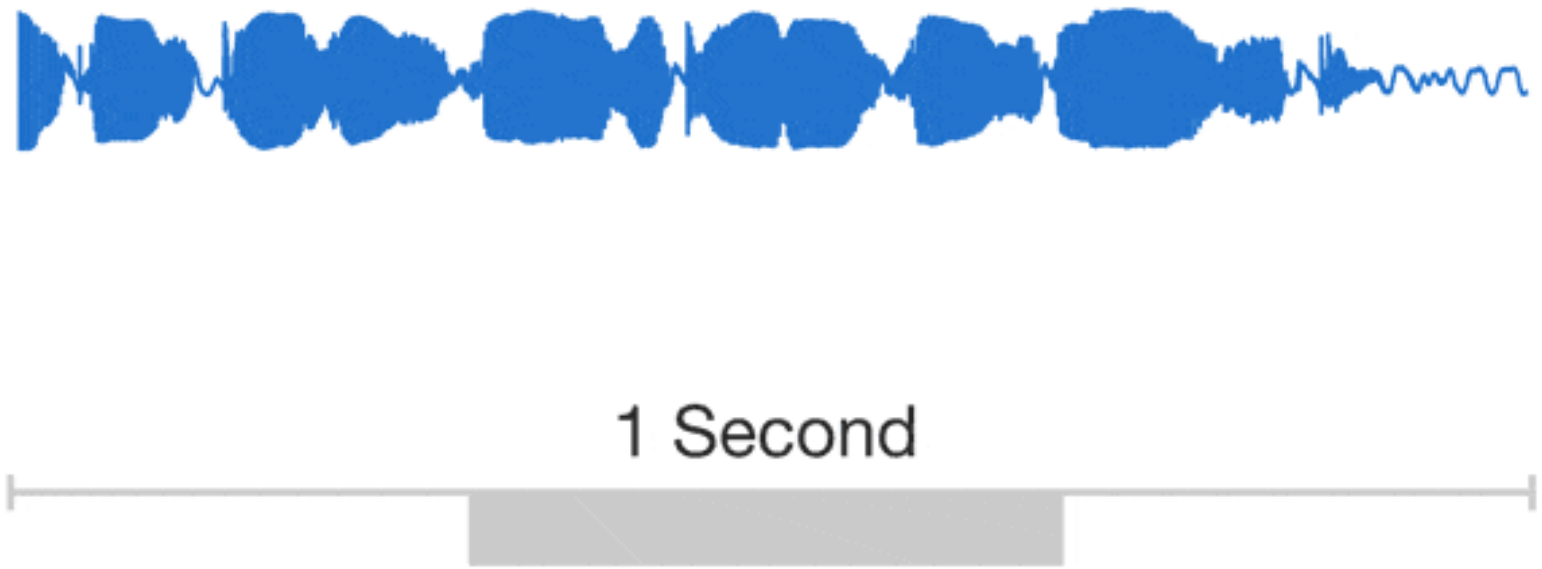
Piano roll

Waveform

Spectrogram

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

Generating Waveforms using a Neural Network

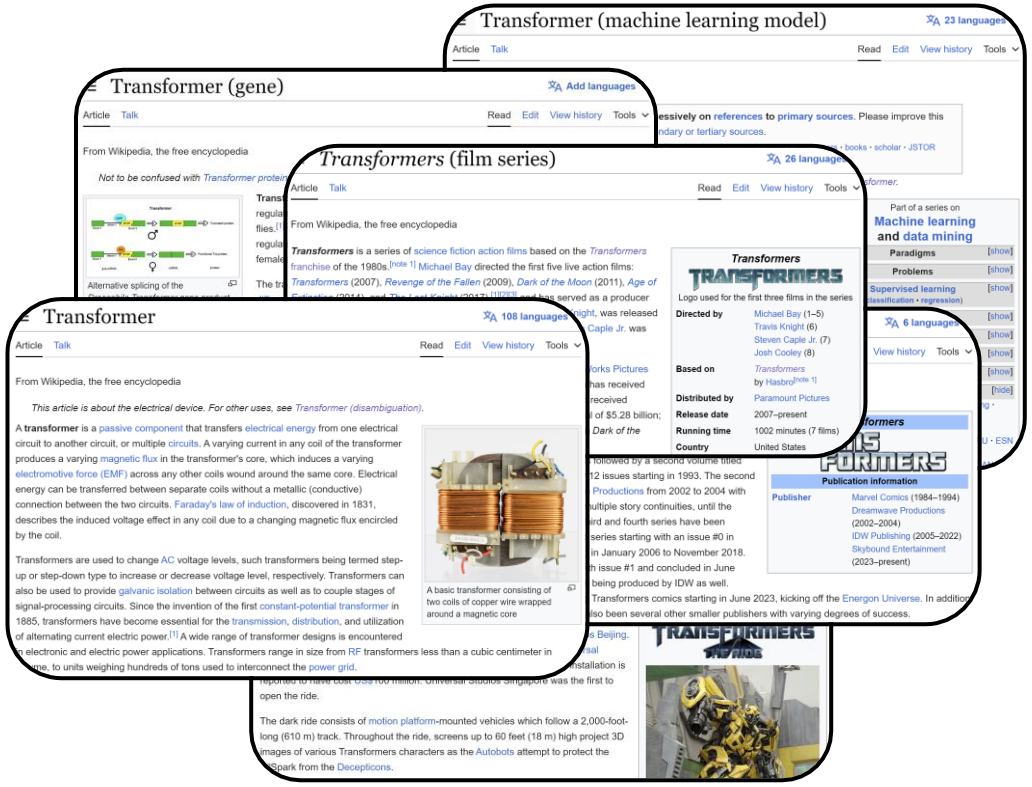


(Source: van den Oord et al., 2016)

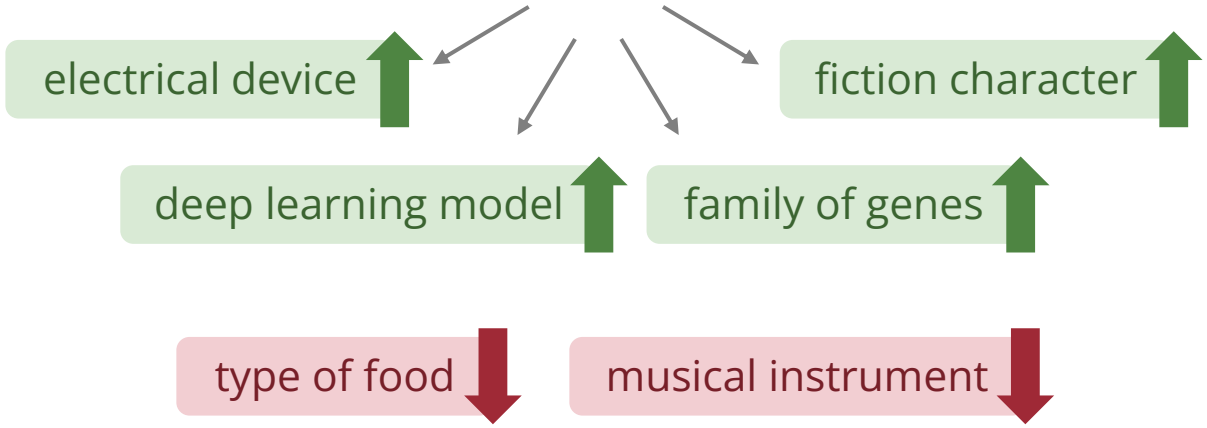
Aaron van den Oord and Sander Dieleman, "WaveNet: A generative model for raw audio," *DeepMind Blog*, September 8, 2016.

Language Models

- Predicting the next word given the past sequence of words



A transformer is a _____



Language Models (Mathematically)

- A class of machine learning models that learn the next word probability

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

Next word — Previous words

$P(\ \text{electrical} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{character} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{gene} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{model} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{food} \mid \text{A transformer is a}\ )$ ↓

$P(\ \text{musical} \mid \text{A transformer is a}\ )$ ↓

Autoregressive Models (Mathematically)

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

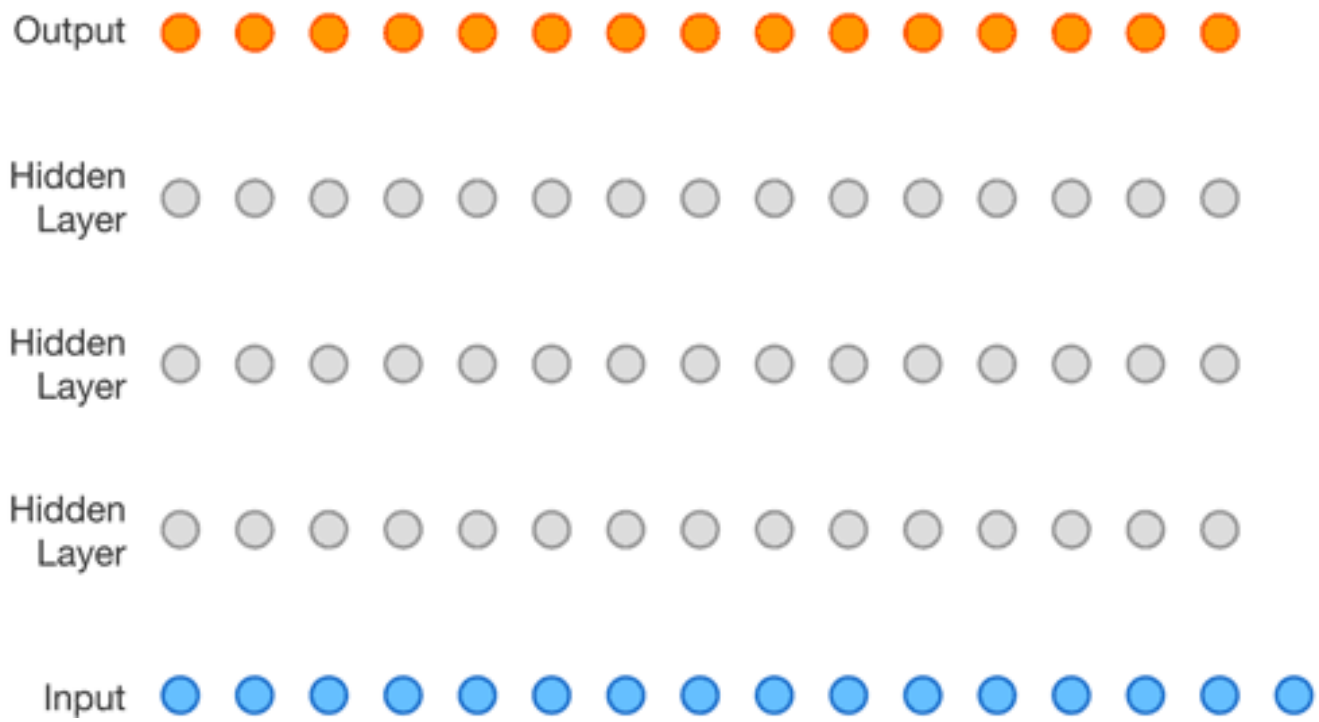
$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

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.

