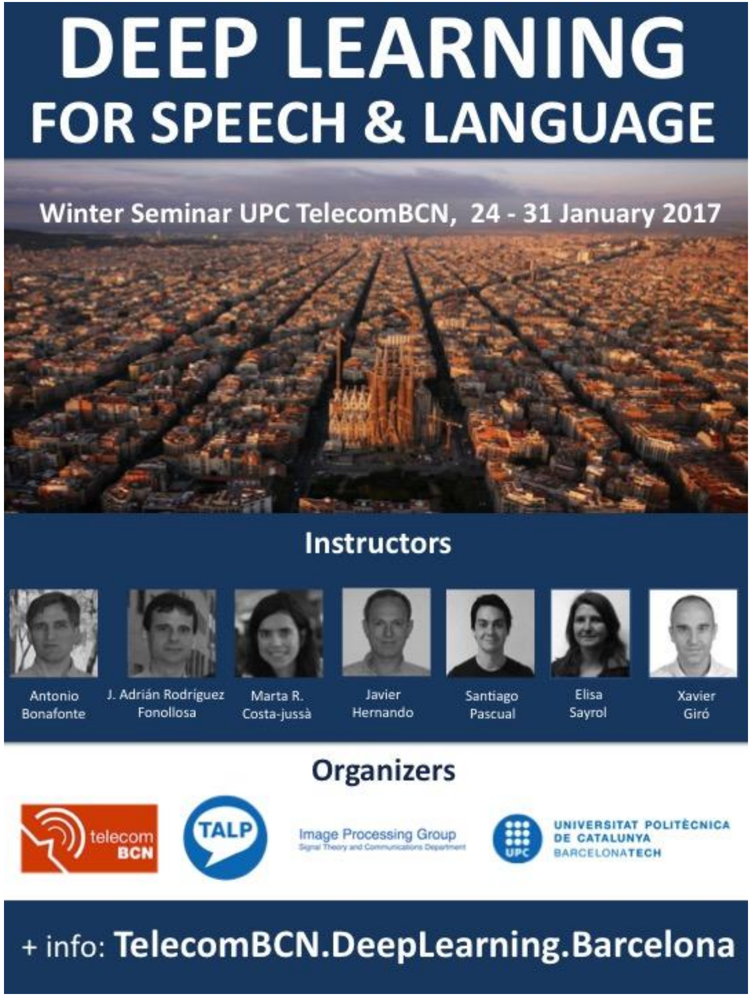


[course site]

Day 4 Lecture 3

Speech Synthesis: WaveNet

Antonio Bonafonte

WaveNet: A Generative Model for Raw Audio

Aäron van den Oord **Sander Dieleman** **Heiga Zen**[†]

