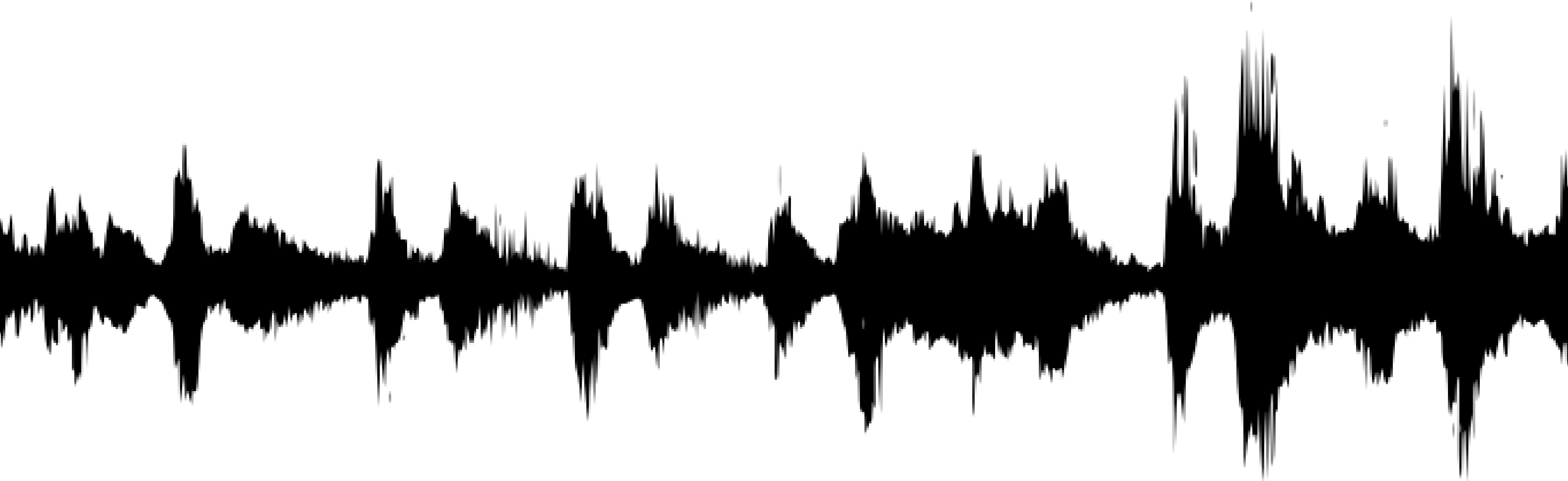
WaveNet

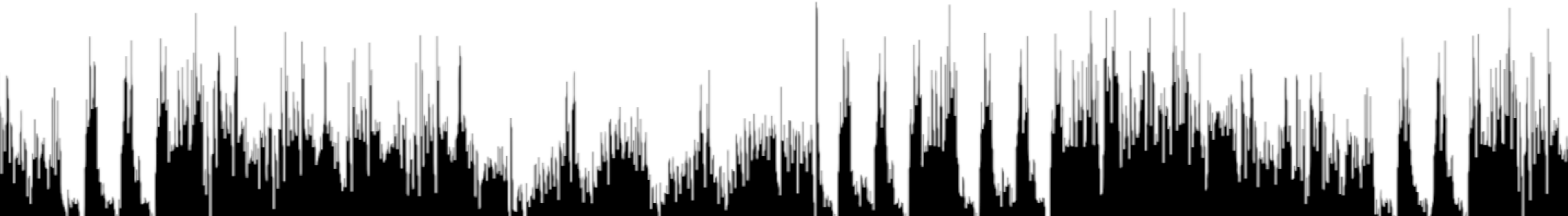


Von: Jonas Zimmer
Seminar Deep-Learning
Fakultät INFM
Hochschule Offenburg

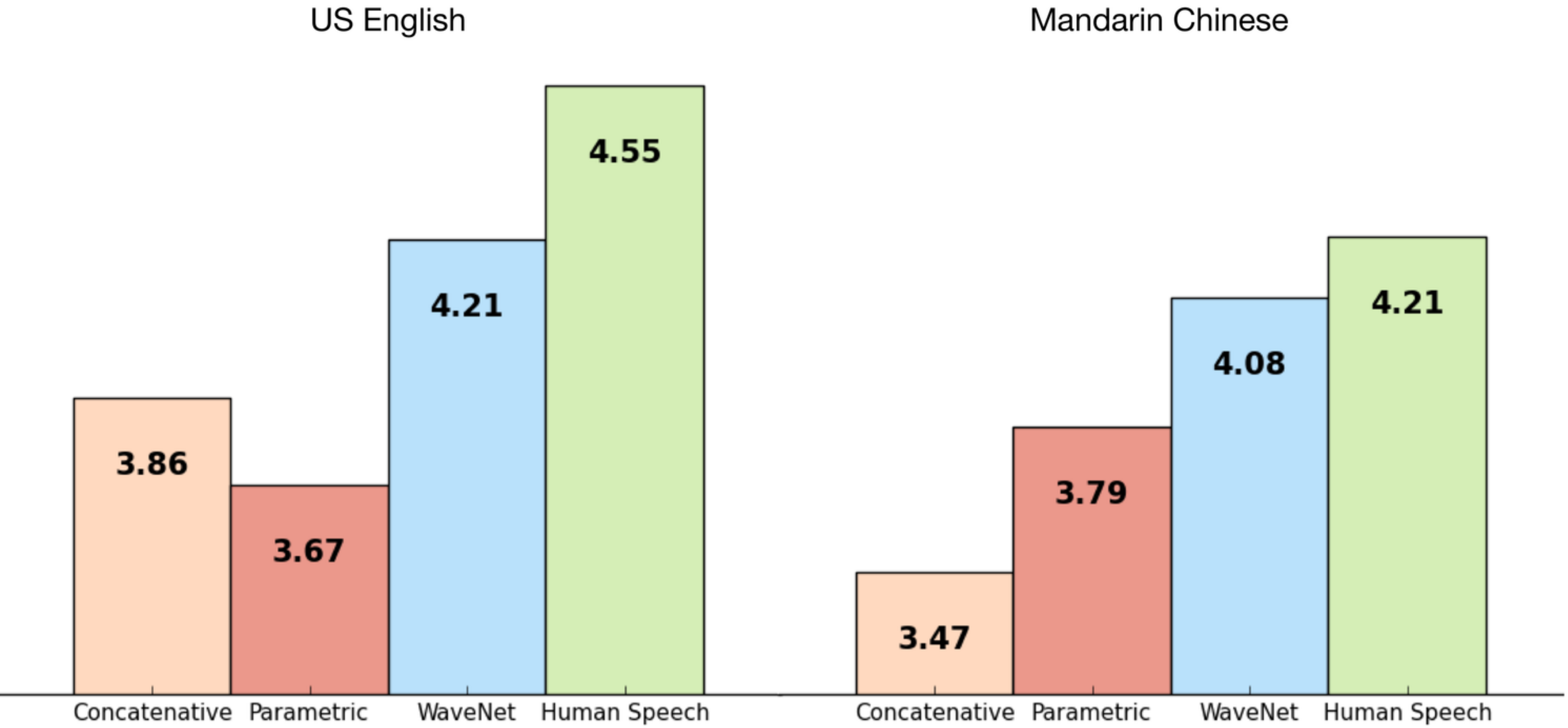
Text To Speech before Wavenet

- Concatenative