WaveNet (van den Oord et al., 2016)



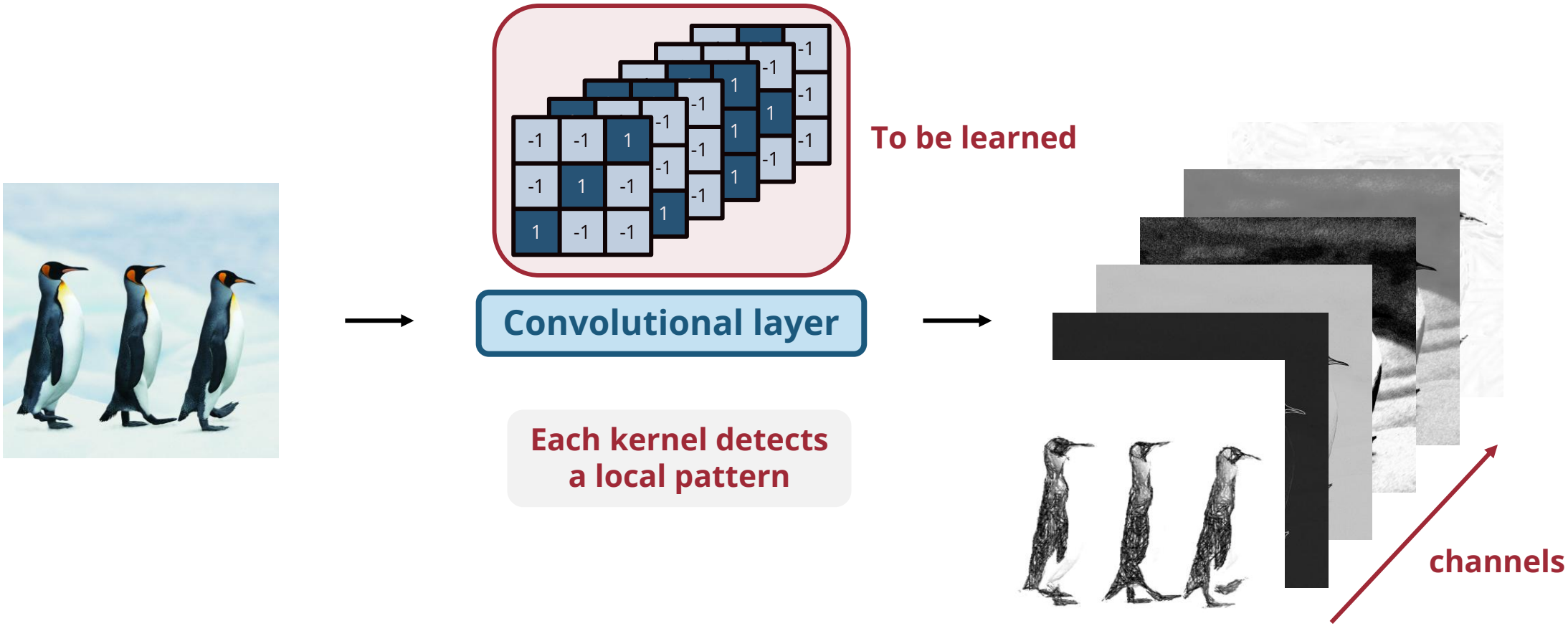
(Source: van den Oord et al., 2016)

A convolutional neural network for raw waveform generation

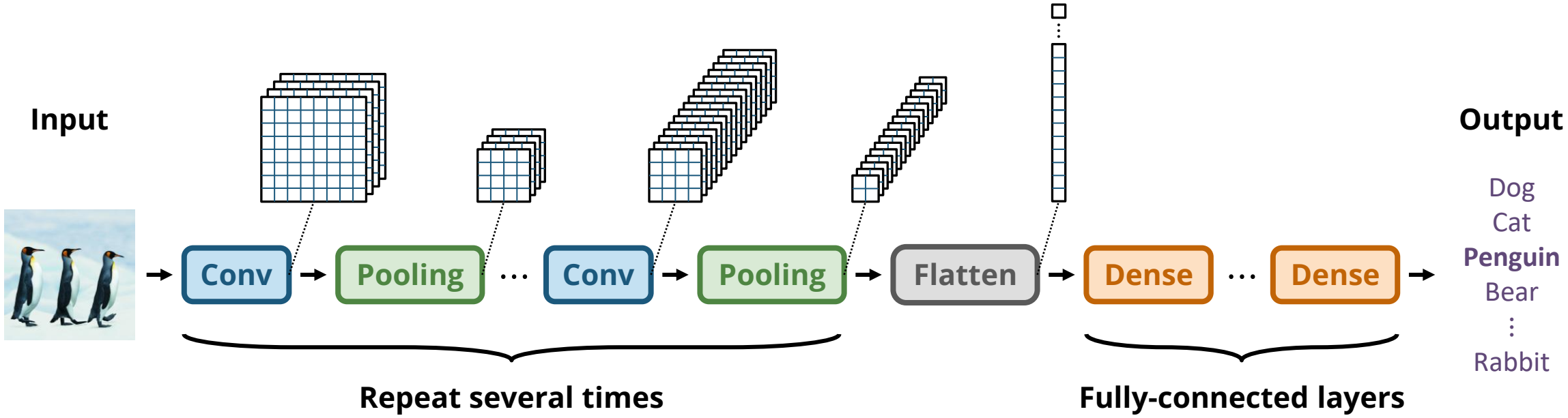
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

Convolutional Layer

- A convolutional layer consists of many **learnable kernels** (channels)

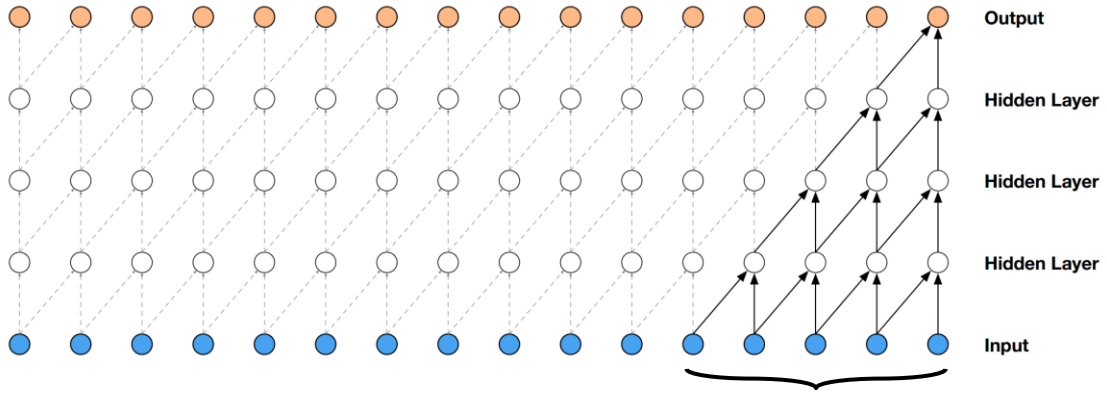


Convolutional Neural Network (CNNs)



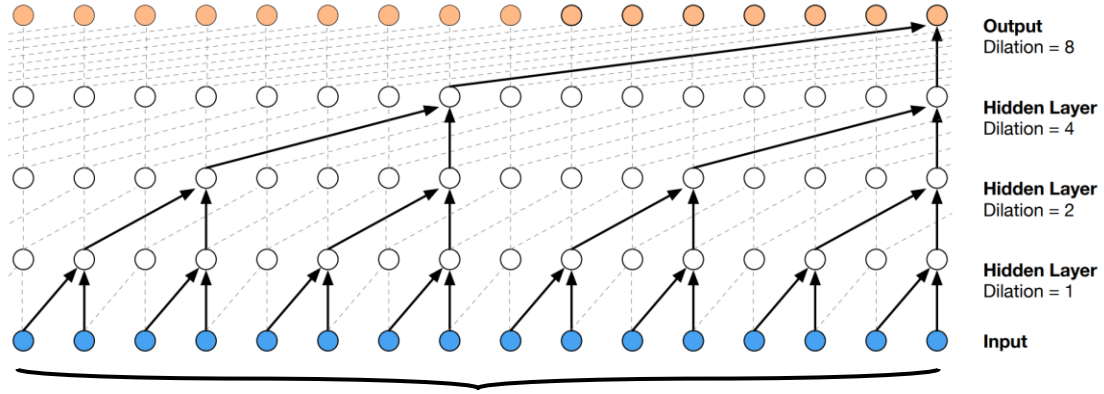
WaveNet (van den Oord et al., 2016)

Standard convolution



Smaller "receptive field"

Dilated convolution



Larger "receptive field"

deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio

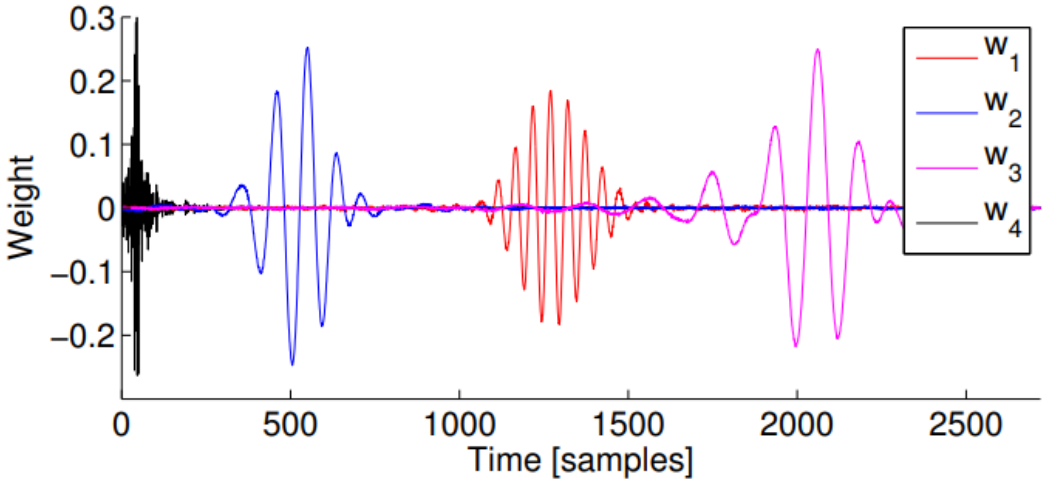
Example of generated music

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

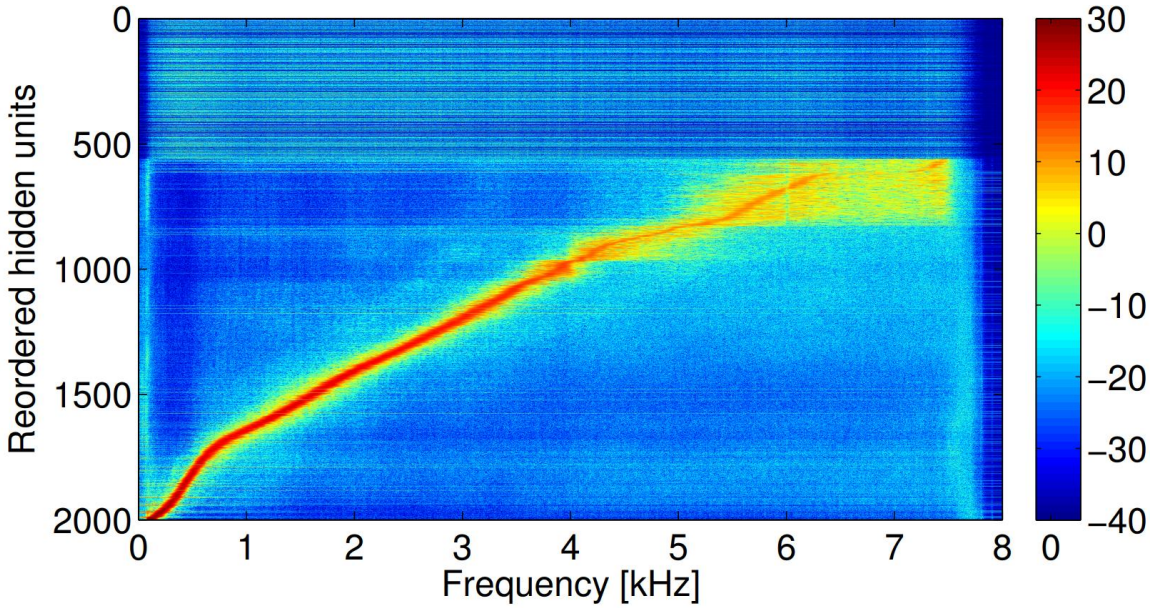
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

