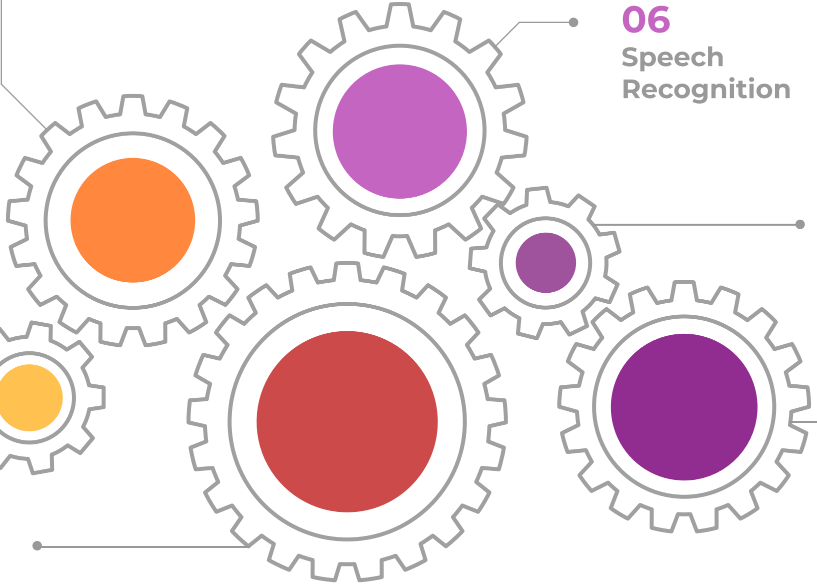
WAVENET:
A GENERATIVE MODEL FOR RAW AUDIO

Mahshid Alinoori

Table of Contents



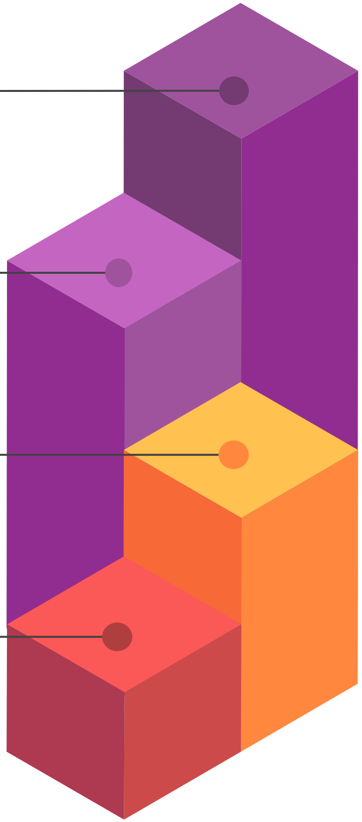
Why Wavenet came into existence?

Autoregressive
models gained popularity for modeling images and text

PixelRNN and PixelCNN successfully modelled the joint probabilities using the production of conditional probabilities

Possibility
of applying these models for generating audio waveforms

WaveNet
is introduced based on PixelCNN

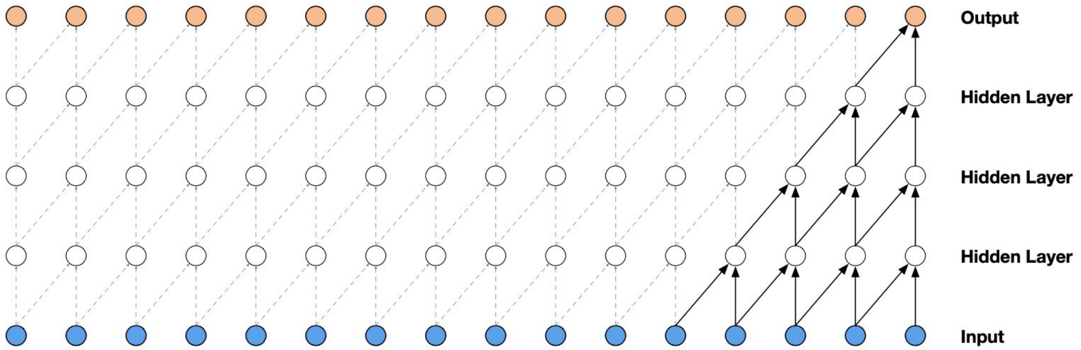


WaveNet: Causal Convolutions

- Audio waveform: $\{x_1,...,x_T\}$
- Modelled as:
  $p(x) = \prod p(x_i|x_1, ..., x_{i-1})$
- Output from a **softmax** layer
- Maximizing the **log-likelihood**

- **Causal convolutions** are the counterpart of **masked convolutions** in PixelCNN
- Used to remove the dependency on **future timesteps**

- Implemented by **shifting** the output
- Needs **many layers** or **larger kernels** to increase the receptive field which imposes **computational cost**

