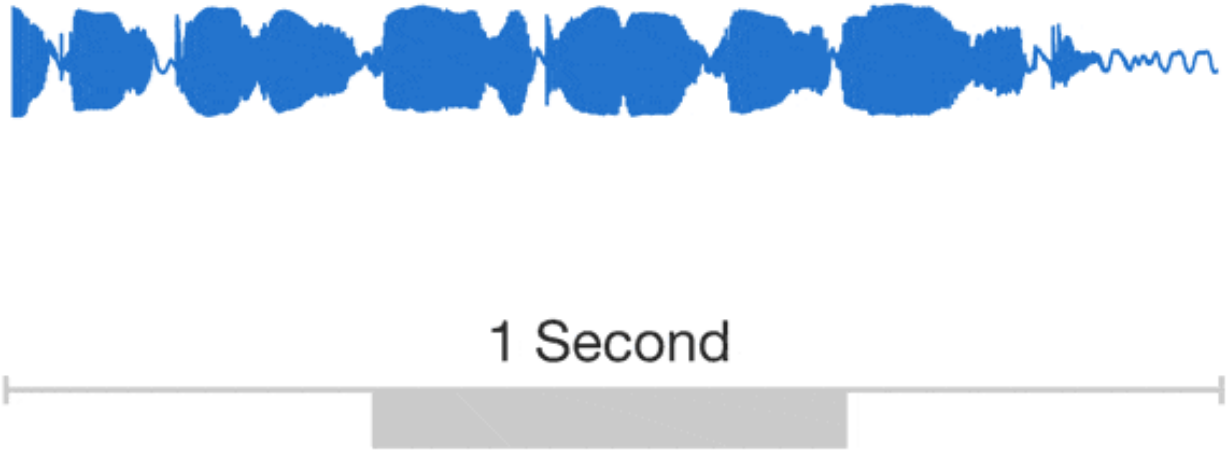
WaveNet

A generative model for raw audio

What we will talk about?

- WaveNet in a few words;
- Before WaveNet;
- Architecture (Based on PixelCNN);
- Dilated causal convolutions;
- More about architecture: softmax, GAU, residual&skip connections;
- Conditional WaveNets;
- Experiments: TTS, music&speech generation;
- Cons => development of WaveNet.

WaveNet in a few words



A second of generated speech

What is it?

Fully probabilistic & autoregressive model based on the PixelCNN architecture.

What it does?

Directly modelling the raw waveform of the audio signal, one sample at a time.

Why?

Naturalness of produced speech.