1D CNNs & Fourier Transform

Convolution kernels learned



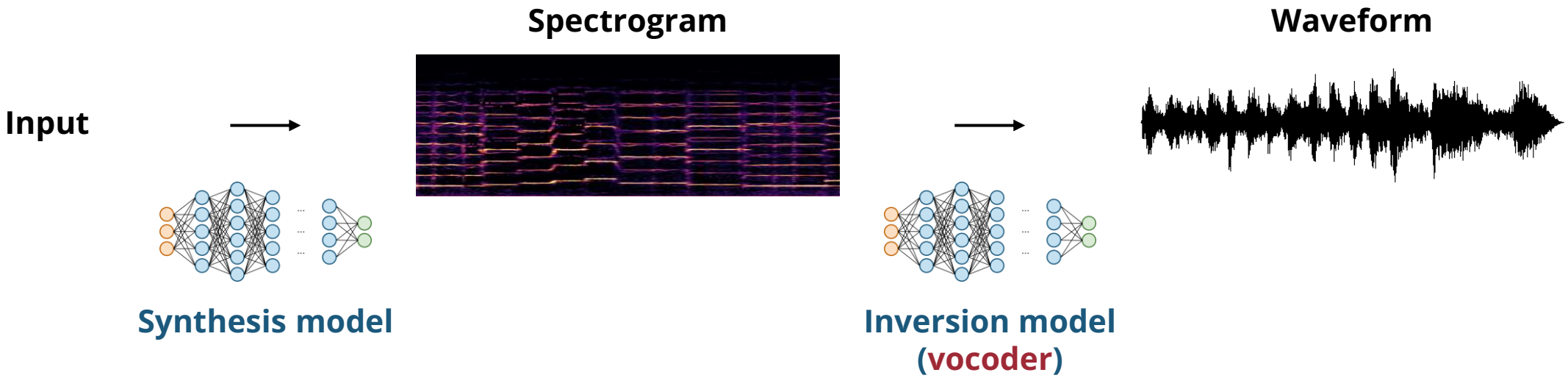
Peak frequency detected by the learned kernels



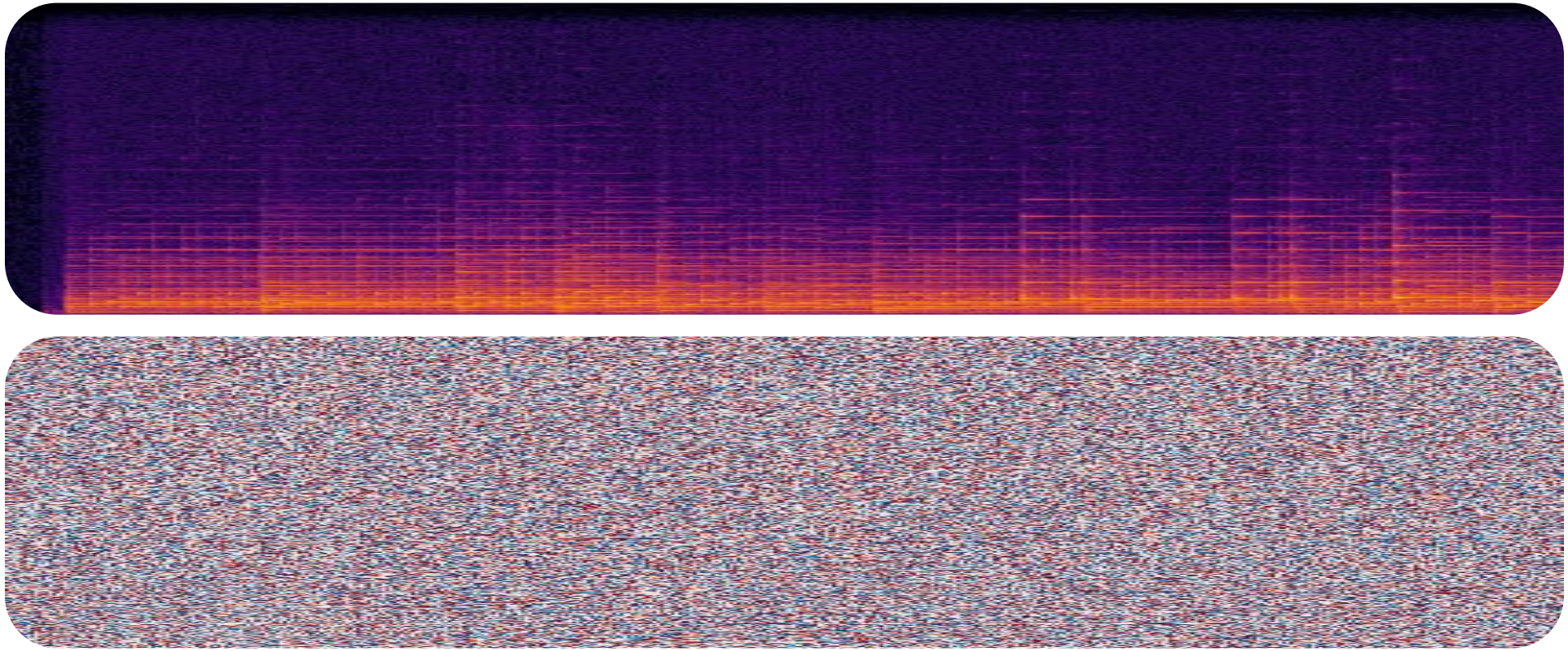
Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," *INTERSPEECH*, 2015.

Frequency-domain Audio Synthesis

Frequency-domain Audio Synthesis



Importance of the Phase Information

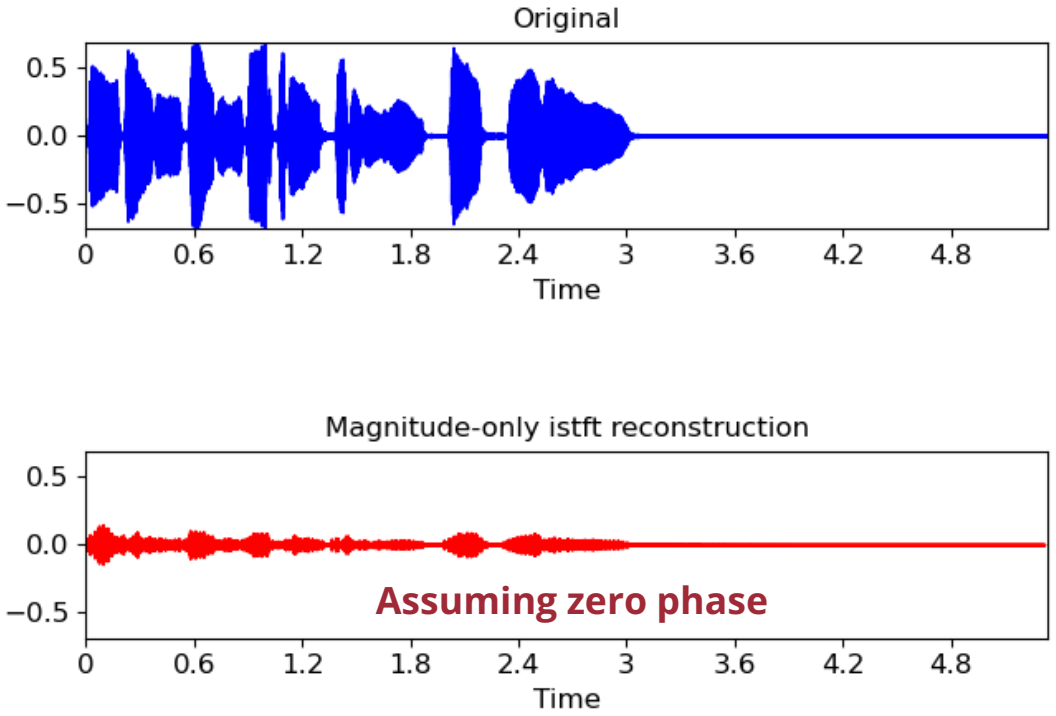


(Source: Dieleman et al., 2020)

Real phase **Random phase**

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

Inverse STFT without Phase Information



(Source: librosa documentation)

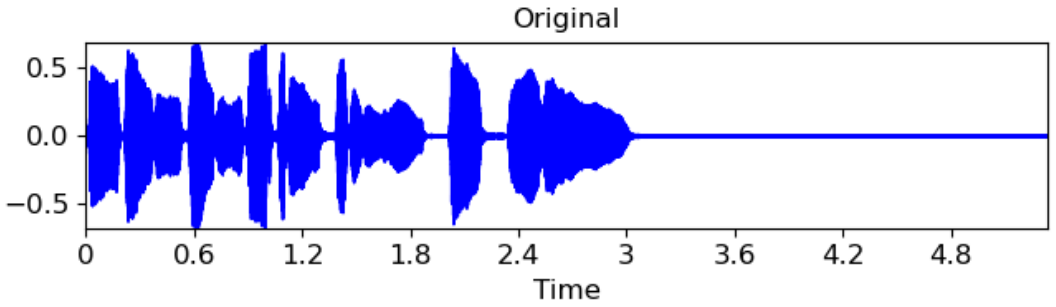
Complex-valued STFT matrix

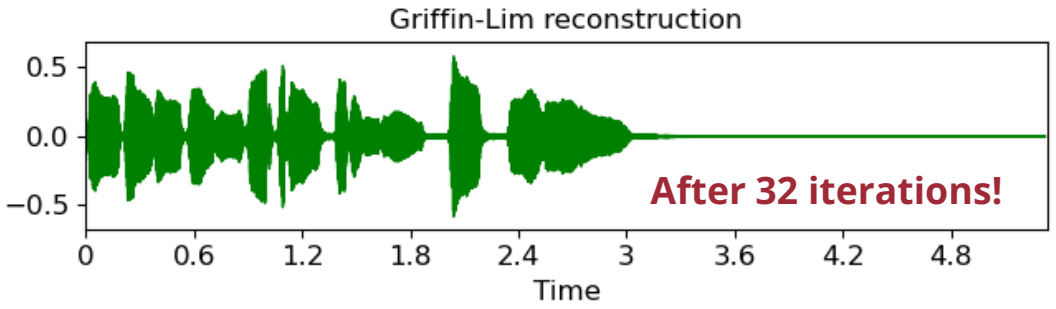
$$\text{ISTFT}(M) = \arg\min_{y} (M - \text{STFT}(y))^2$$

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)

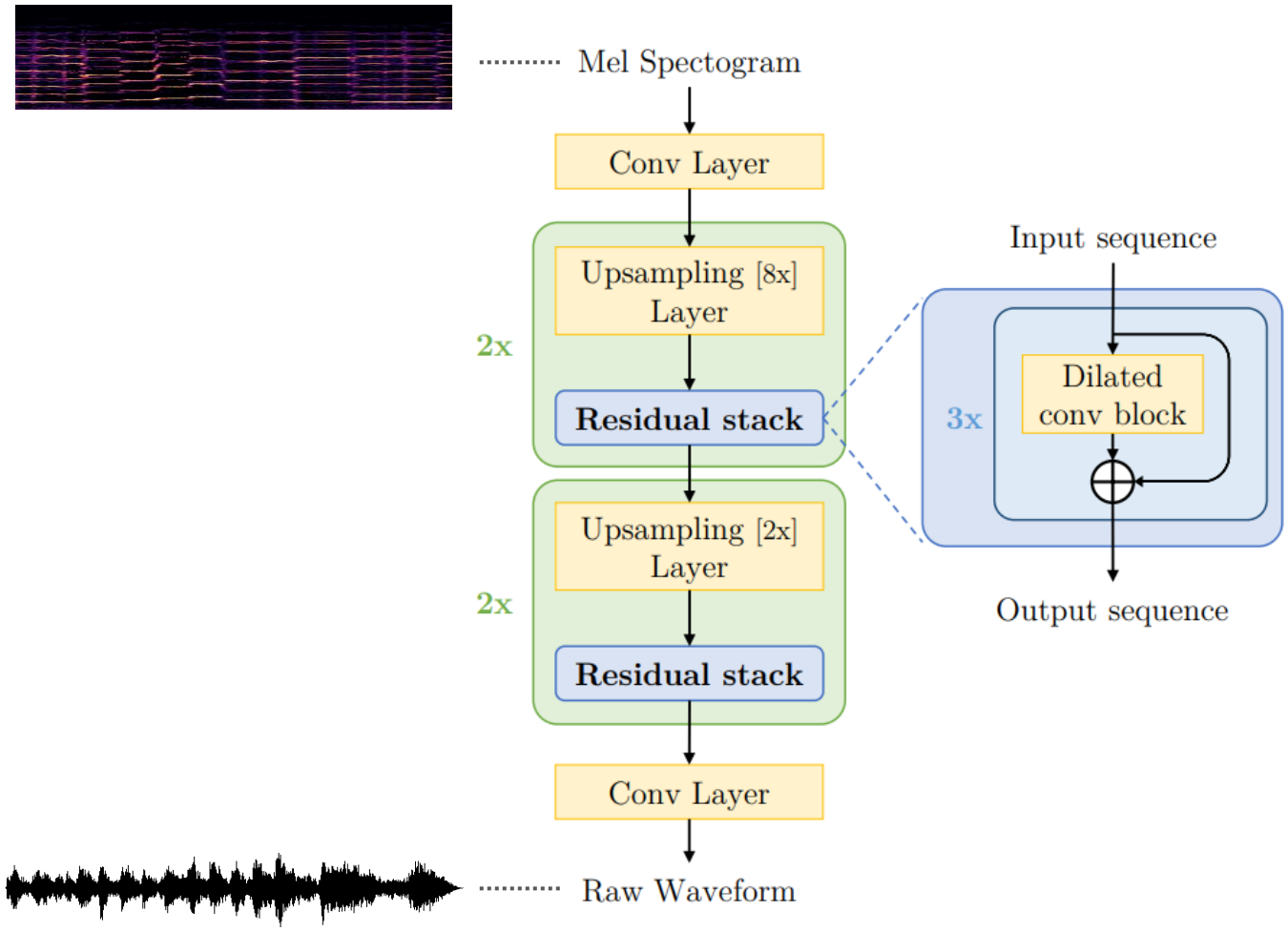
↓

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

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

MelGAN (Kumar et al., 2019)