- Parametric



https://towardsdatascience.com/wavenet-google-assistants-voice-synthesizer-a168e9af13b1



What makes WaveNet so interesting?

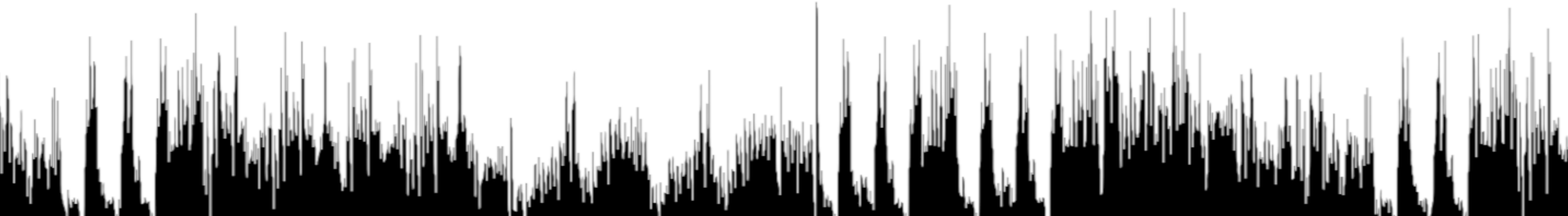
- Conditioning to different features:
  - Speech
  - Speaker
  - Music