Before WaveNet

Concatenative TTS

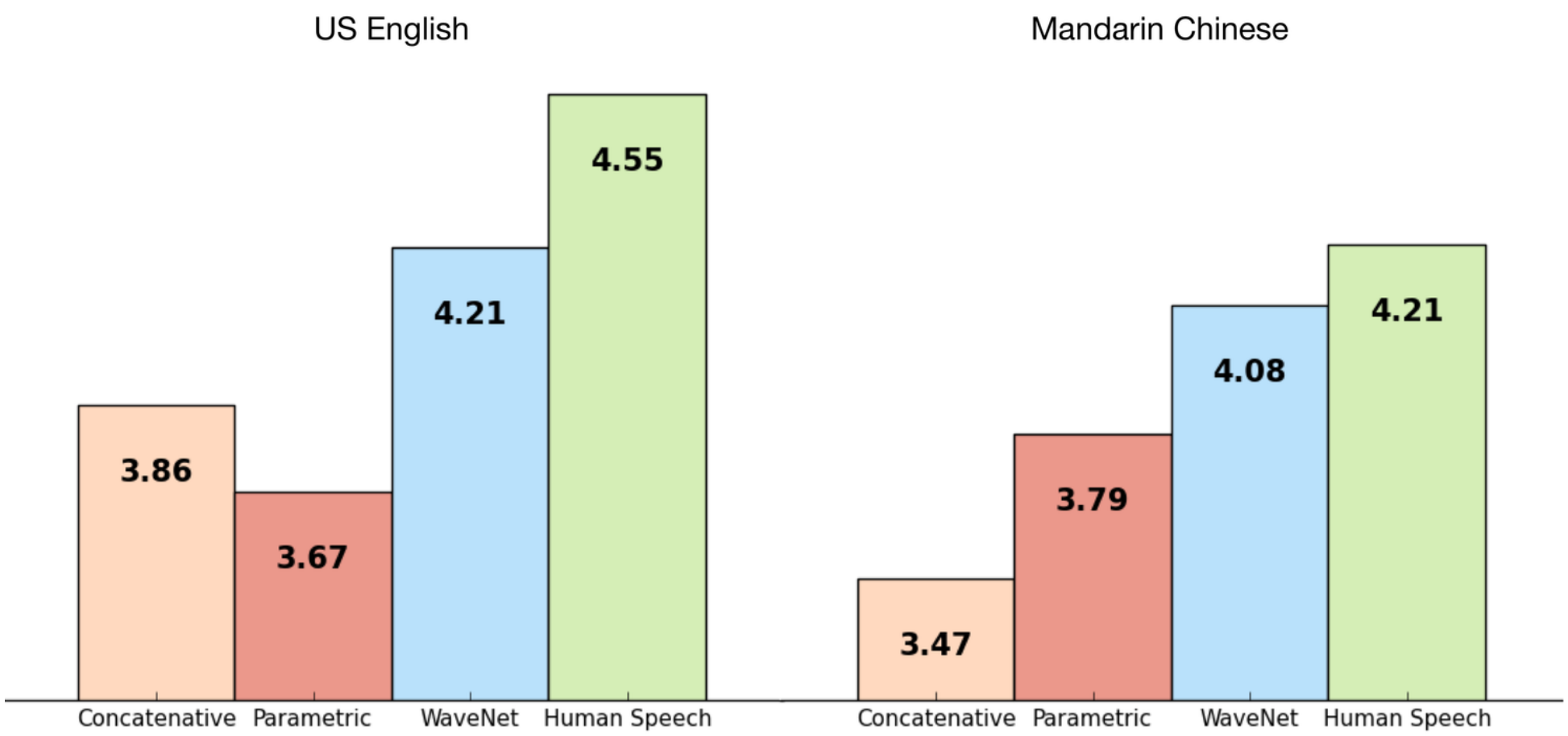
Very large database of short speech fragments are recorded from a single speaker and then recombined to form complete utterances.

Parametric TTS

All the information required to generate the data is stored in the parameters of the model, and the contents and characteristics of the speech can be controlled via the inputs to the model.

Let's compare:

WaveNet cooler? Let's observe MOS

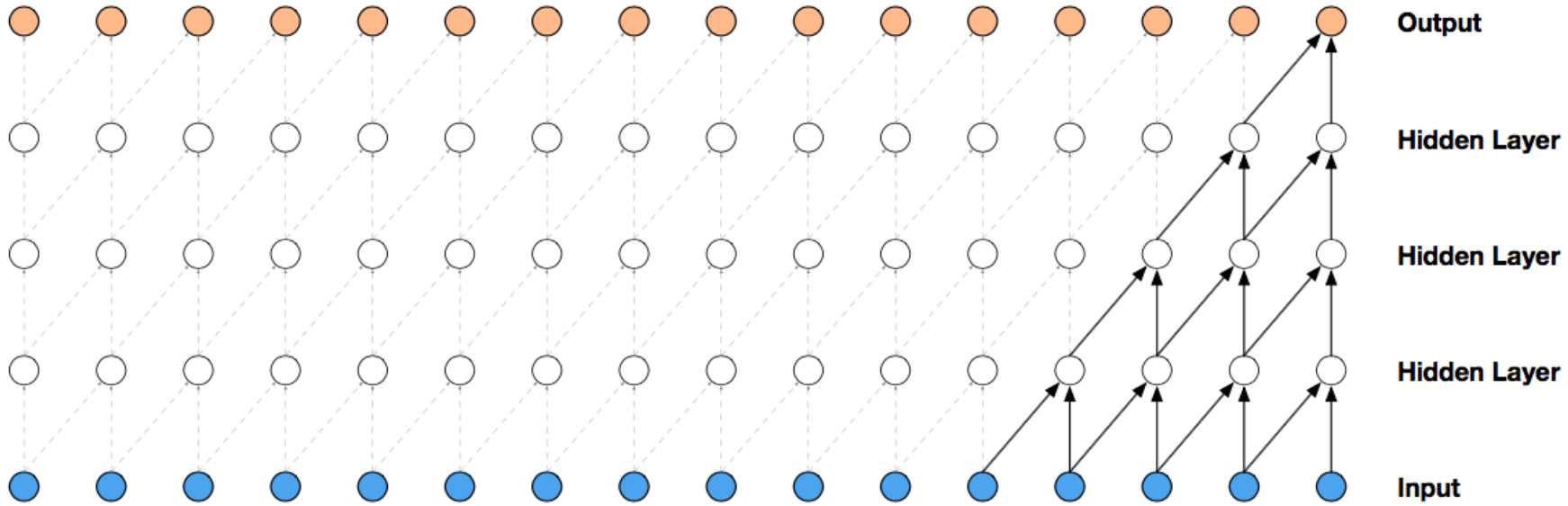


Architecture (Based on PixelCNN)

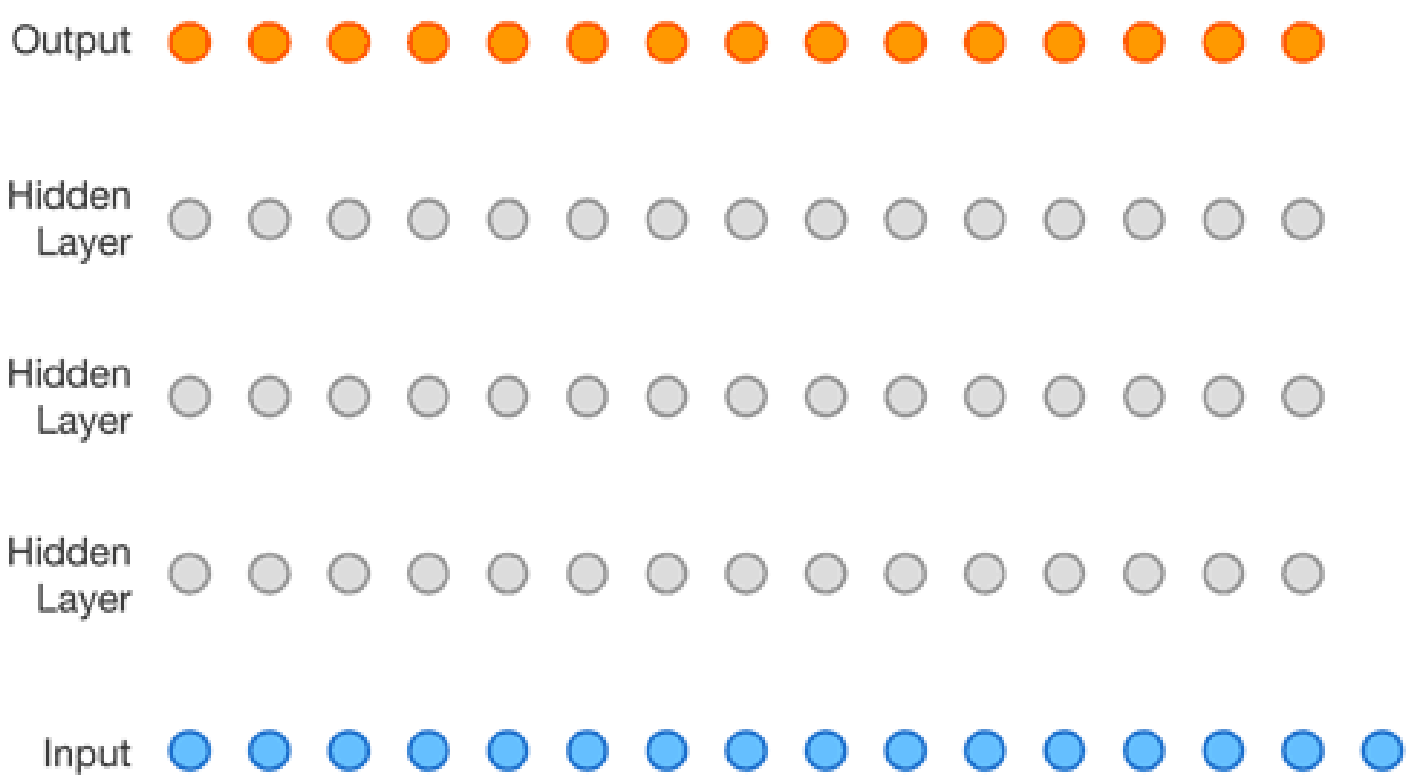
The joint probability of a waveform $x = \{x_1, \ldots, x_T\}$:

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

Each audio sample x_t is therefore conditioned on the samples at all previous timesteps

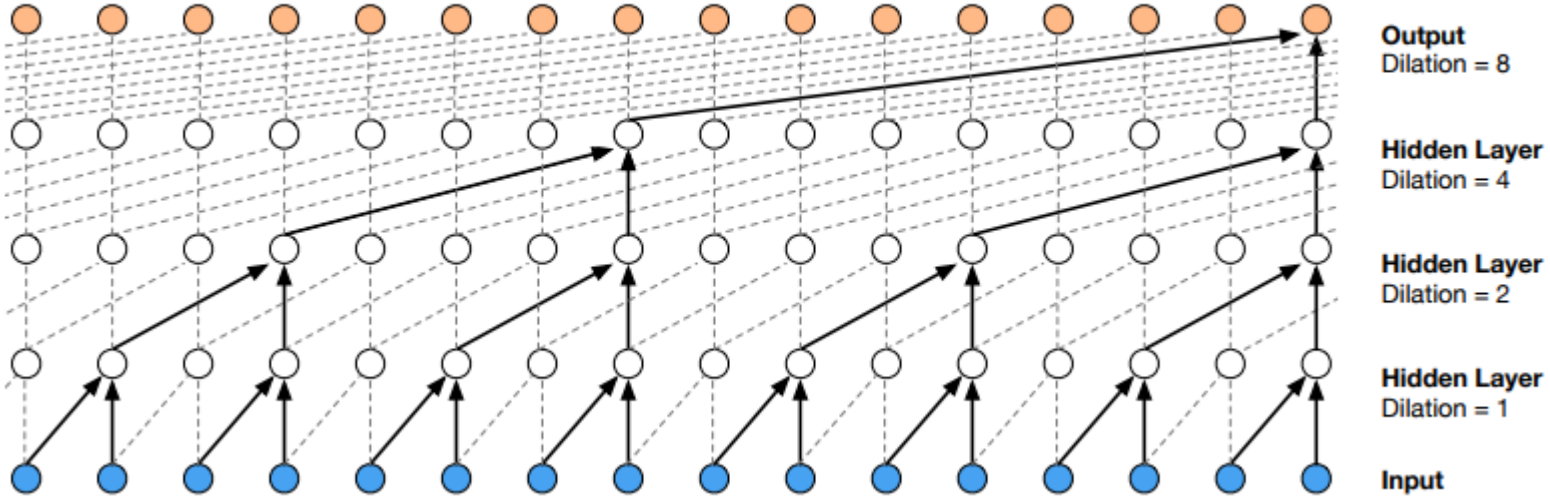


Dilated causal convolutions: what?



Dilated causal convolutions: why?

The reason we need such convolutions is receptive field size increase without greatly increasing computational cost:
A **dilated convolution** effectively allows the network to operate on a coarser scale than with a normal convolution.