Karen Simonyan **Oriol Vinyals** **Alex Graves**

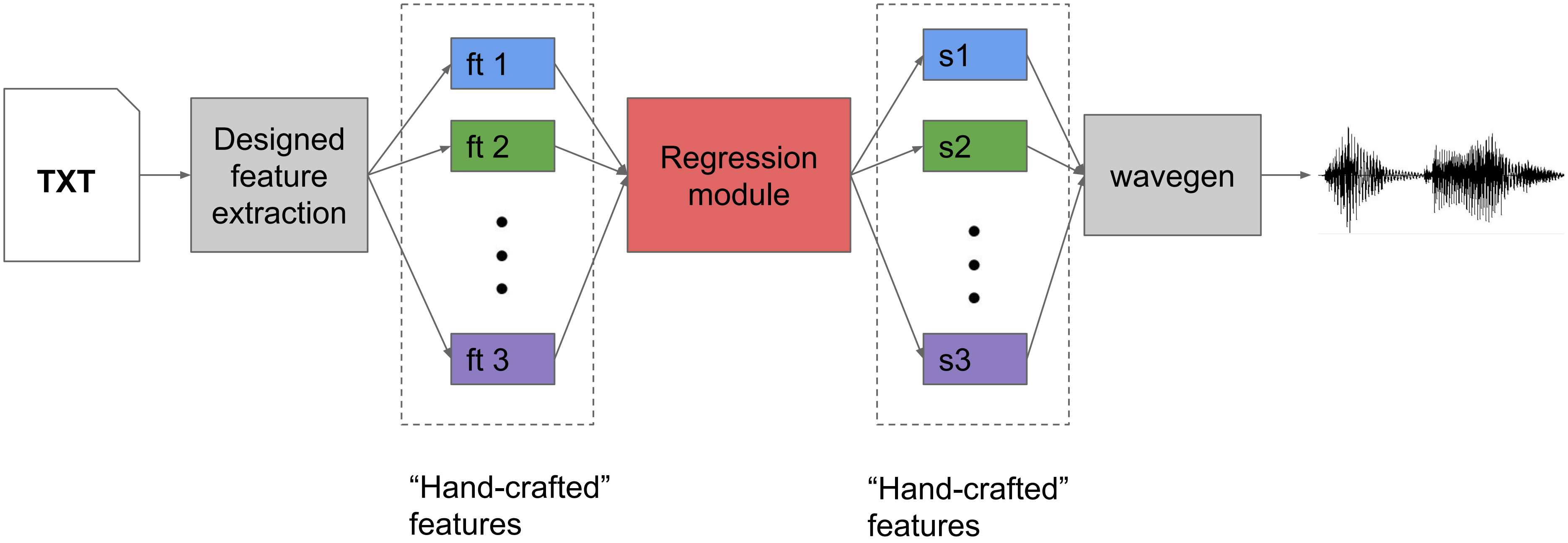
Nal Kalchbrenner **Andrew Senior** **Koray Kavukcuoglu**

