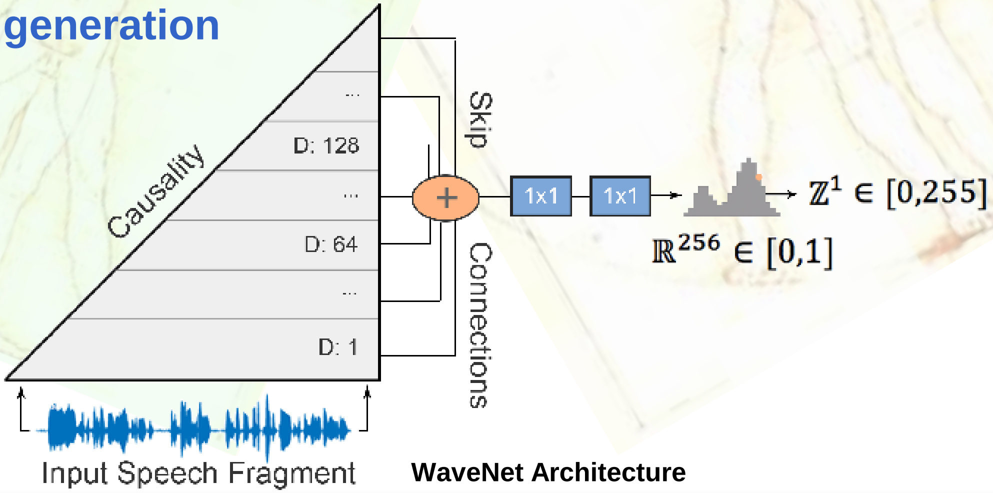
WaveNet

Pattern Recognition & Machine Learning Laboratory
Ji-Hoon Park
Aug 11, 2021

WaveNet: A Generative Model for Raw Audio [A. Oord et al., 2016] (1/5)

- **Goal**
  - Generating wideband raw speech signals with subjective naturalness by developing new architecture
- **Motivation**
  - Unnaturalness of the existing speech generation model
- **Contribution**
  - Applying PixelCNN[*A. Oord et al*., 2016] to speech generation model to get more natural sound
  - Introducing dilated causal convolutions to exhibit very large receptive fields in order to deal with long-range temporal dependencies needed for raw audio generation

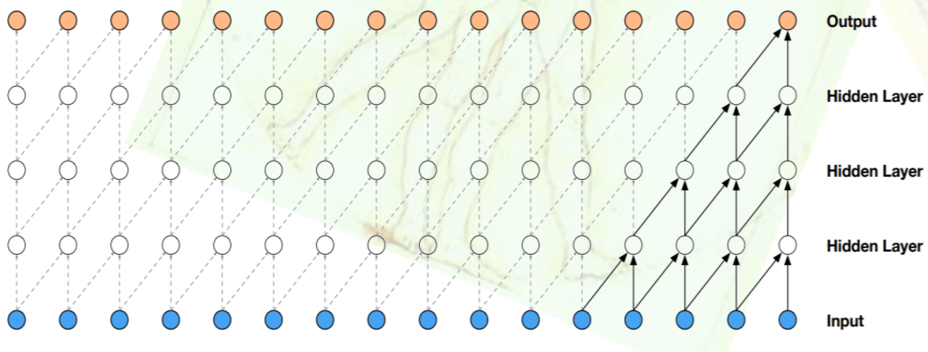


WaveNet Architecture

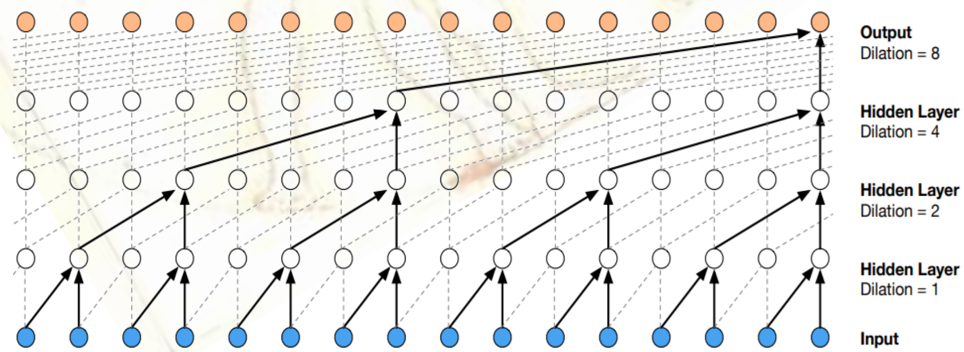
A. Oord et al., "WaveNet: A generative model for raw audio," *Arxiv*, 2016.
A. Oord et al., "Pixel Recurrent Neural Networks," *PMLR*, 2016.