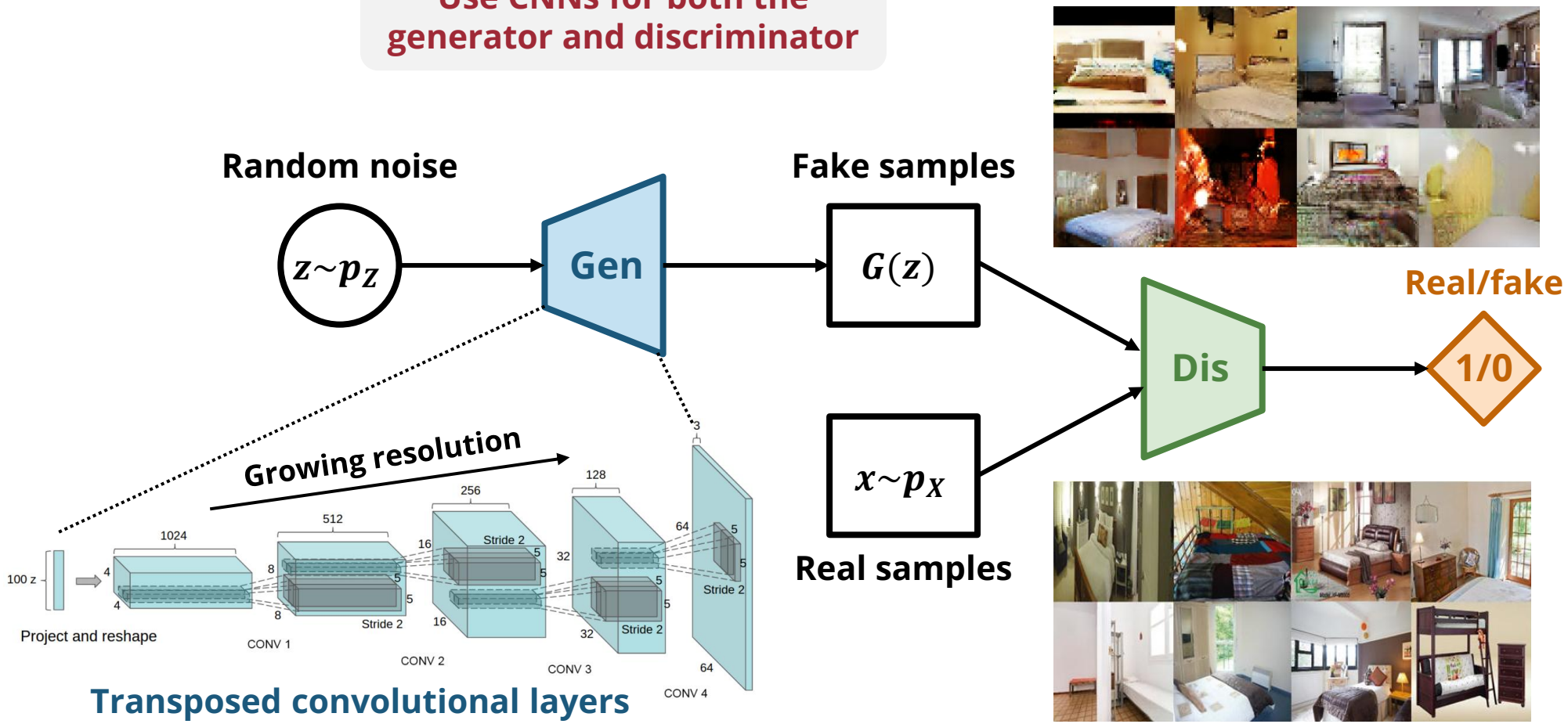


(Source: Kumar et al., 2019)

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.

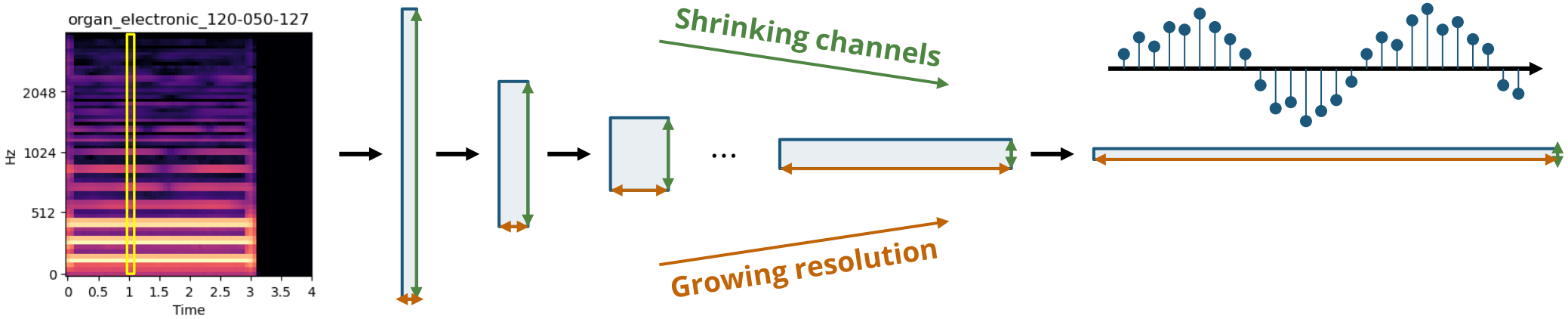
Deep Convolutional GANs (DCGANs) (Radford et al., 2014)

Use CNNs for both the generator and discriminator

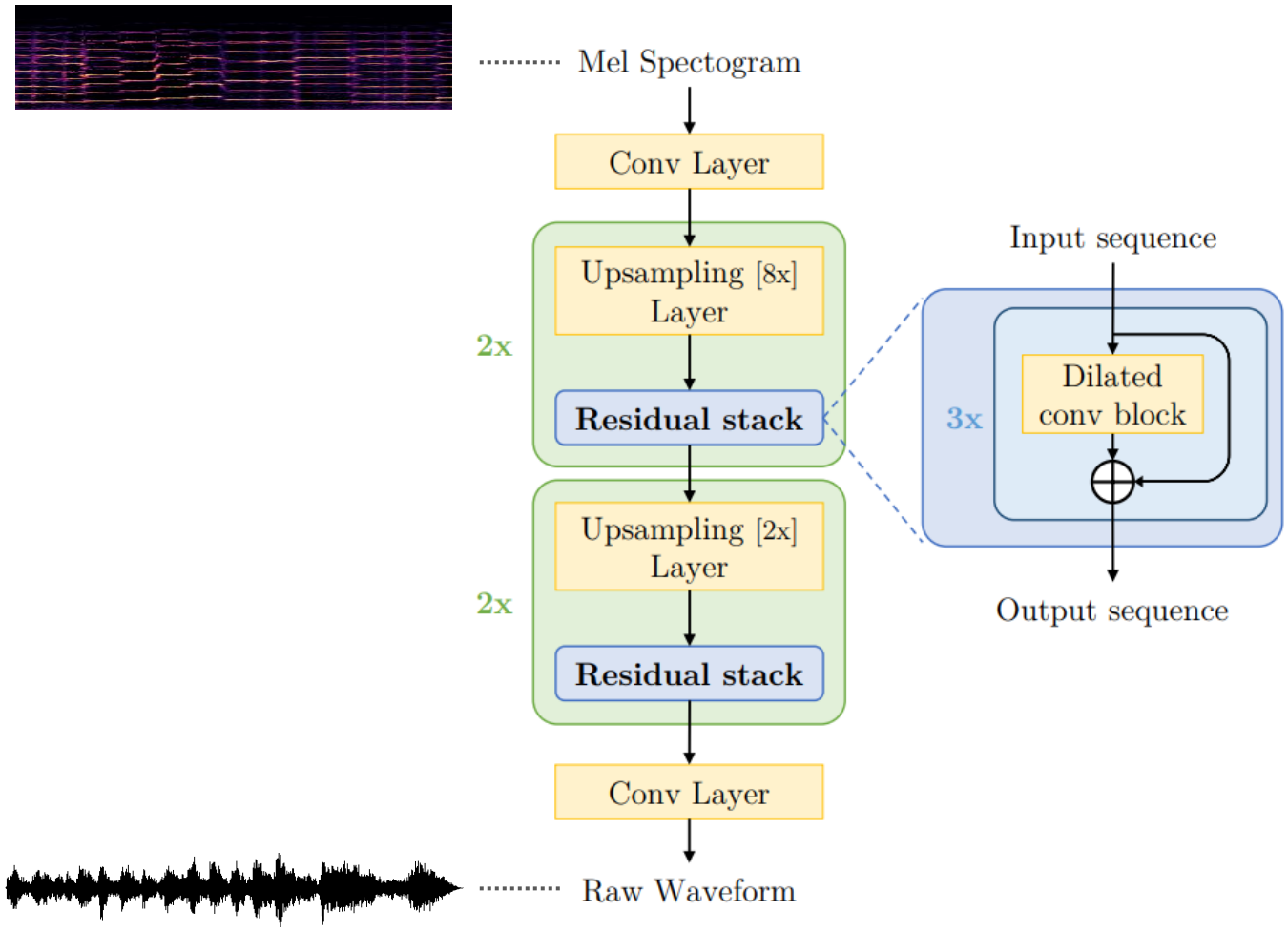


Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

Upsampling for Vocoders



MelGAN (Kumar et al., 2019)

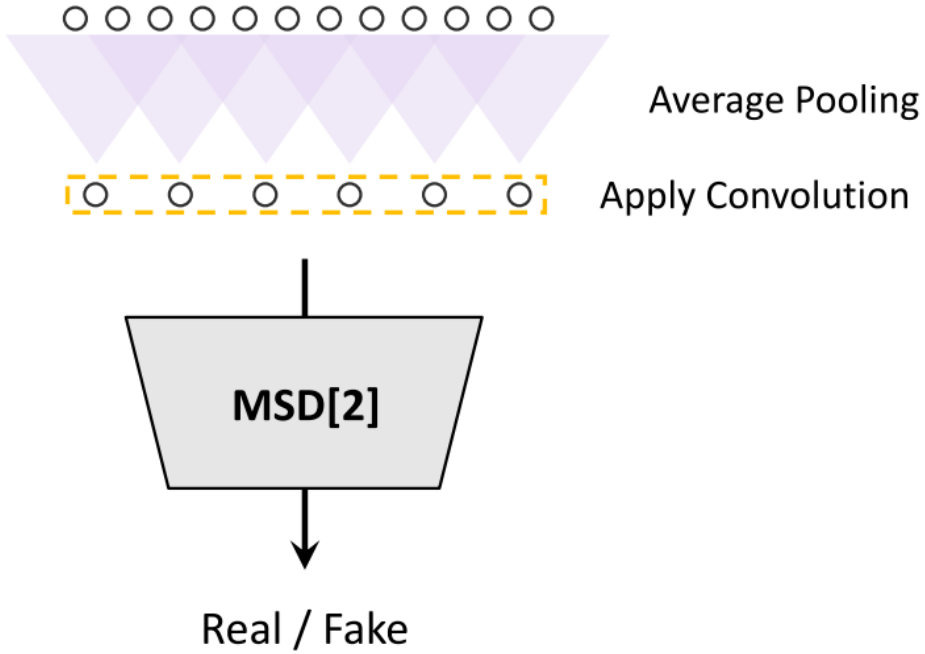


(Source: Kumar et al., 2019)

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.