Overview

- similar to PixelCNN

- based on previous data points

- Probability of successive data points



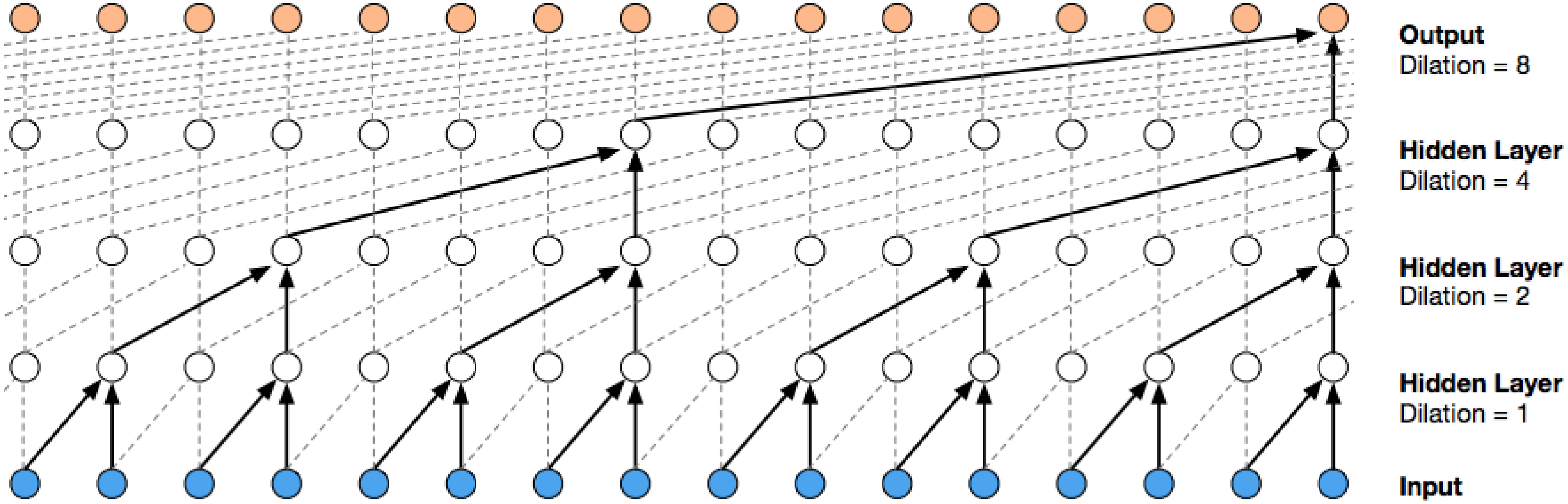
CNN

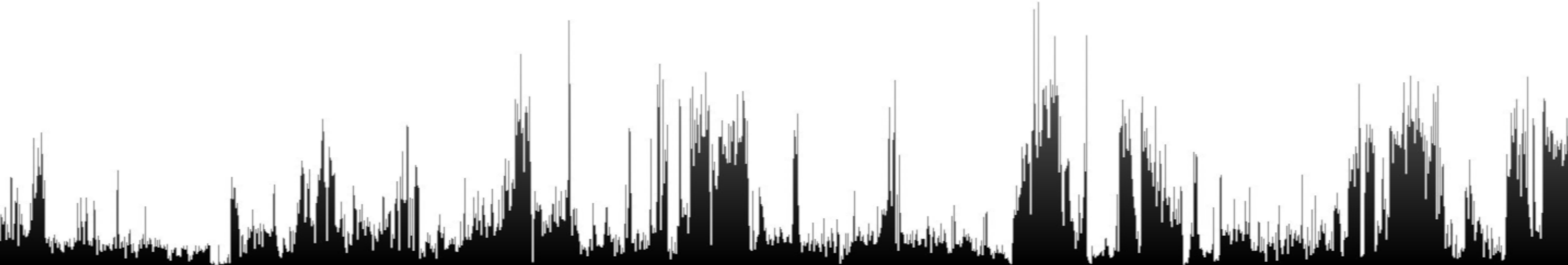
- faster trained then RNNs or LSTMs for 1-D Sequences