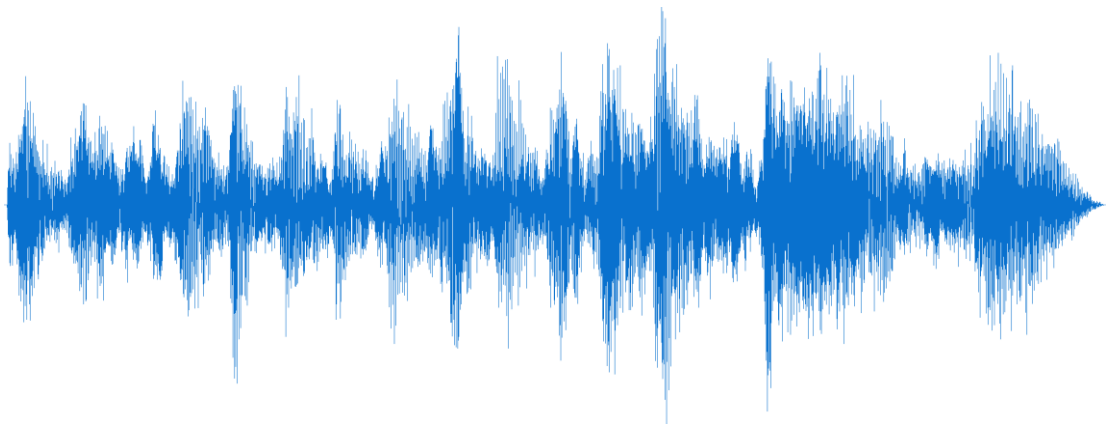


Stack of dilated causal convolutional layers

This paper:
1, 2, 4, . . . , 512,
1, 2, 4, . . . , 512,
1, 2, 4, . . . , 512.

Softmax distribution

$$p(x_t|x_1, \ldots, x_{t-1}) = \ \ldots\ MDN?\ MCGSM?\ Softmax\ distribution!$$



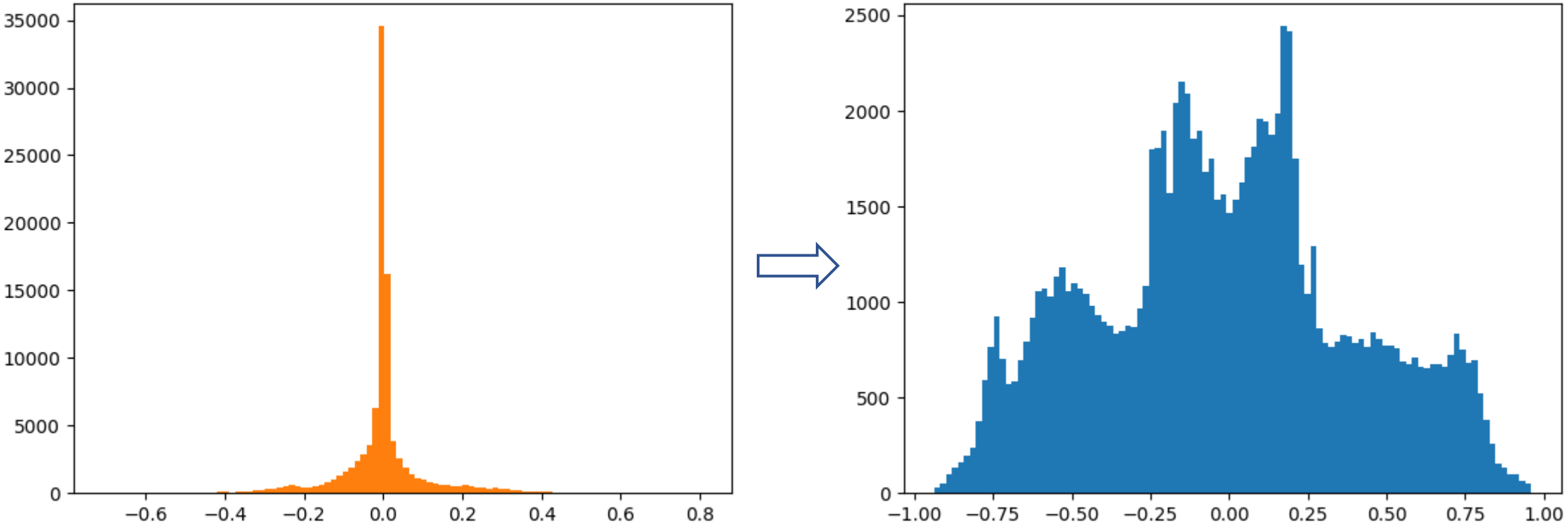
Sequence of 16-bit integer values
(one per timestep)

⇩

65,536 probabilities per timestep

$$f(x_t) = sign(x_t)\frac{\ln(1+\mu\,|x_t|)}{\ln(1+\mu)}, where\ -1 < x < 1\ and\ \mu = 255$$