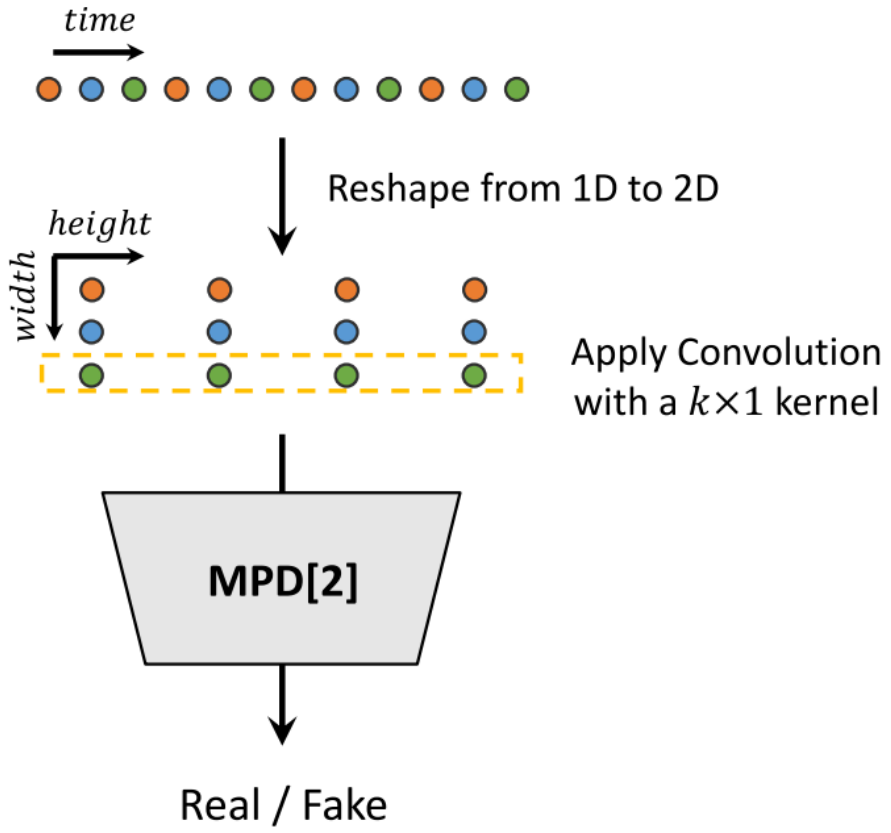
MelGAN (Kumar et al., 2019)

Multi-scale discriminator



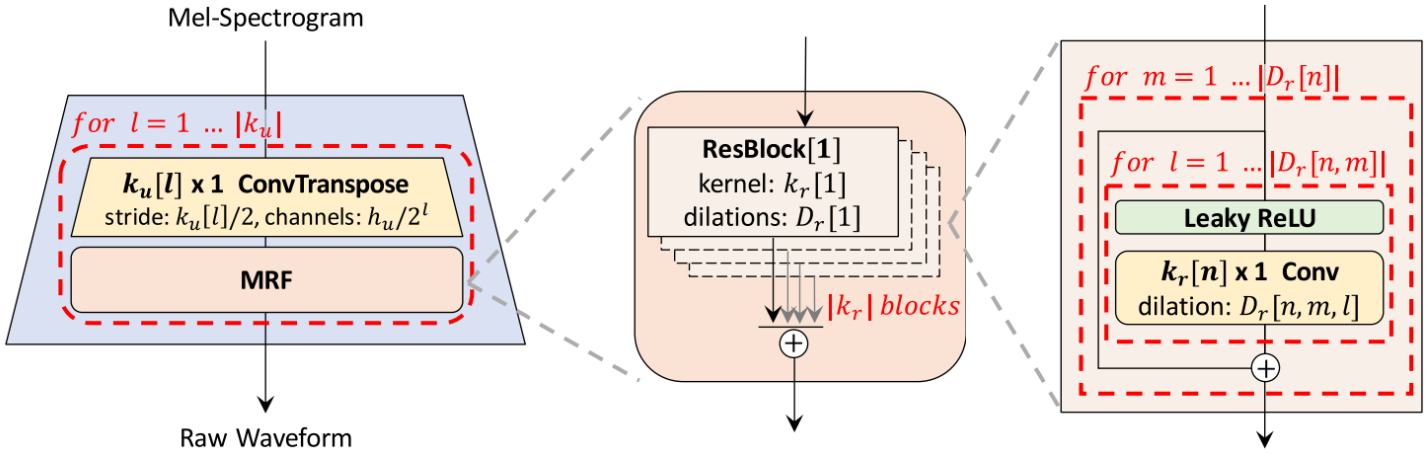
(Source: Kong et al., 2019)

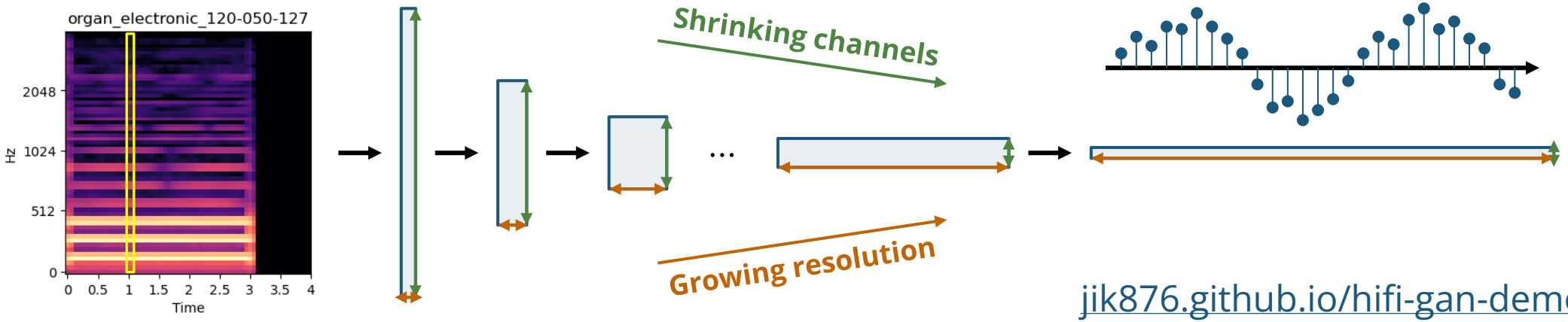
Multi-period discriminator



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

Hifi-GAN (Kong et al., 2020)

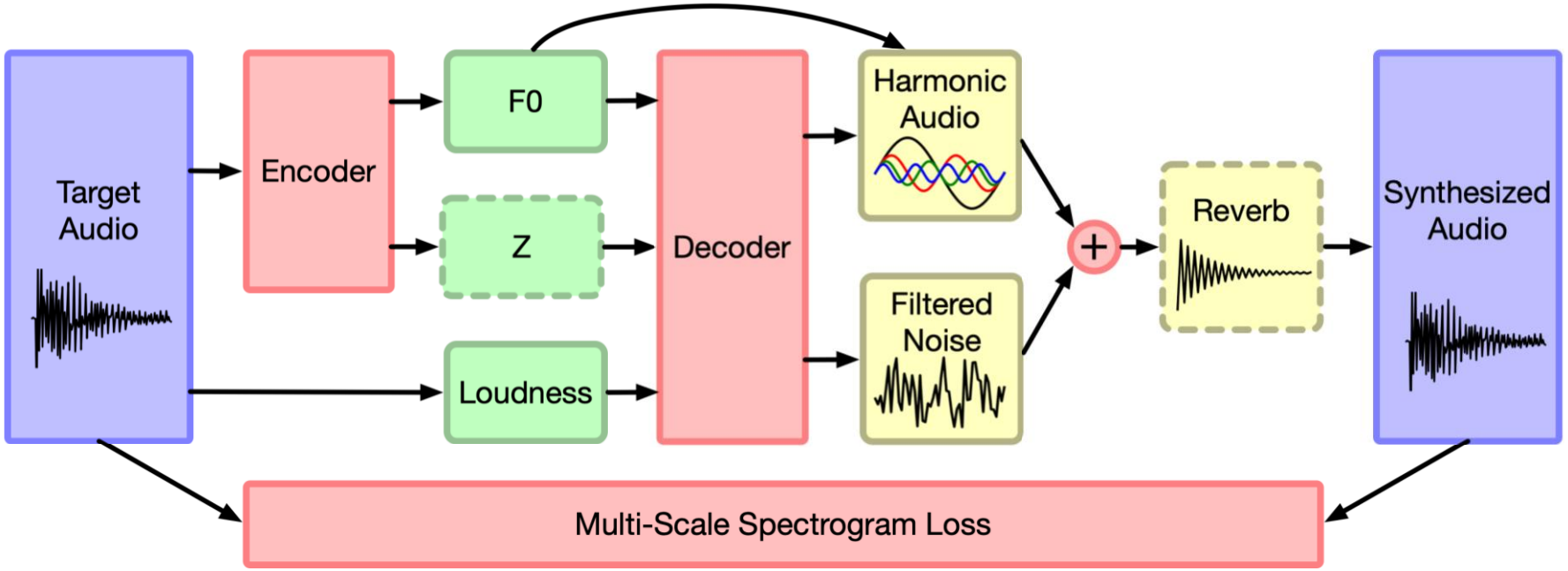




jik876.github.io/hifi-gan-demo

Differentiable Digital Signal Processing (DSP)

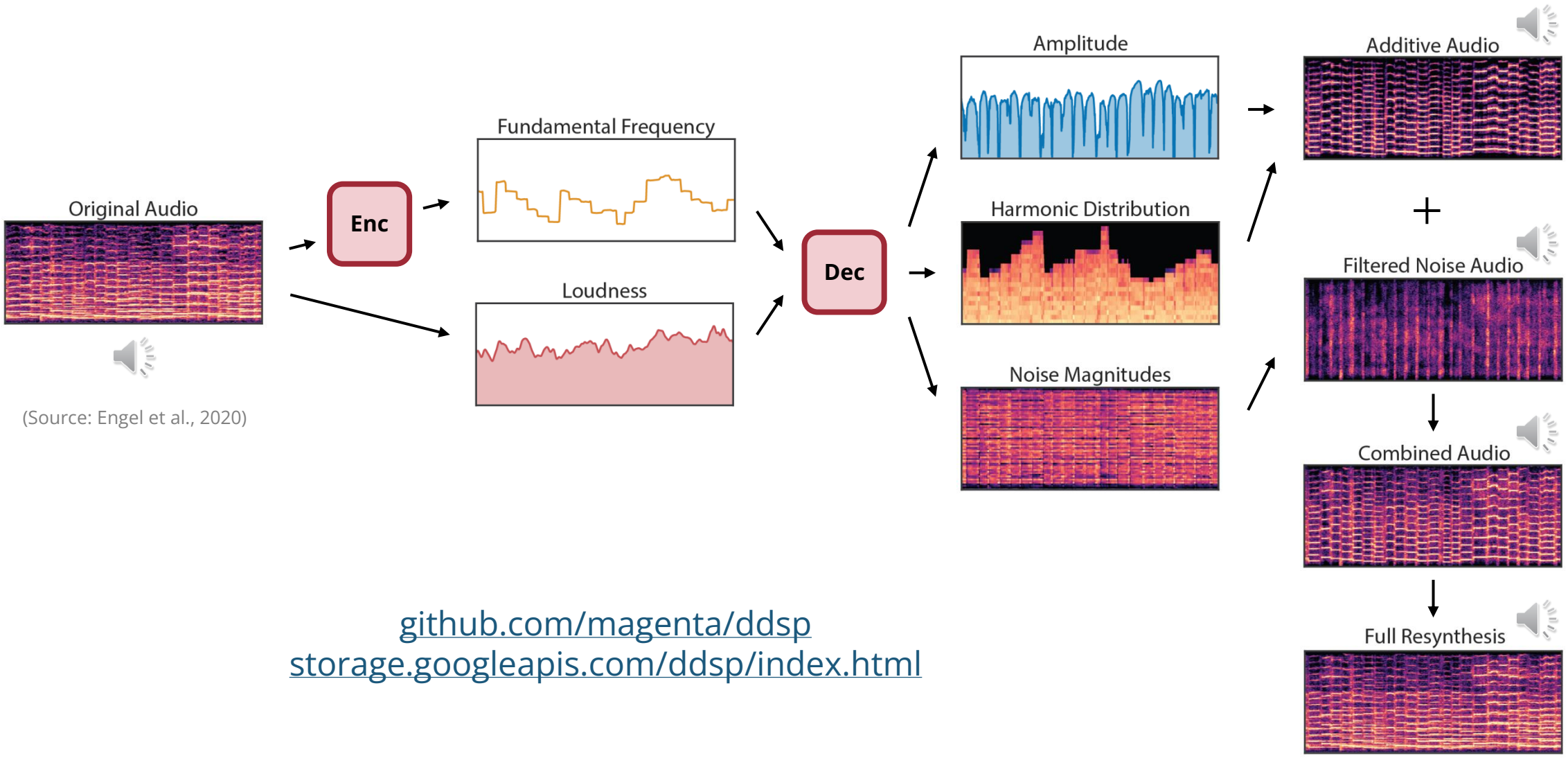
Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