- Multiple Dilated causal Convolution layers stacked on top of each other

→ longer time dependencies

https://miro.medium.com/max/1145/1*9ig2pWTfqGcEBAgEUG12WA.png





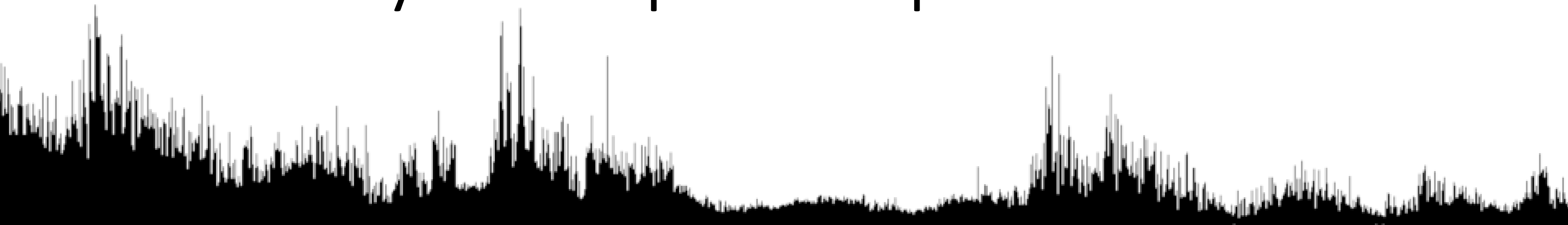
Softmax Distribution

- Categoric distribution

- Similar to Sigmoid

Problem:

- raw-audio is 16Bit (–32,768 to 32,767)
- For every timestep 65.535 possible Values



Mu-law

- Reduce bitdepth

- Logarithmic digitalization

Why?

- Problem with low amplitudes when rounding off
- more quantization steps at lower amplitudes

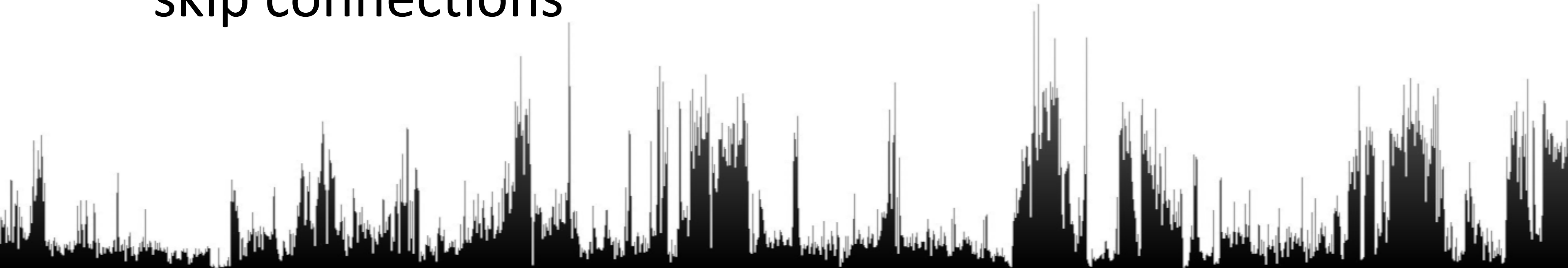
https://miro.medium.com/max/460/1*YT9ZtQ90dJavJzv7HwlXjw.png