WaveNet: A Generative Model for Raw Audio [A. Oord et al., 2016] (2/5)

- **Dilated causal convolution**
  - Enabling networks to have very large receptive fields with just a few layers
  - Causal convolution
    - Convolution layer that only depends on past timesteps
      - $p(x_{t+1}|x_1, \dots, x_t)$
    - Making sure the model cannot violate the ordering
    - Training faster than RNN
      - Without recurrent connections
    - Limitation: require many layers and filters to increase the receptive field
  - Dilated convolution
    - Skipping input values with a certain step
    - Enabling networks to have very large receptive fields with just a few layers



Causal convolution layers

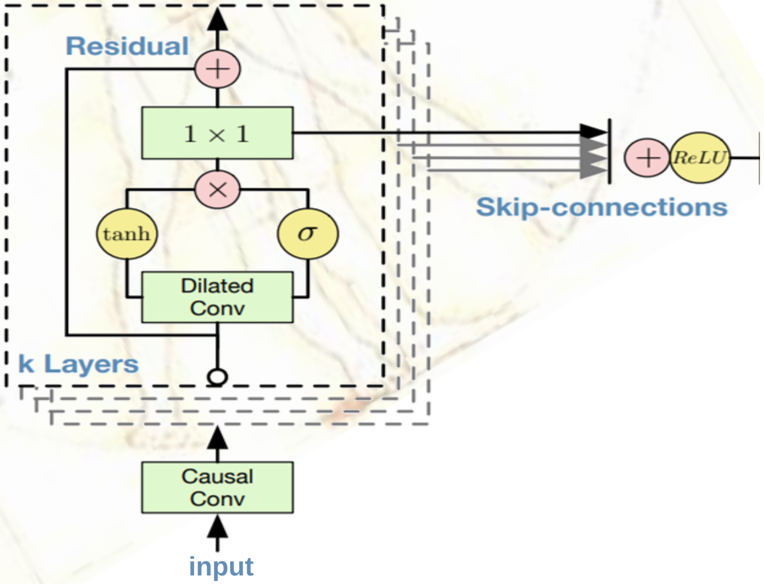


Dilated causal convolution layers

A. Oord et al., "WaveNet: A generative model for raw audio," *Arxiv*, 2016.

WaveNet: A Generative Model for Raw Audio [A. Oord et al., 2016] (3/5)

- **Modeling the conditional distribution**
  - Softmax distribution
    - Categorical distribution is more flexible
    - Easily modeling arbitrary distribution
      - Making no assumptions about distribution shape
  - μ-law companding transformation
    - Softmax layer need to output 65,536 probabilities (16-bit integer values)
    - Quantize it to 256 possible values (8-bit values)
      - $f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$
        - x_t: raw audio value $(-1 < x_t < 1)$
          μ: parameter (255)
- **Gated activation units**
  - Non-linearity activation function
    - $z = \tanh(W_{f,k} * \mathrm{x}) \odot \sigma(W_{g,k} * \mathrm{x})$
      - $*$: convolution operator
        $\odot$: element-wise multiplication
        σ: sigmoid function
        f: filter
        g: gate
        W: learnable convolution filter
        k: layer index