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

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



(Source: Engel et al., 2020)

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

Entering Demons & Gods by Yaboi Hanoi (2022)

youtu.be/PbrRoR3nEVw

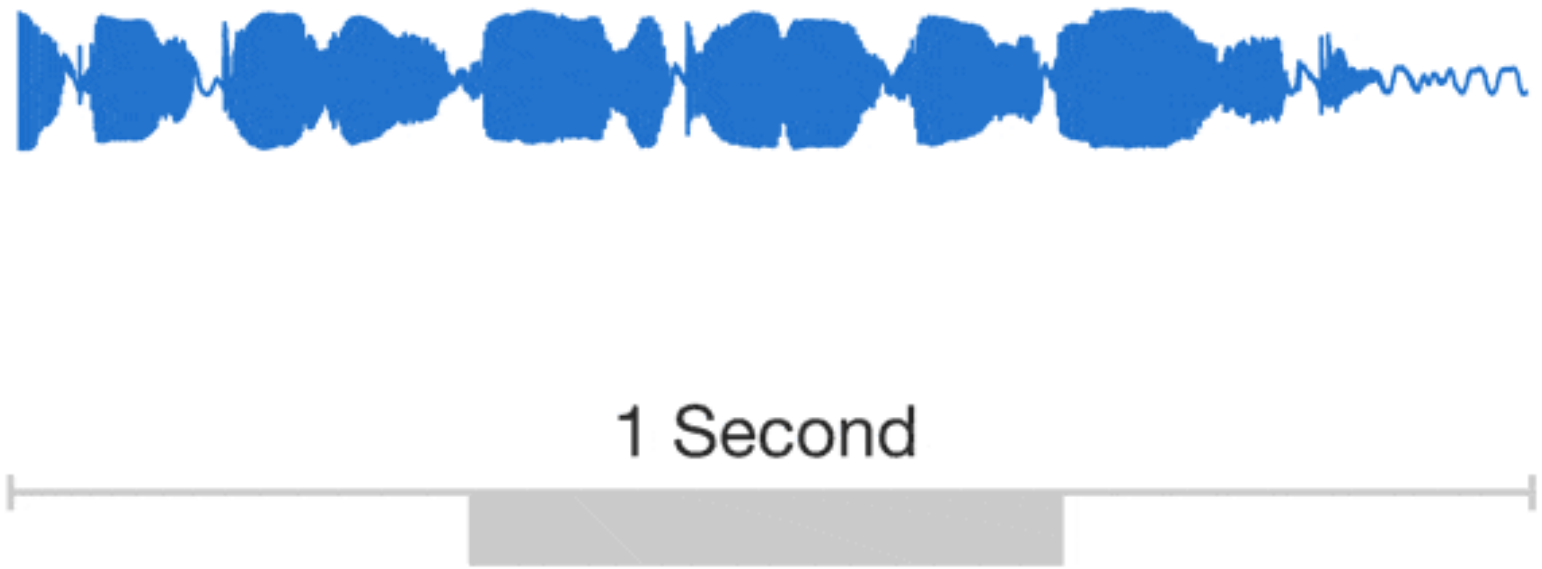
soundcloud.com/yaboihanoi/enter-demons-and-gods

Optional Reading

- A very nice blog on “**Generating music in the waveform domain**” by Sander Dieleman: sander.ai/2020/03/24/audio-generation

Recap

Generating Waveforms using a Neural Network



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

Aaron van den Oord and Sander Dieleman, "WaveNet: A generative model for raw audio," *DeepMind Blog*, September 8, 2016.

Autoregressive Models (Mathematically)

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

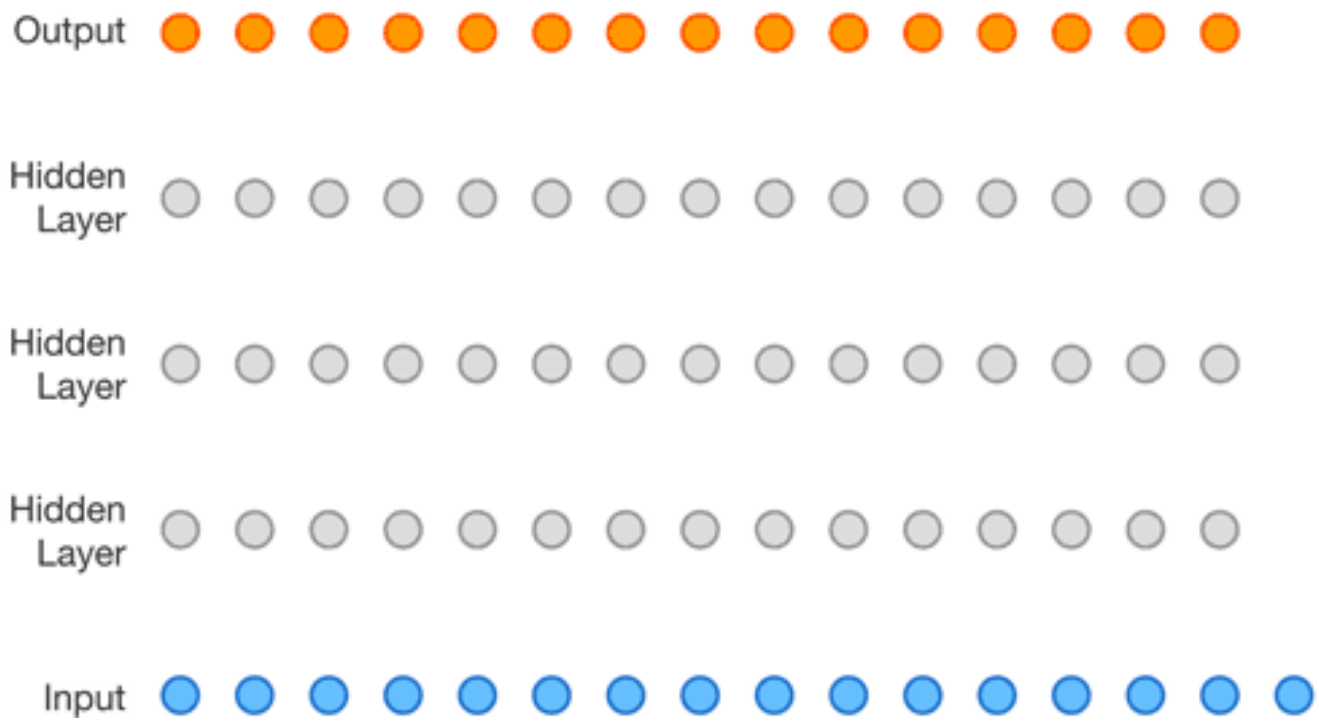
$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

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.

WaveNet (van den Oord et al., 2016)



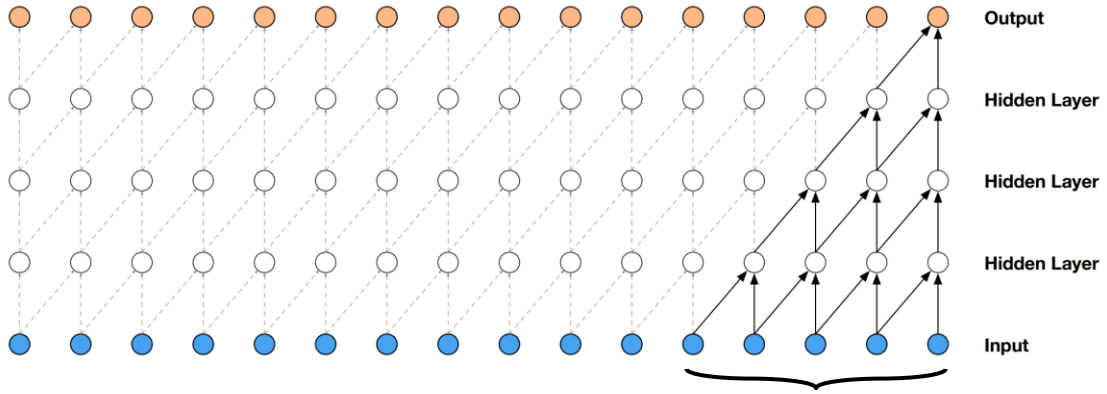
(Source: van den Oord et al., 2016)

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.

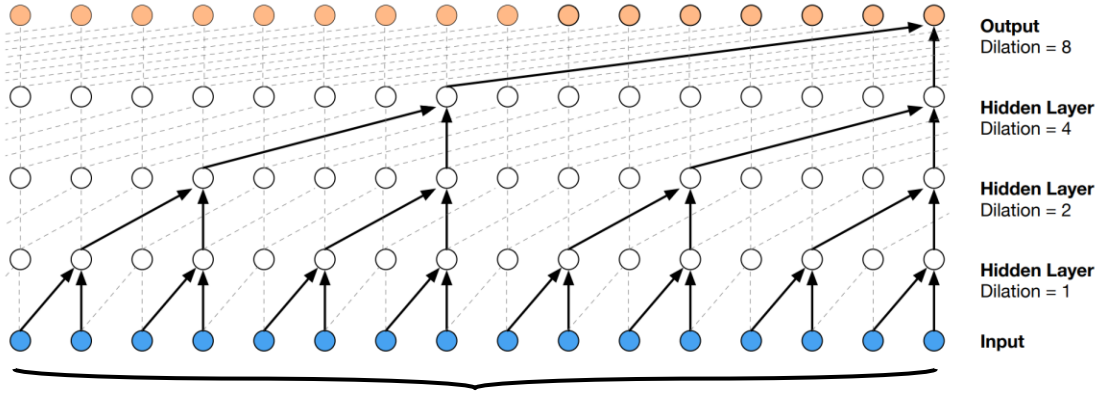
WaveNet (van den Oord et al., 2016)

Standard convolution



Smaller "receptive field"