deepmind.com/blog/wavenet-generative-model-raw-audio/

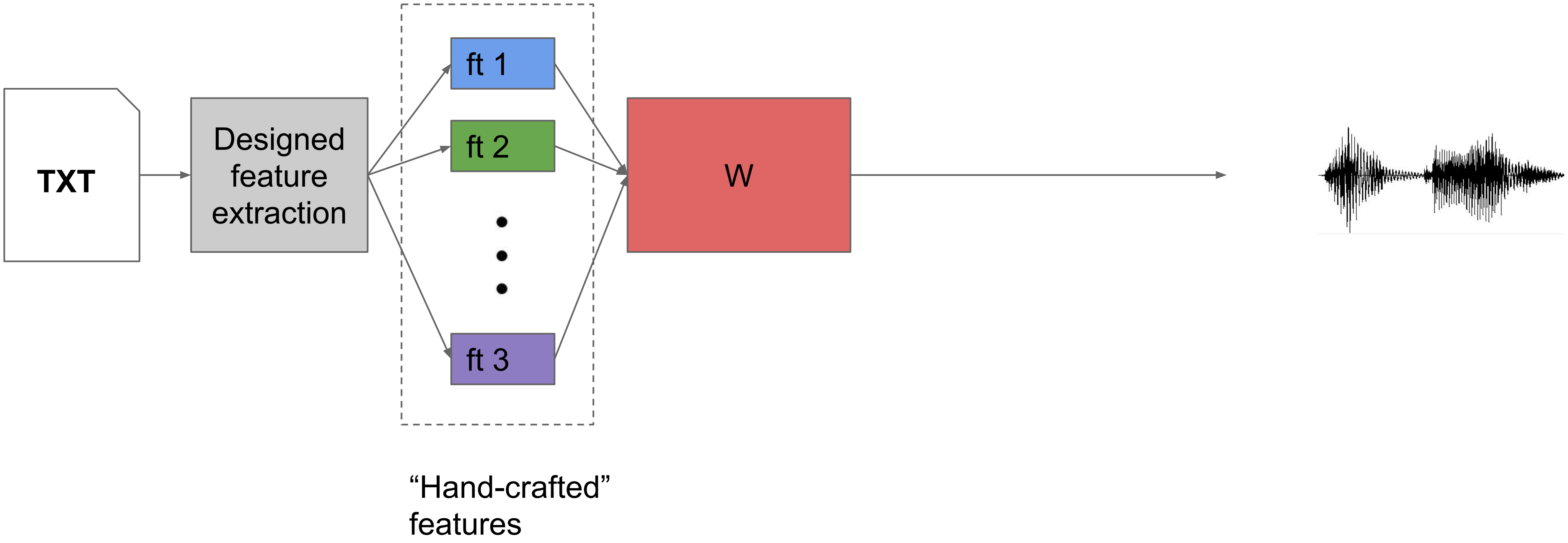
September 2016

Deep architectures ... but not deep (yet)

Text to Speech: Textual features → Spectrum of speech (many coefficients)



Text-to-Speech using WaveNet



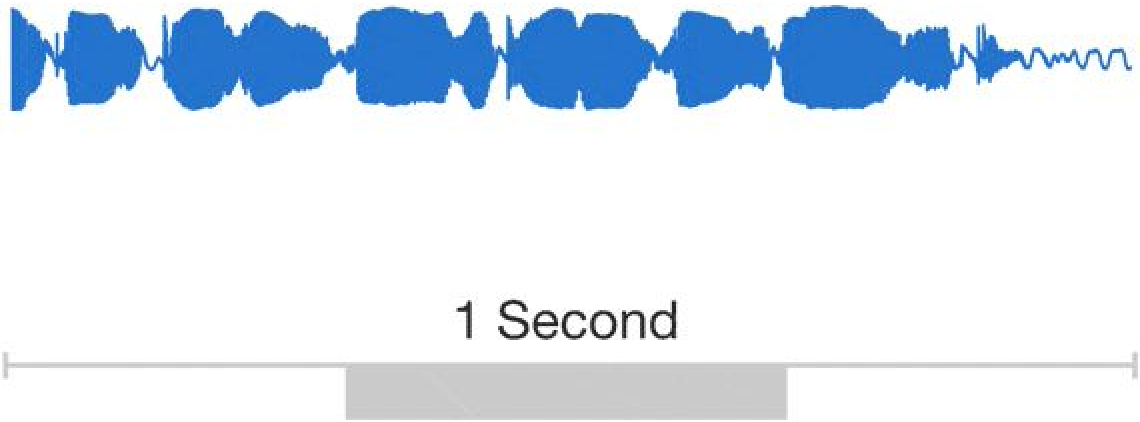
Introduction

- Based on PixelCNN
- Generative model operating directly on audio samples
- Objective: factorised joint probability

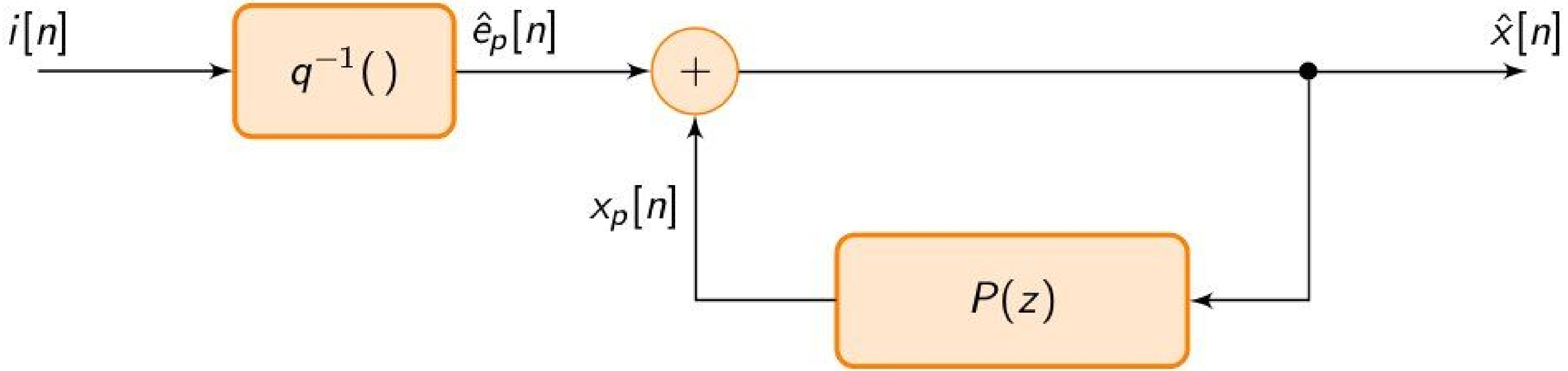
$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

