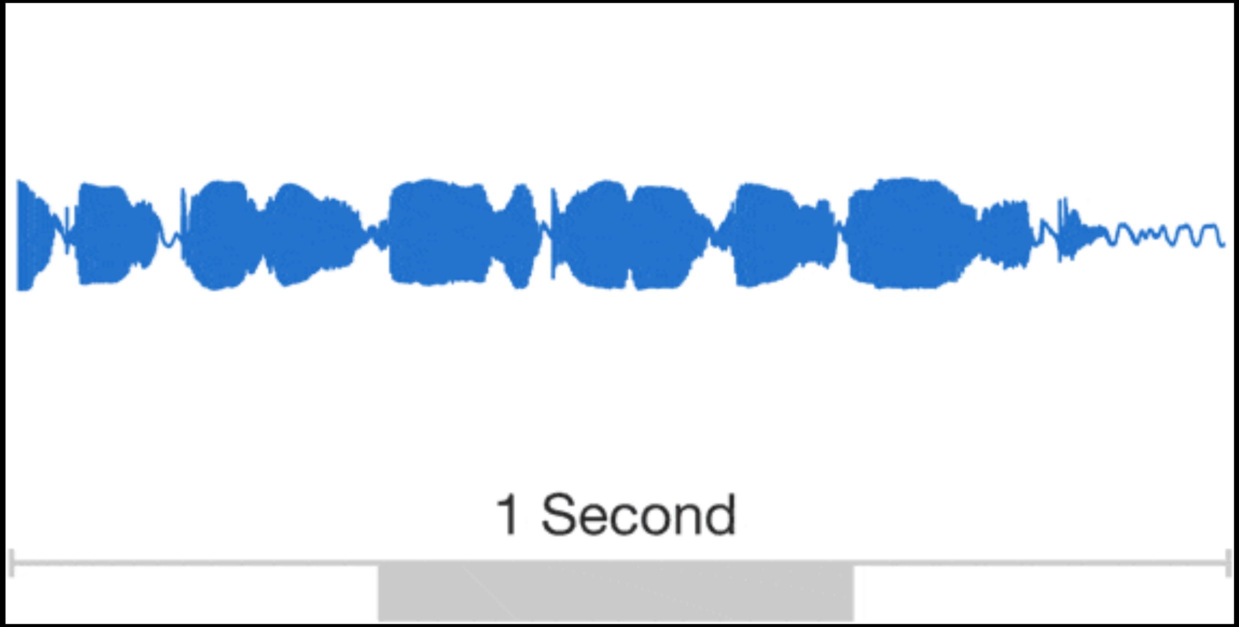
Waveform-based deep learning research at Dolby

Jordi Pons (@jordiponsdotme / www.jordipons.me)
Representing the recent work of the Applied AI team!