Dilated convolution



Larger "receptive field"

deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio

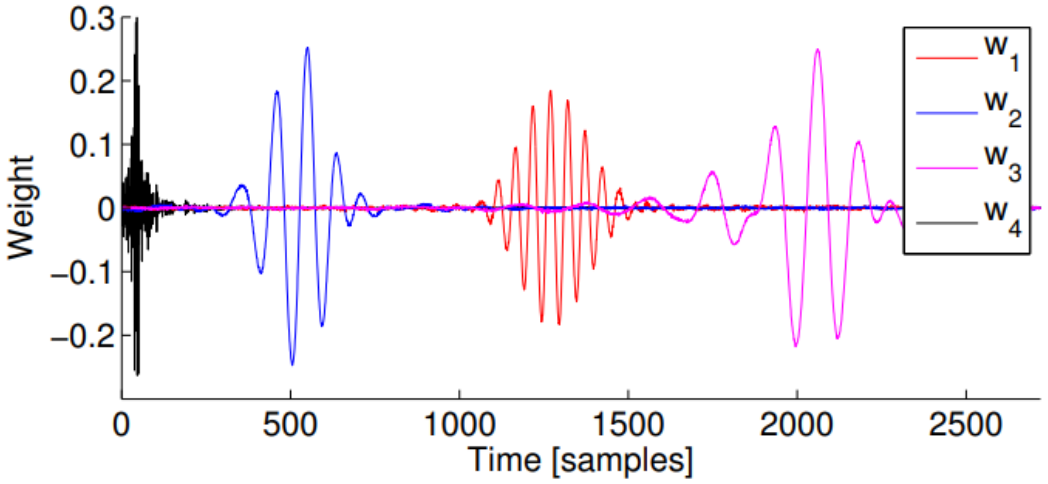
Example of generated music

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

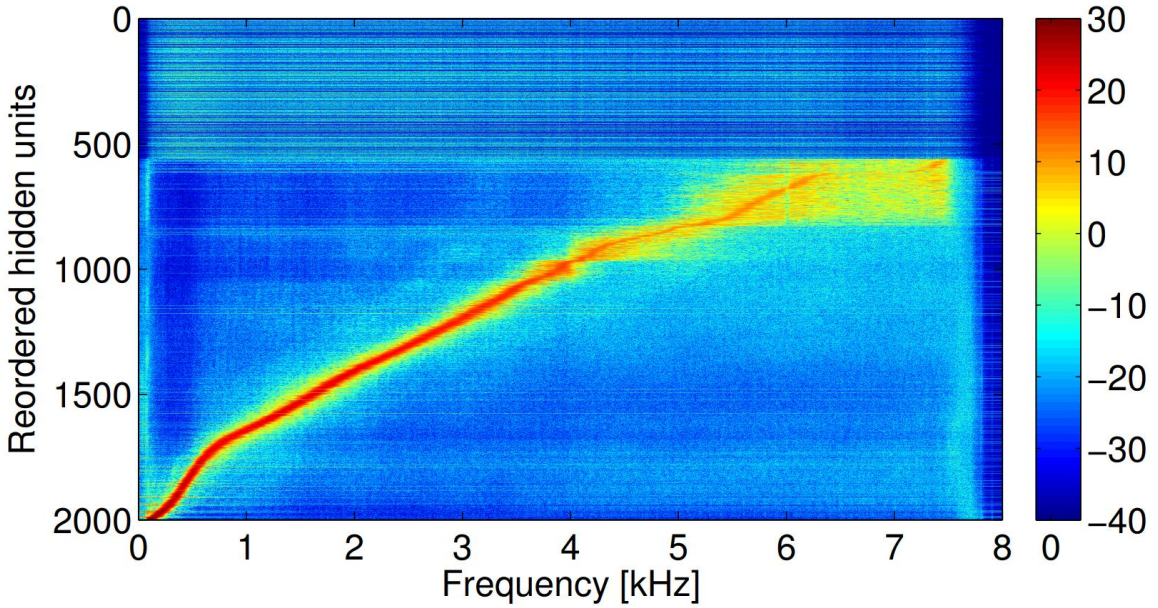
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

1D CNNs & Fourier Transform

Convolution kernels learned

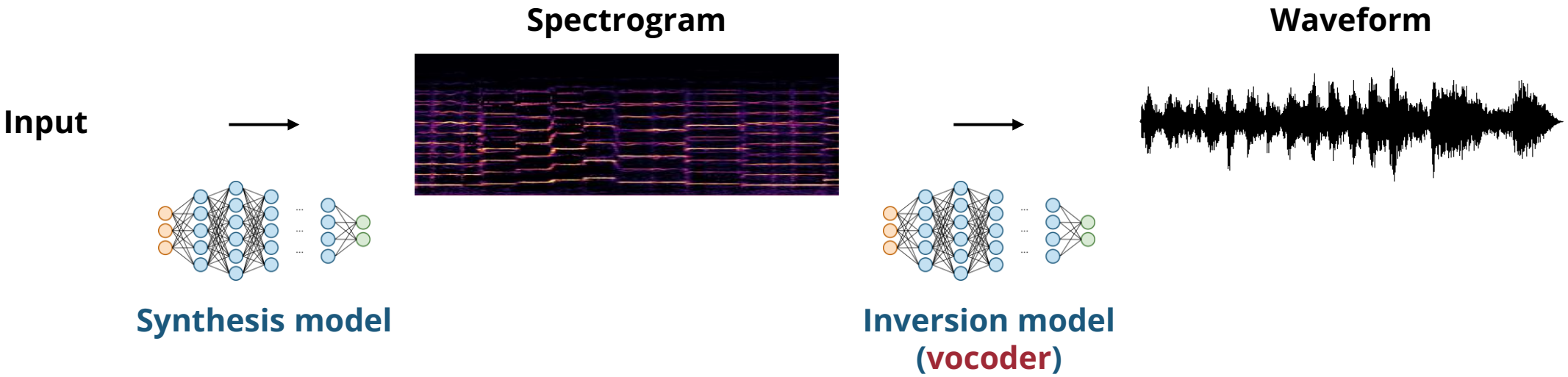


Peak frequency detected by the learned kernels

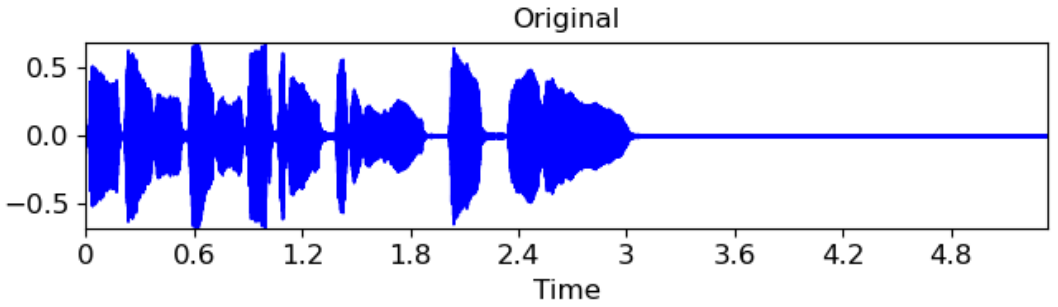


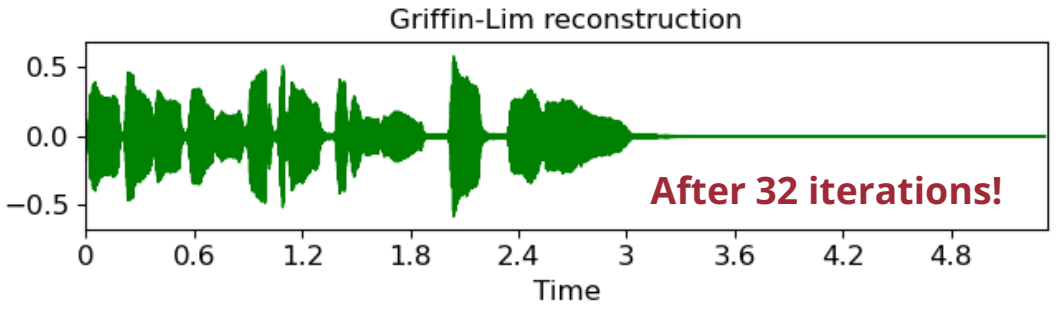
Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," *INTERSPEECH*, 2015.

Frequency-domain Audio Synthesis



Griffin-Lim Algorithm (Griffin & Lim, 1984)





(Source: librosa documentation)

Given a magnitude-only STFT matrix

↓

Randomly initialize the phase

↓

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

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

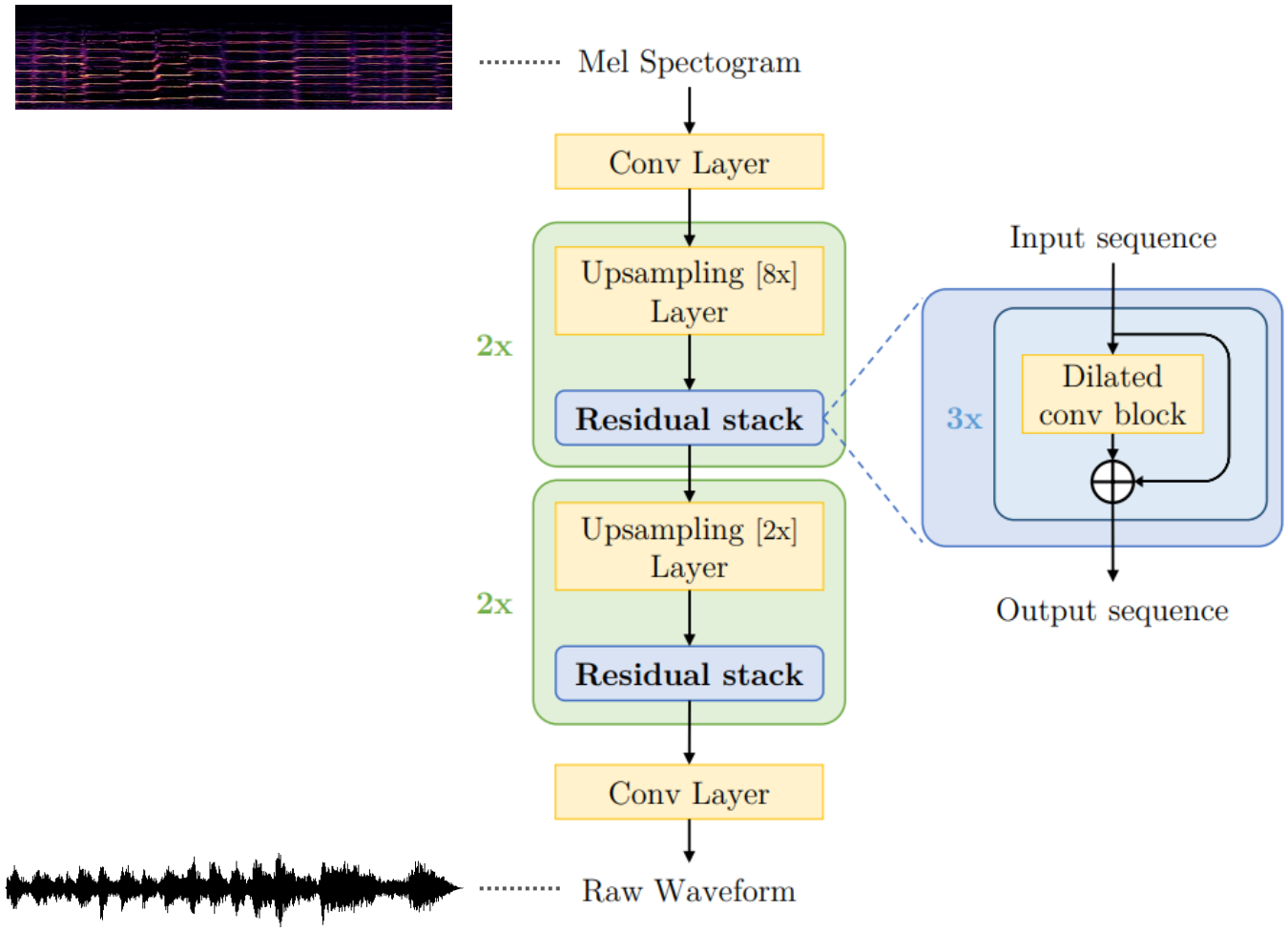
↓

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

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

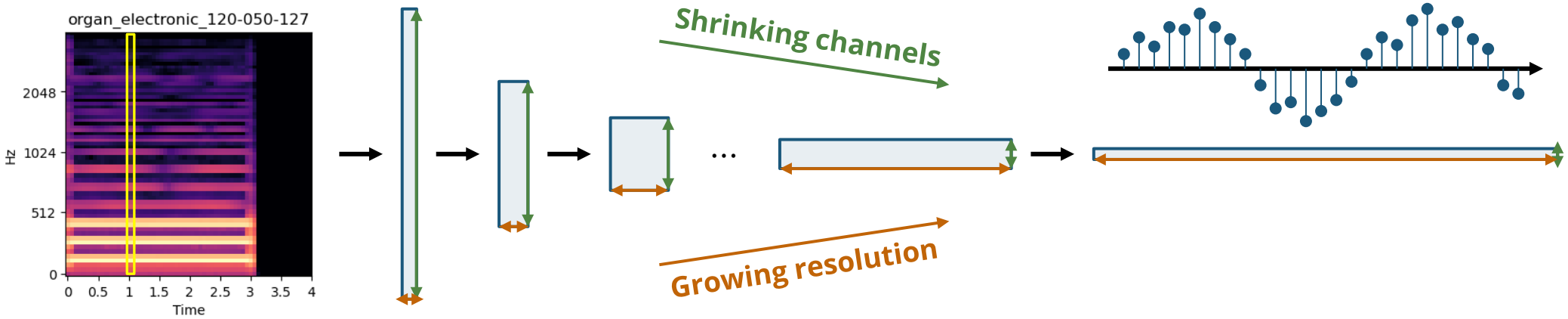
MelGAN (Kumar et al., 2019)