- Stack of convolutional networks
- Output: categorical distribution → softmax
- Hyperparameters & overfitting controlled on validation set

High resolution signal and long term dependencies

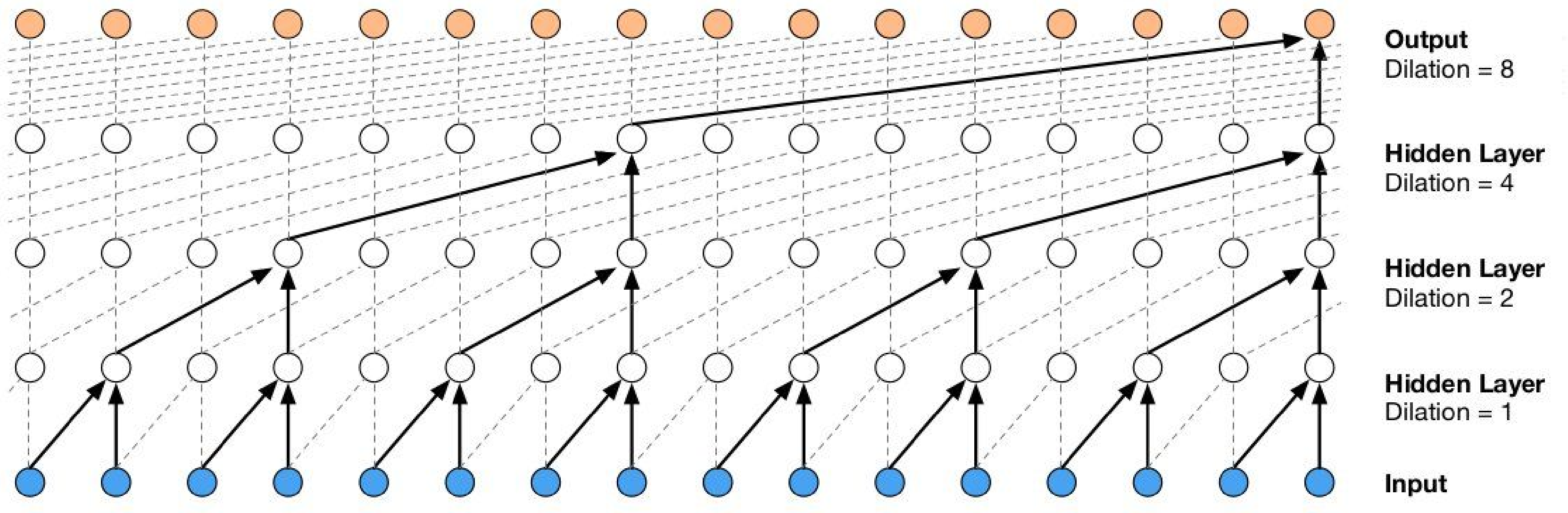


Autoregressive model



DPCM decoder: next sample is (almost) reconstructed from linear causal convolution of past samples