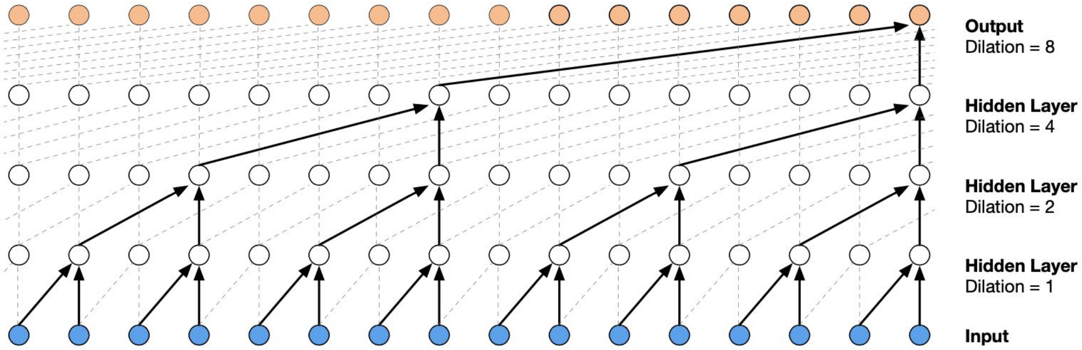


WaveNet: Dilated Casual Convolutions

- Dilated convolution is a convolution with **holes**
- Increases the receptive field by **skipping input values**

- Large receptive fields are achieved with **few layers**
- Dilating the original filter with **zeros**

- **Dilation 1** is the standard convolution
- The dilation is **doubled** for every layer up to a limit and then **repeated**: 1, 2, 4, ..., 512, 1, 2, 4, 512



WaveNet and PixelCNN Commonality

Gated Activation

- Same gated activation units as PixelCNN to mimic the complexity in RNNs: $\mathbf{y = \tanh(W_{k,f} * x) \odot \sigma(W_{k,g} * x)}$
- Works better than **RELU**

Global Conditionality

- The conditional distribution **p(x|h) = ∏ p(xi|x1, ..., xi−1, h)** is used to apply some **characteristics**
- Global condition h influences the output distribution in all timesteps:
- Modelled as:

$$\mathbf{y = \tanh(W_{k,f}*x+V_{k,f}^{T} h)\odot\sigma(W_{k,g}*x+V_{k,g}^{T} h)}$$

Local Conditionality

- Second time series h_t with lower sampling frequency
- Mapped to a new time series **y = f(h)** using **transposed convolutional network**

$$\mathbf{y = \tanh(W_{k,f} *x+V_{k,f} *y)\odot\sigma(W_{k,g} *x+V_{k,g} *y)}$$

More Details on WaveNet

Softmax Distribution

- Similar to PixelCNN, **categorical distribution** replaces the continuous distribution because of its flexibility
- Non-linear quantization using μ-law

Residual and Skip

- Residual and parameterized skip connections are adopted for **faster convergence**
- Helps when having **many layers** in the network

Context Stacks

- Another approach to **increase the receptive field**
- Smaller context stacks process a **long part of audio** and condition a large WaveNet that processes a small part of audio

01

Multi-Speaker Speech Generation

WaveNet Applications

- English **multi-speaker** corpus as the dataset
- **Conditional model** with the **speaker** as the external condition with promising results
- No condition on any text or content
  - **Non-existent words** but human-like intonation
- Capturing **acoustic quality, breathing, and mouth movement** in addition to the speaker's voice

02

Text to Speech

WaveNet Applications

- Single-speaker speech database
- Two conditional model:
  - **local** conditionality with **linguistic features** as the condition
  - logarithmic fundamental frequency **(log F_0)** in addition to the linguistic features
- Evaluation: **subjective paired comparison** and **mean opinion score**
- In both cases WaveNet with **two conditions** outperformed the rival models

03

Music

WaveNet Applications

- Experimented on two music datasets: **MagnaTagATune** and **YouTube piano dataset**
- Subjectively **sounding musical** is dependent on the large receptive fields
- **Conditional models** are used to set some quality tags for the output like the genre or instrument

WaveNet Applications

04

Speech Recognition

- WaveNet is also used in **discriminative** audio tasks
- Speech recognition by WaveNet is applied to the **raw audio**
- **Long-term dependencies** addressed by LSTMs are now taken care of using dilated convolutions
- Experimented on **TIMIT dataset**
- Trained with **two loss terms**

Conclusion

- Implemented with similar components t as in **gated PixelCNN and conditional PixelCNN**
- Applied to **TTS, music, and speech recognition** with the possibility of conditioning on some features
- Quantitative evaluation on TTS and qualitative evaluation on other tasks showed promising results in general

Thanks!