Gated Activations

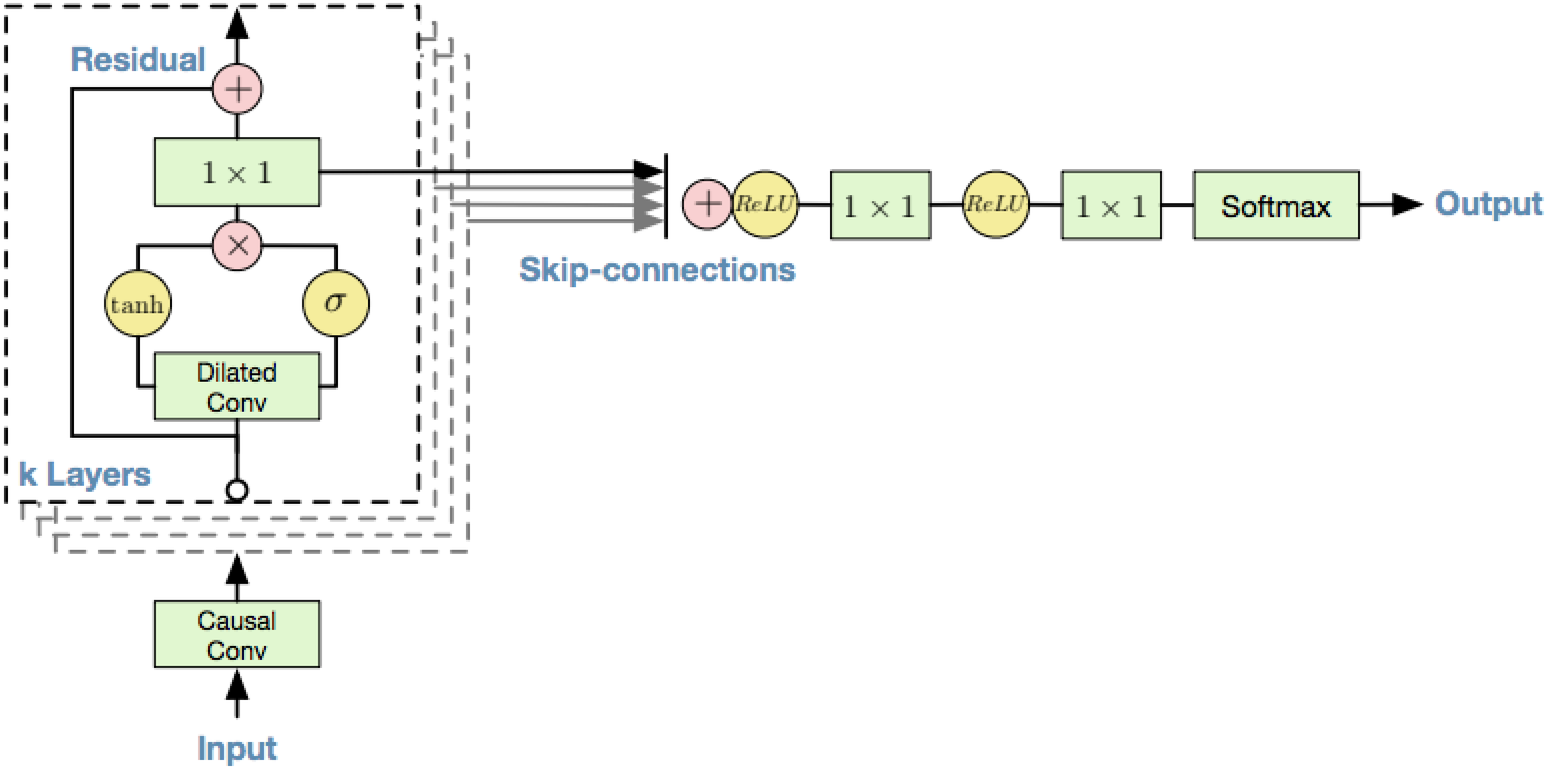
- previously Relu
- After experiments: tan-hyperbolic gated with Sigmoid-Activation works better

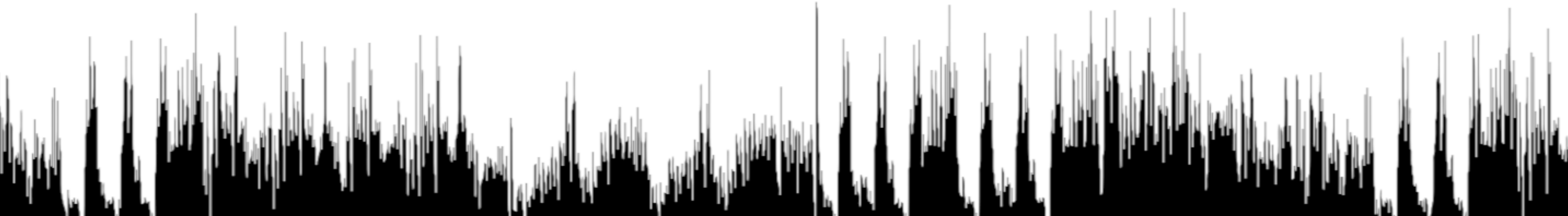
$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