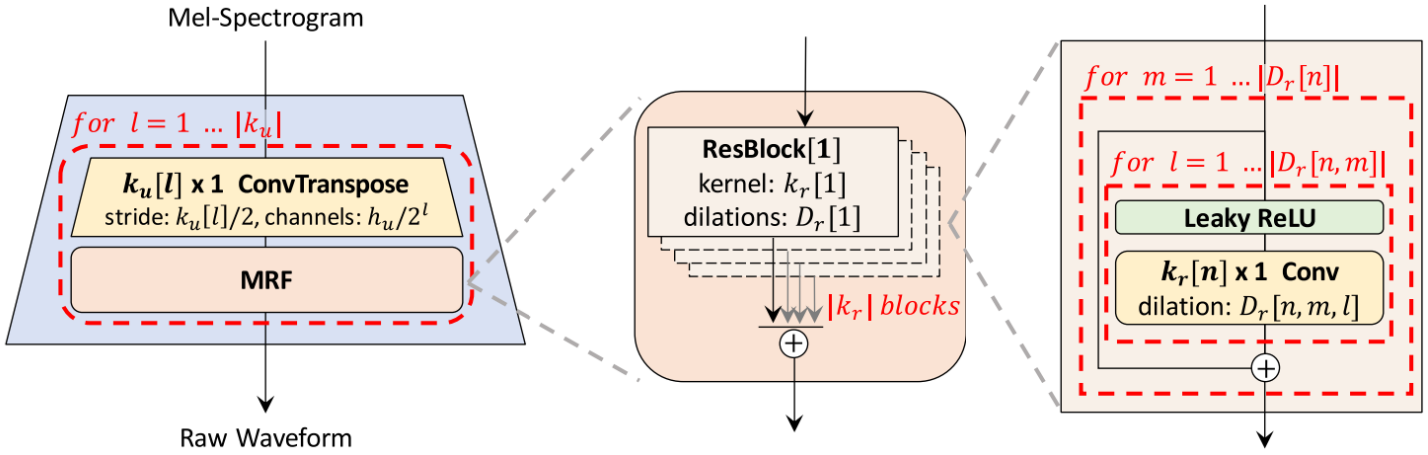
(Source: Kumar et al., 2019)

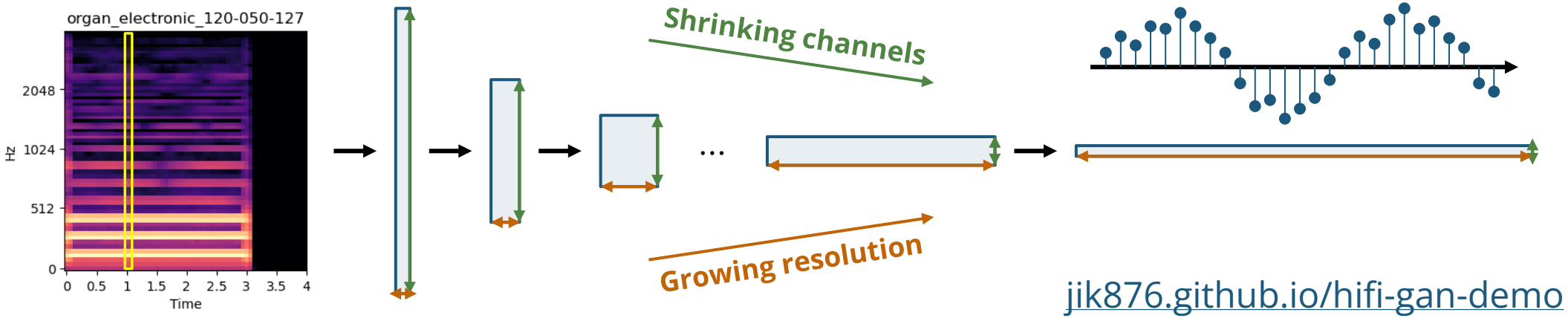
Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.

Upsampling for Vocoders



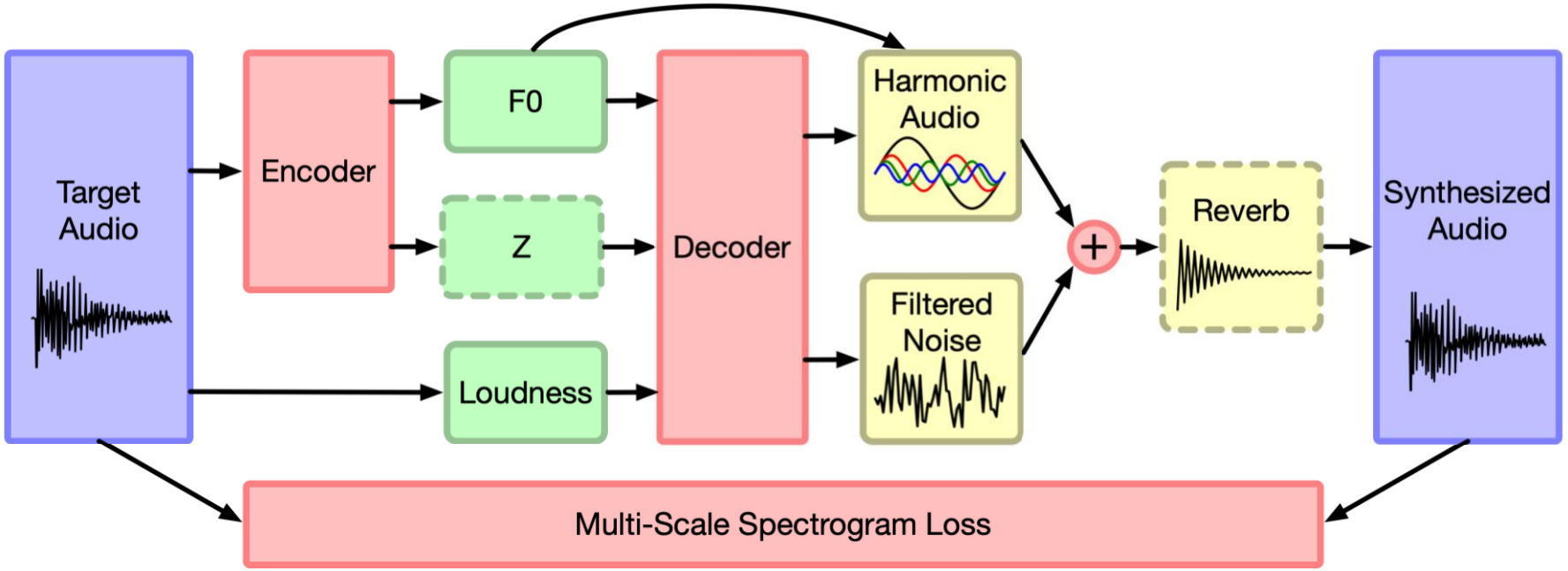
Hifi-GAN (Kong et al., 2020)





jik876.github.io/hifi-gan-demo

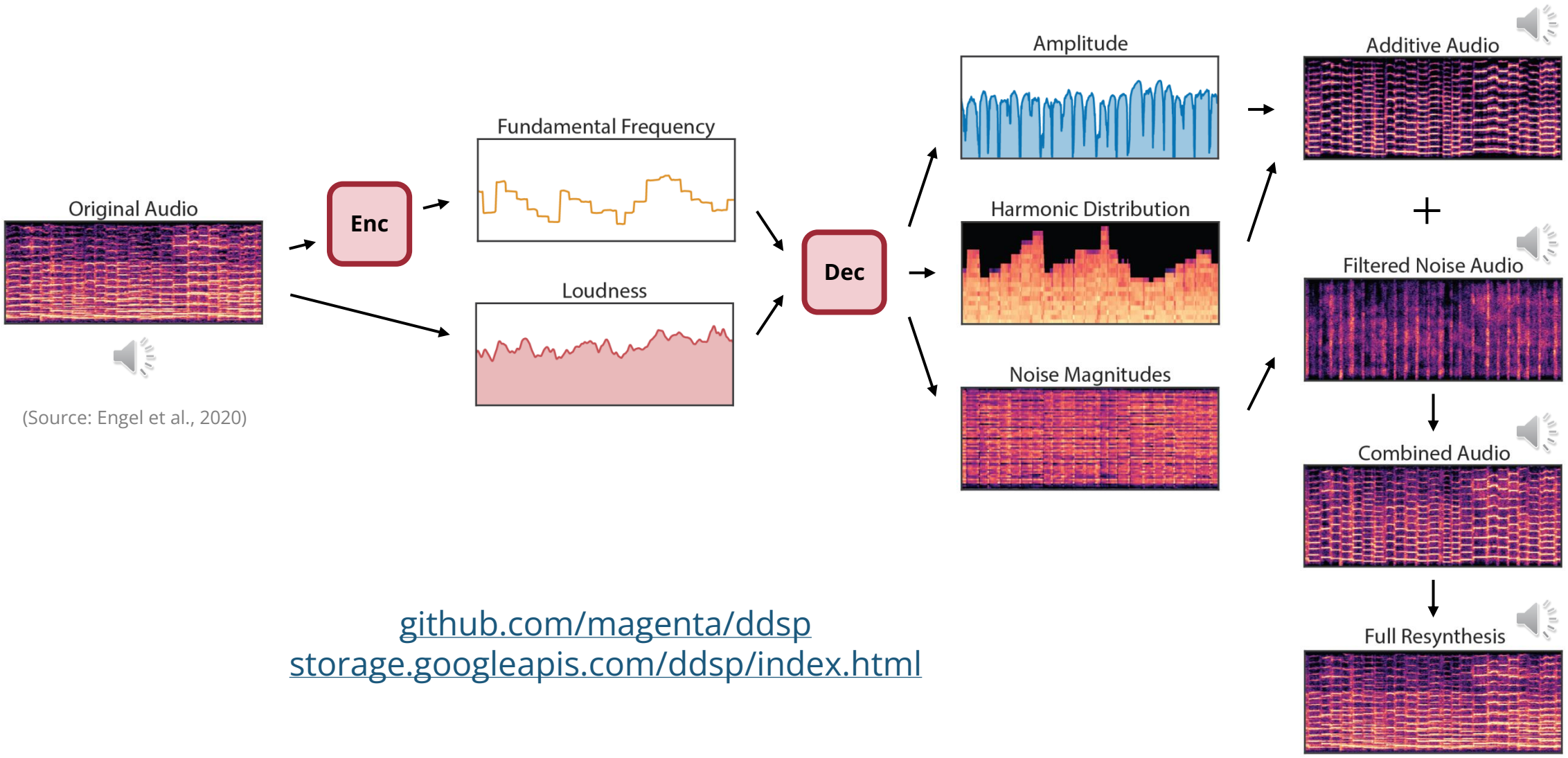
Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

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

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



(Source: Engel et al., 2020)

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

Entering Demons & Gods by Yaboi Hanoi (2022)

youtu.be/PbrRoR3nEVw

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

Next Lecture

Latent-based Music Generation