Dilated causal convolutions

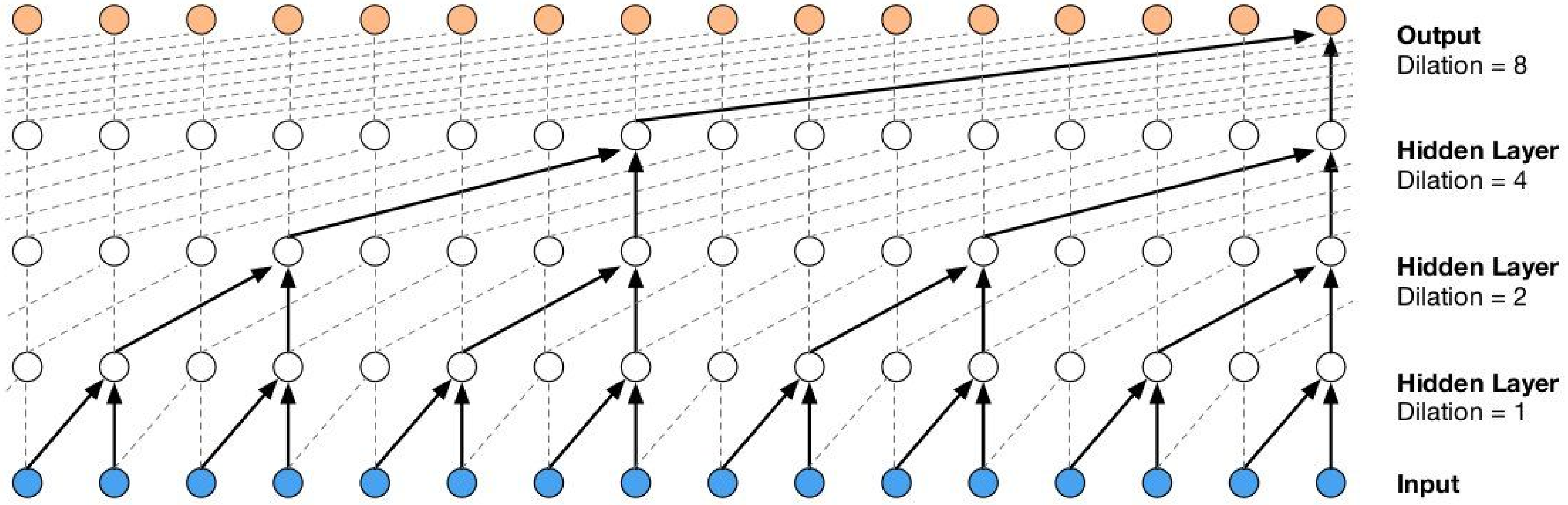


Stacked dilated convolutions:

Eg: 1, 2, 4, . . . , 512, 1, 2, 4, . . . , 512, 1, 2, 4, . . . , 512

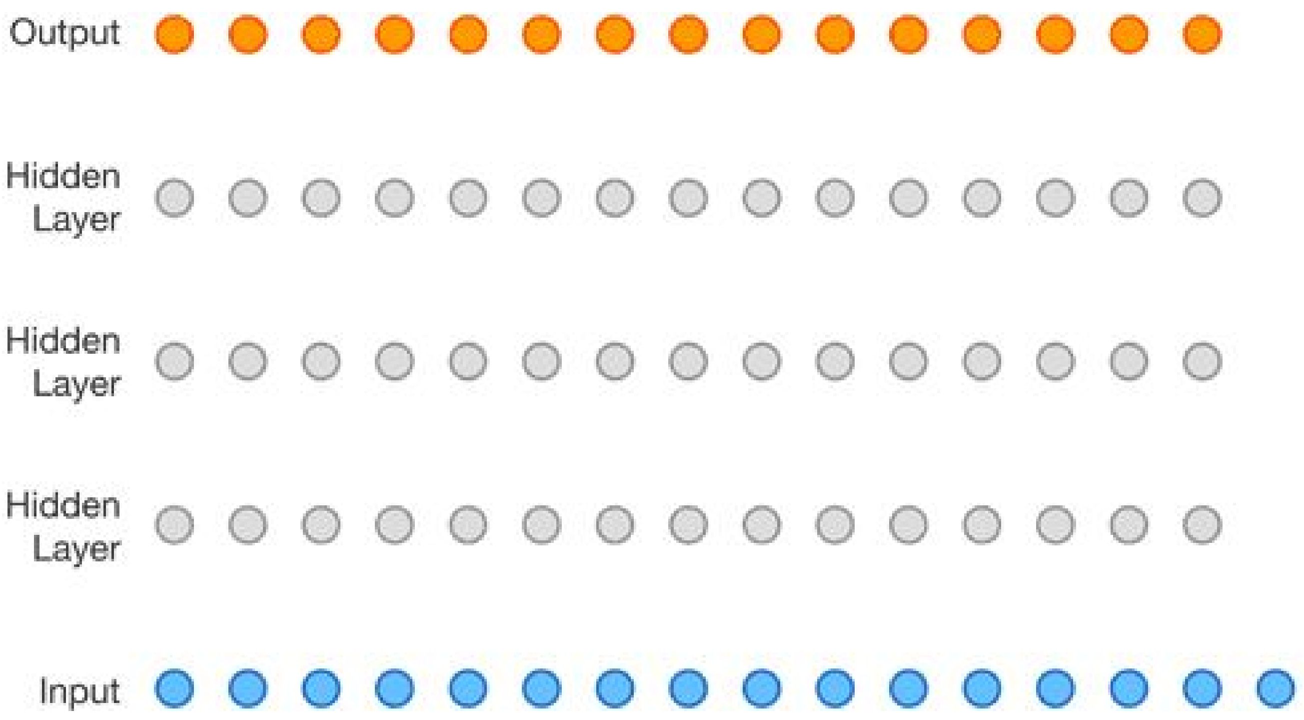
Receptive field: 1024 x 3 → 192 ms (at 16kHz)

Dilated causal convolutions



In training: all convolutions can be done in parallel

Dilated causal convolutions



Generating: predictions are sequential (~ 2min. per second)

Modeling pdf

$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

- Not MSE
- Not Mixture Density Networks (MDN)
- But categorical distribution, softmax (classification problem)

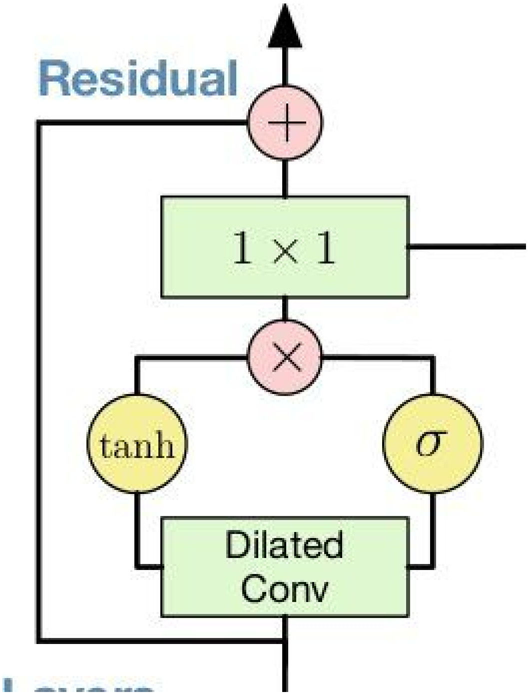
Modeling pdf

A softmax distribution tends to work better, even when the data is implicitly continuous (as is the case for image pixel intensities or audio sample values)

Van den Oord et al. 2016

Signal represented using mu law: 16 bits → 8 bits (256 categories)