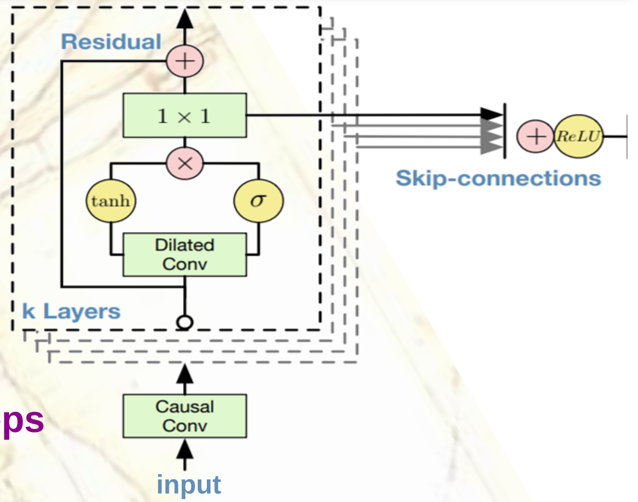


WaveNet Architecture

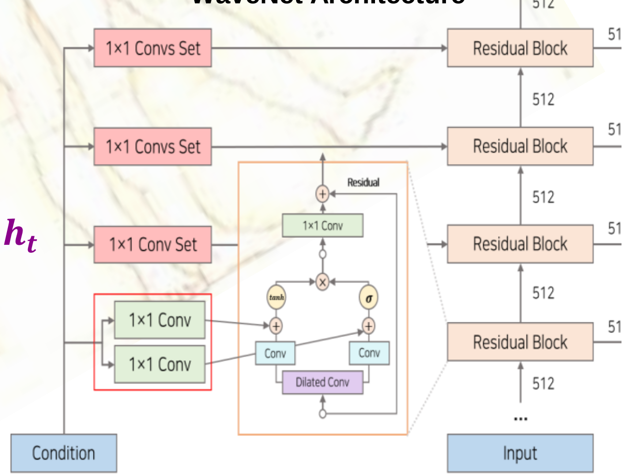
A. Oord et al., "WaveNet: A generative model for raw audio," *Arxiv*, 2016.

WaveNet: A Generative Model for Raw Audio [A. Oord et al., 2016] (4/5)

- **Residual and Skip connection**
  - ➢ Increasing convergence speed
  - ➢ Enabling training of much deeper models
- **Conditional WaveNet**
  - ➢ Conditioning the model on other input variables
  - ➢ Global conditioning
    - Influencing the output distribution across all timesteps
    - Conditioning Speaker identity on model
    - Activation function
      - – $\boldsymbol{z} = \tanh\left(W_{f,k} * \boldsymbol{x} + V_{f,k}^{T}\boldsymbol{h}\right) \odot \sigma\left(W_{g,k} * \boldsymbol{x} + V_{g,k}^{T}\boldsymbol{h}\right)$
        - » $\boldsymbol{h}$: conditioning parameter
          $\boldsymbol{V}_{*,k}$: learnable linear projection
  - ➢ Local conditioning
    - Influencing the output distribution across timesteps $\boldsymbol{h}_t$
    - Transforming time series using a transposed convolution network
    - Conditioning Linguistic features on model
    - Activation function
      - – $\boldsymbol{z} = \tanh\left(W_{f,k} * \boldsymbol{x} + V_{f,k}^{T}\boldsymbol{h} * \boldsymbol{y}\right) \odot \sigma\left(W_{g,k} * \boldsymbol{x} + V_{g,k}^{T}\boldsymbol{h} * \boldsymbol{y}\right)$
        - » $\boldsymbol{y}$: new time series $\boldsymbol{y} = f(\boldsymbol{h})$, $V_{g,k}^{T}\boldsymbol{h} * \boldsymbol{y}$: 1×1 convolution operation



WaveNet Architecture



Conditional WaveNet Architecture

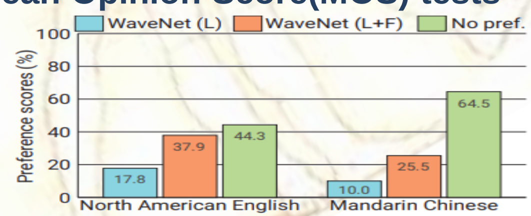
A. Oord et al., “WaveNet: A generative model for raw audio," *Arxiv*, 2016.

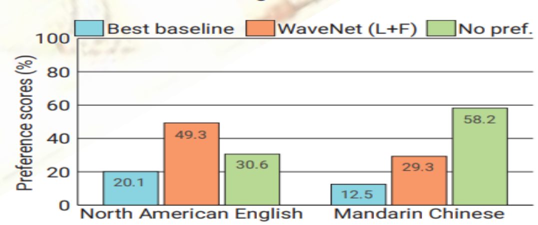
WaveNet: A Generative Model for Raw Audio [A. Oord et al., 2016] (5/5)

- **Experiments**
  - Multi-speaker speech generation
    - Conditioning model on a one-hot encoding of a speaker
    - Dataset
      - CSTR Voice Cloning Toolkit (VCTK)
  - Text to Speech (TTS)
    - Comparing WaveNet to HMM (Hidden Marcov Model) and LSTM-RNN (Long Short-Term Memory Recurrent Neural Network)
    - Datasets
      - Google's North America English and Mandarin Chinese dataset
    - Tests
      - Subjective paired comparison tests and Mean Opinion Score(MOS) tests
  - Music and Speech recognition
    - Datasets
      - MagnaTagATune dataset, YouTube piano dataset, and TIMIT dataset

Mean Opinion Score test results

Speech samples	Subjective 5-scale MOS in naturalness: North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	**4.21** ± 0.081	**4.08** ± 0.085
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071





Subjective paired comparison test results

WaveNet(L+F): Conditioned on linguistic features and $\log F_0$

F_0: Basic frequency

A. Oord et al., "WaveNet: A generative model for raw audio," *Arxiv*, 2016.