μ – law visulization



Gated Activation Units

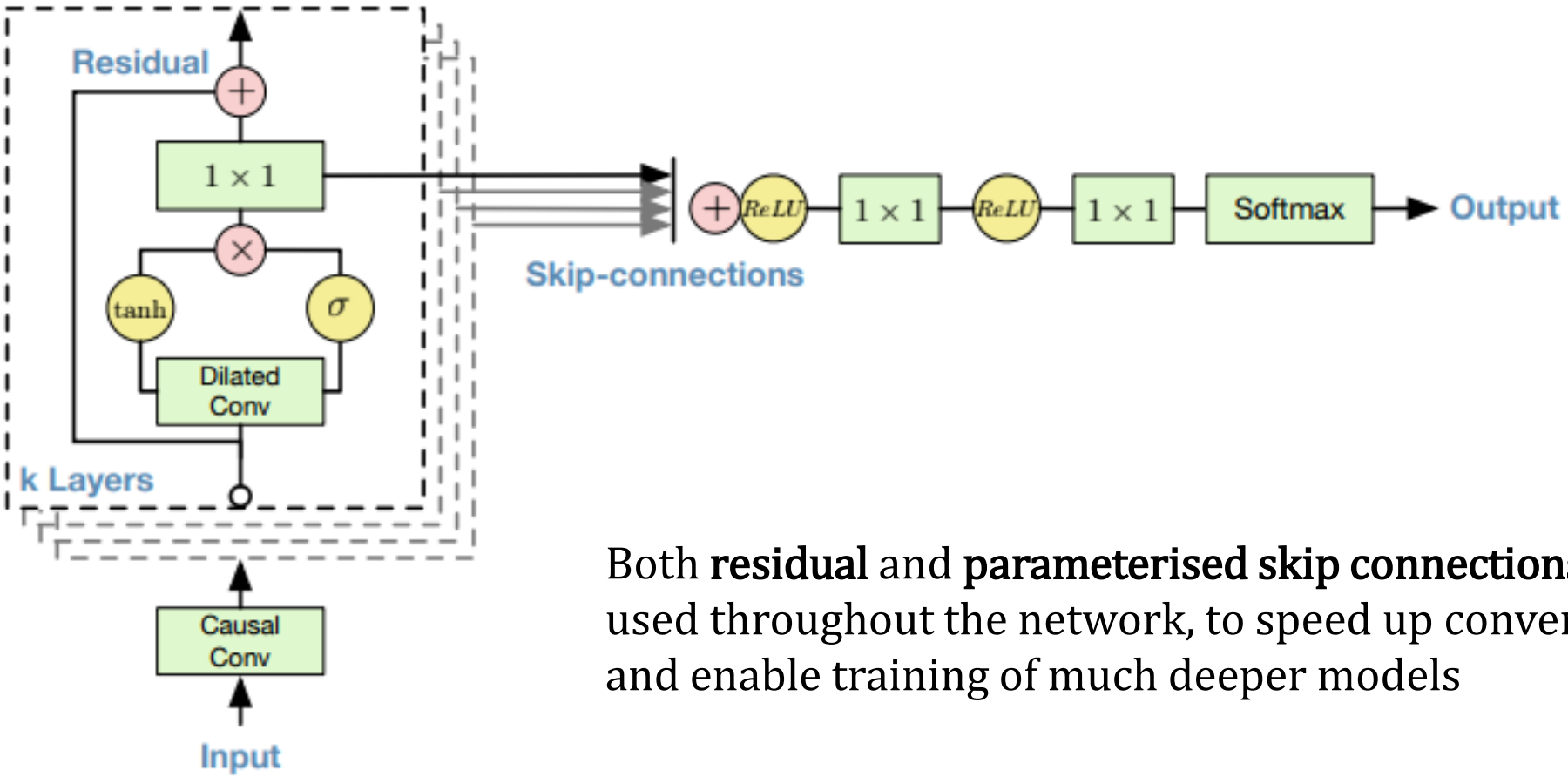
Same as used in gated PixelCNN:

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

$* - convolution\ operator,$
$\odot - element - wise\ multiplication\ operator,$
$\sigma(\cdot) - a\ sigmoid\ function,$
$k\ - layer\ index,$
$f\ and\ g\ - filter\ and\ gate\ respectively.$

Interesting detail: this non-linearity worked significantly better than the rectified linear activation function for modeling audio signals.