- Reduction of Convergence time with residual and skip connections



https://miro.medium.com/max/899/1*pV-gxj6g-fJ9doWtKyvErw.png





Conditioning

- By additional Input variable

- Global
  - With one feature

  Speaker → Multi-Speaker Audio

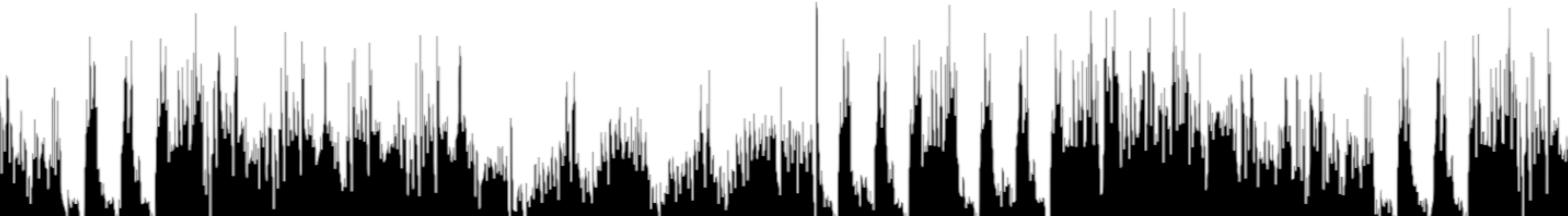
- Local
  - With multiple features

  Speech → Text to Speech

Testing WaveNet

- With Voice
  - After 12 hours of training: 🔊
  - After 3 days of training: 🔊

Batch-size: 1 Learning Rate: 0.0001



Testing WaveNet

- With Music
  - Violin 🔊
  - Piano 🔊

Batch-Size: 1-20

Learning Rate: 0.0001-1

Epochs: 10

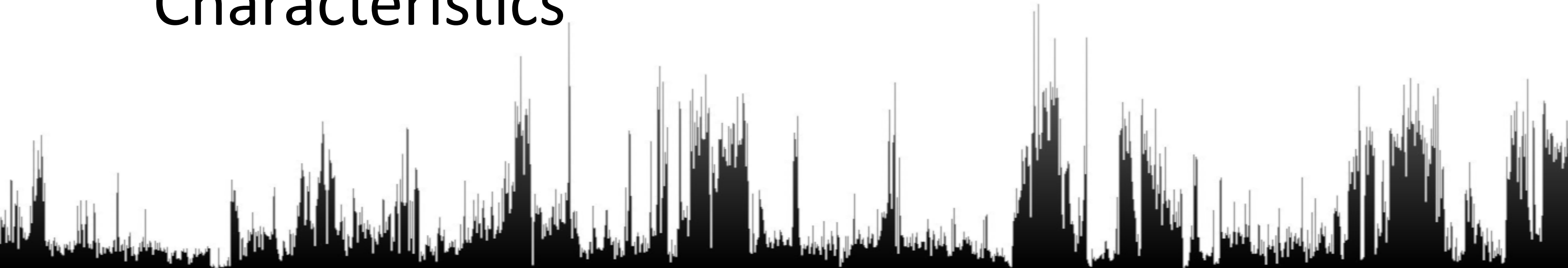
Layers: 10

How it can sound

- Examples from the Web
  - No Language 🔊
  - English 🔊
  - Mandarin 🔊
  - Piano 🔊

Summary

- Learning probability distributions
- With Data of the previous Timesteps
- And a Input variable to condition on different Characteristics



Sources

- https://towardsdatascience.com/wavenet-google-assistants-voice-synthesizer-a168e9af13b1

- https://github.com/vincentherrmann/pytorch-wavenet
- https://github.com/ibab/tensorflow-wavenet

- https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

- https://arxiv.org/pdf/1609.03499.pdf