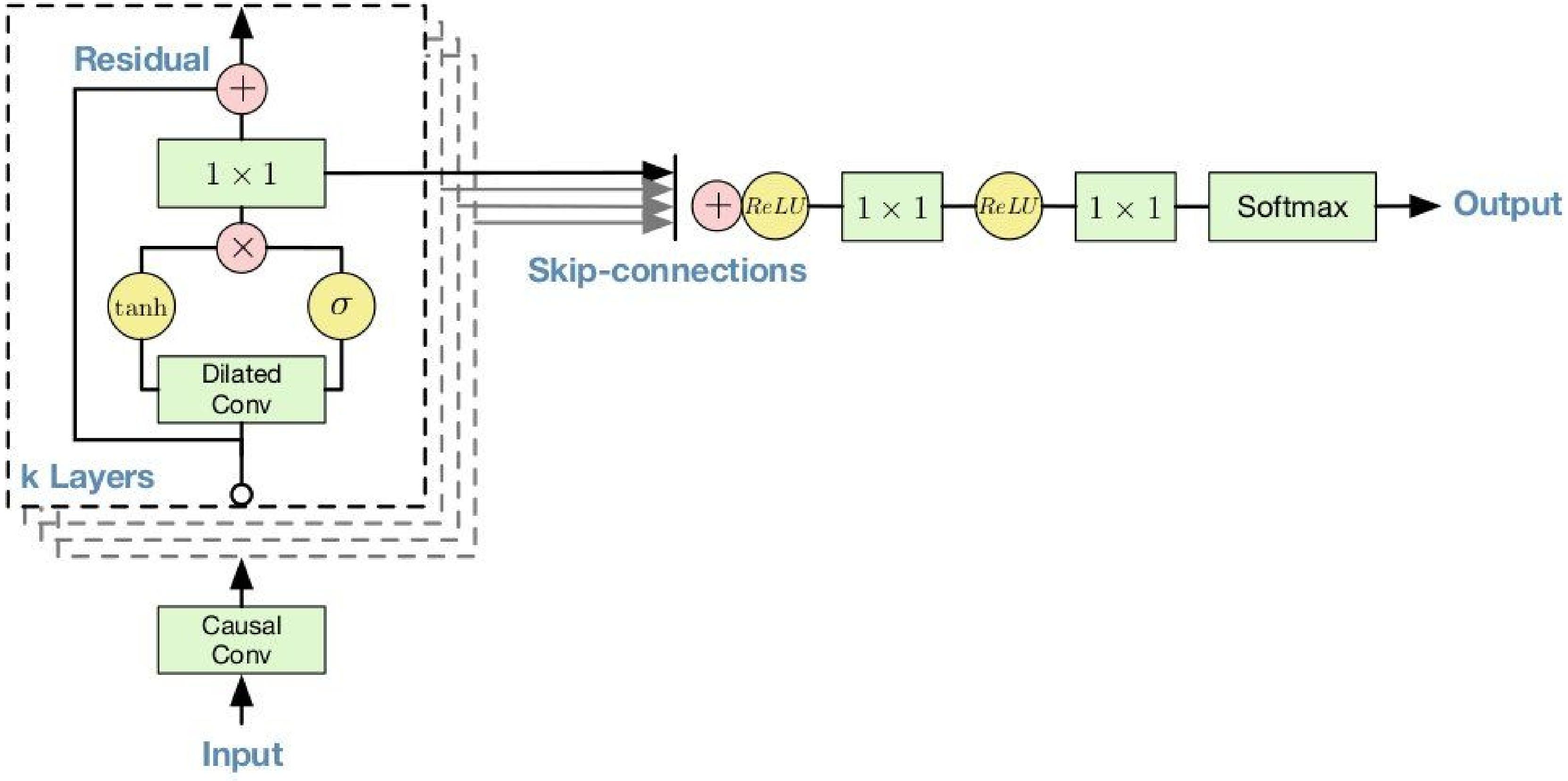
Gated Activation Units



$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

Residual Learning

Architecture



Conditional WaveNet

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

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

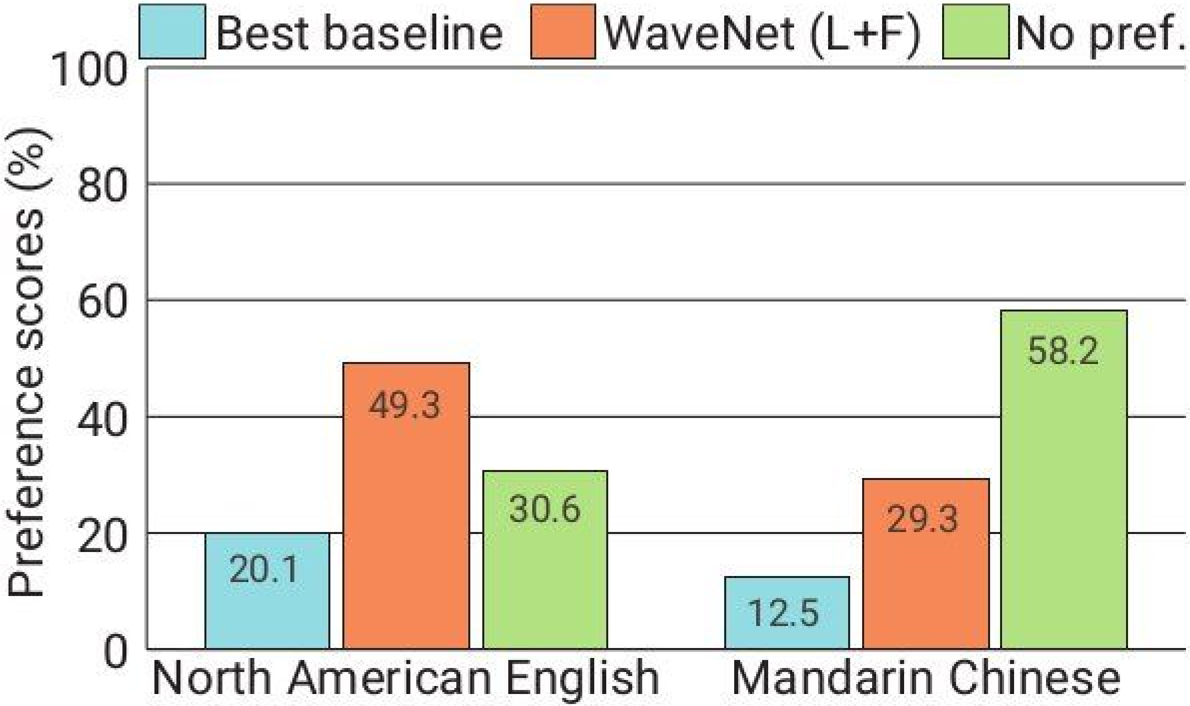
They show results with **h:**

- Speaker ID
- Music genre, instrument
- **TTS: Linguistic Features +F0**. (duration model needed to switch condition phoneme to phoneme.

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)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
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

Results



Listen yourself!

Discussion

- Wavenet: deep generative model of audio samples
- Convolutional nets: faster than RNN
- Outperforms best TTS systems
- Autoregressive model: sequential model in generation

GANs were designed to be able to generate all of x in parallel, yielding greater generation speed

Ian Goodfellow
NIPS 2016 Tutorial: Generative Adversarial Networks

WaveNet: A Generative Model for Raw Audio

Aäron van den Oord Sander Dieleman Heiga Zen[†]
Karen Simonyan Oriol Vinyals Alex Graves
Nal Kalchbrenner Andrew Senior Koray Kavukcuoglu

deepmind.com/blog/wavenet-generative-model-raw-audio/

September 2016