Architecture in one picture



Both **residual** and **parameterised skip connections** are used throughout the network, to speed up convergence and enable training of much deeper models

Conditional WaveNets

WaveNets can model the conditional distribution $p(x|h)$ of the audio given the input h – f.e. we can choose the speaker by feeding the speaker identity to the model as an extra input

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

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

$V_{*,k}$ $-$ *a learnable linear projection*

$V_{*,k}^T h$ $-$ *broadcast over the time dimension*

Global conditioning: single latent representation h

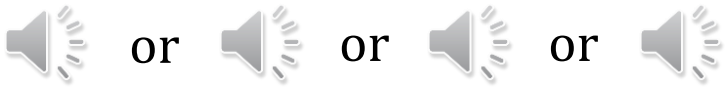
$$z = tanh\left(W_{f,k} * x + V_{f,k} y\right) \odot \sigma\left(W_{f,k} * x + V_{g,k} y\right),$$

$V_{*,k} y$ $-$ 1×1 *convolution*

Local conditioning: second timeseries $h_t \Rightarrow y = f(h)$

Multispeaker speech generator results

Predictions conditioned only on the previous audio samples (not on a text) = bla-bla

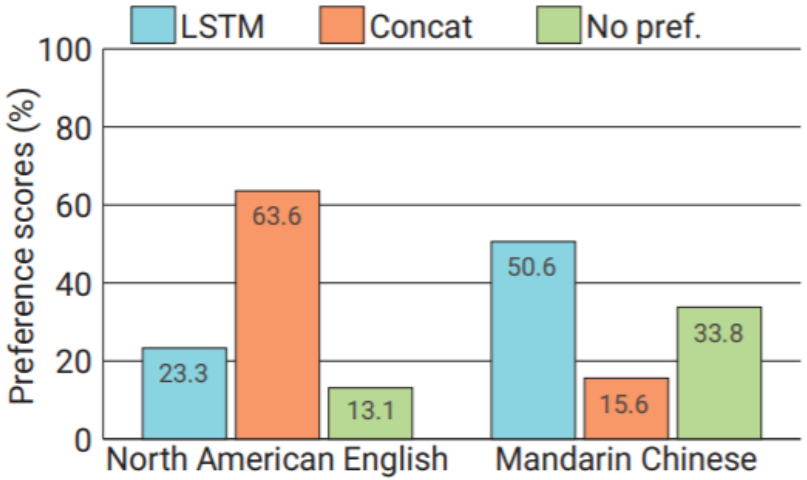
Changing identity of a speaker:  or or or

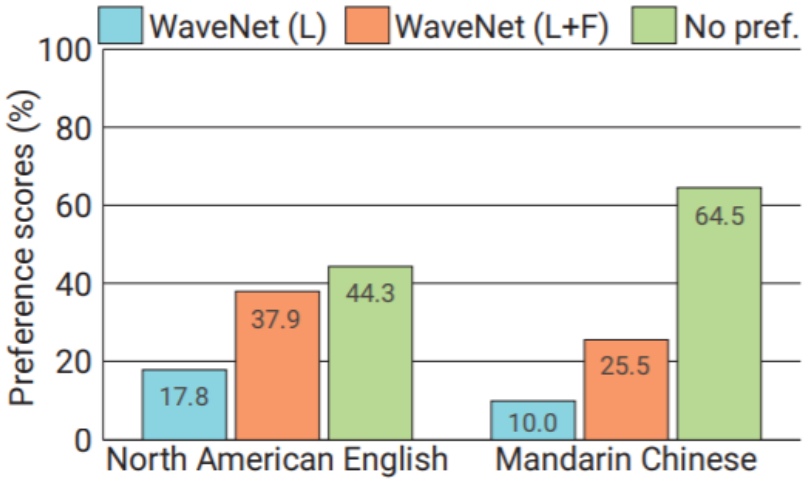
+ emotions

+ accents

Interesting detail: adding speakers resulted in better validation set performance compared to training solely on a single speaker

TTS results





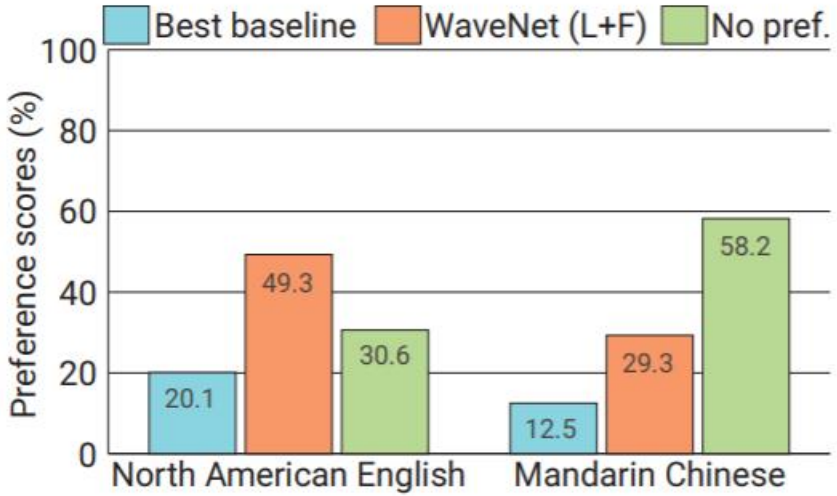


Figure 5: Subjective preference scores (%) of speech samples between (top) two baselines, (middle) two WaveNets, and (bottom) the best baseline and WaveNet. Note that LSTM and Concat correspond to LSTM-RNN-based statistical parametric and HMM-driven unit selection concatenative baseline synthesizers, and WaveNet (L) and WaveNet (L+F) correspond to the WaveNet conditioned on linguistic features only and that conditioned on both linguistic features and $\log F_0$ values.

Music generation results

YouTube piano dataset + MagnaTagATune dataset

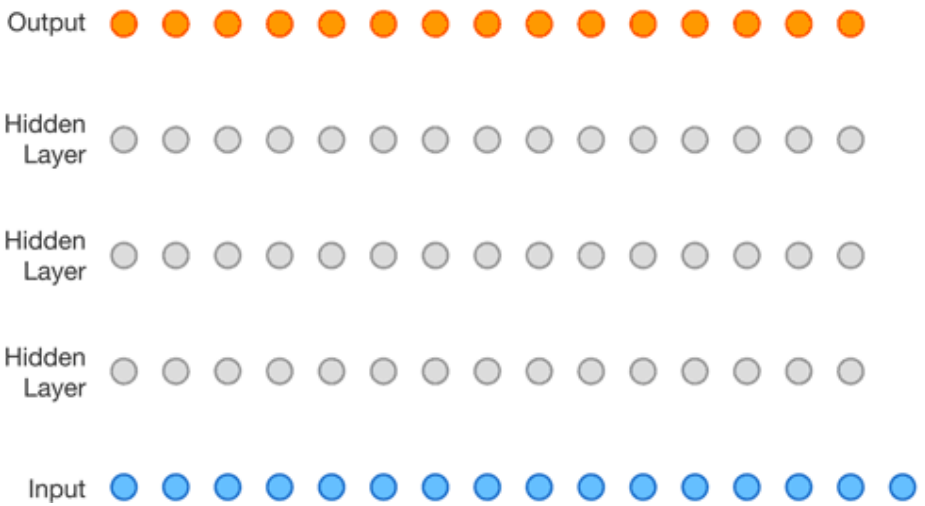
Interesting detail: enlarging the receptive field was crucial to obtain samples that sounded musical. Even with a receptive field of several seconds, the models did not enforce long-range consistency which resulted in second-to-second variations in genre, instrumentation, volume and sound quality

Cons of WaveNet

- Ok, Google, I feel so alone that I don't want to live, what should I do?

WaveNet produces high-quality audio with up to 24,000 samples per second

Hours later…



-  Sounded perfectly natural ;)

Any developments? Parallel WaveNet

Main idea:

Student

Teacher

> 1000 times faster!

References

- https://arxiv.org/pdf/1609.03499.pdf;
- https://deepmind.com/blog/wavenet-generative-model-raw-audio;
- http://www.inference.vc/dilated-convolutions-and-kronecker-factorisation;
- http://sergeiturukin.com/2017/03/02/wavenet.html;
- https://deepmind.com/blog/high-fidelity-speech-synthesis-wavenet;
- https://arxiv.org/pdf/1711.10433.pdf.