Challenges of Deep Learning in Audio

HIGH DIMENTIONALITY

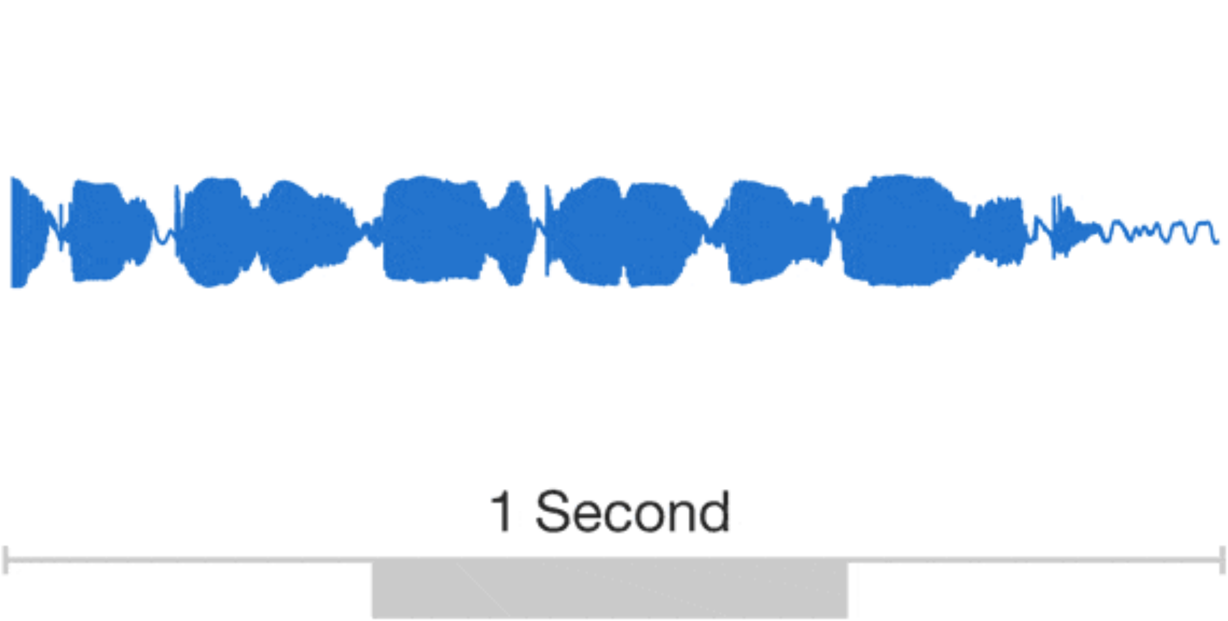


Animation from:
https://deepmind.com/blog/wavenet-generative-model-raw-audio

Challenges of Deep Learning in Audio

HIGH DIMENTIONALITY

MULTI-LEVEL TEMPORAL DEPENDANCY



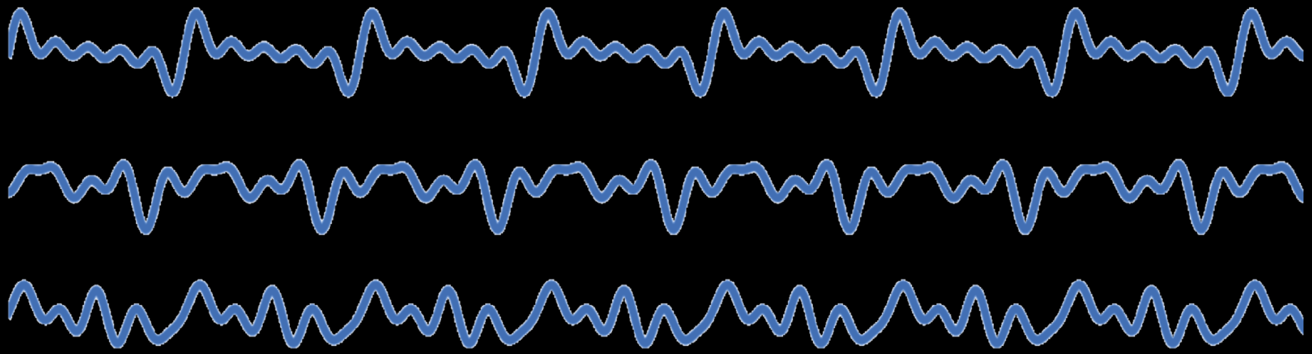
Animation from:
https://deepmind.com/blog/wavenet-generative-model-raw-audio

Challenges of Deep Learning in Audio

HIGH DIMENTIONALITY

MULTI LEVEL TEMPORAL DEPENDANCY

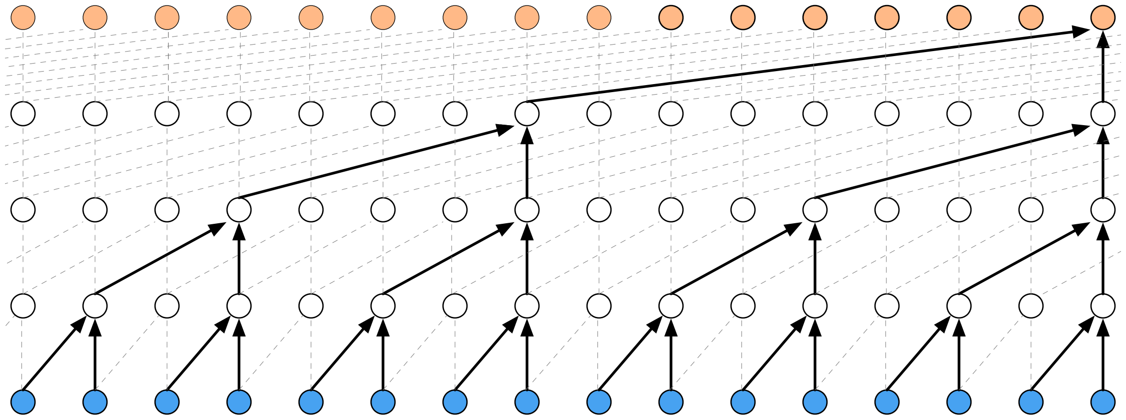
PERCEPTION MATTERS



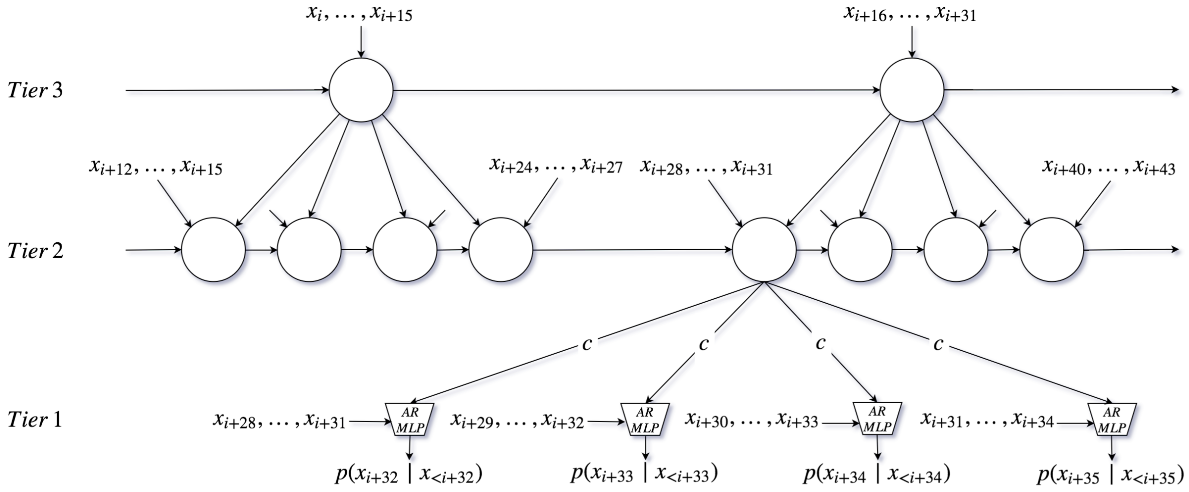
Which of these waves sound different?

Image by Jesse Engel - Problems with WaveNet (DAFx 2019)

Generating Audio Waveforms

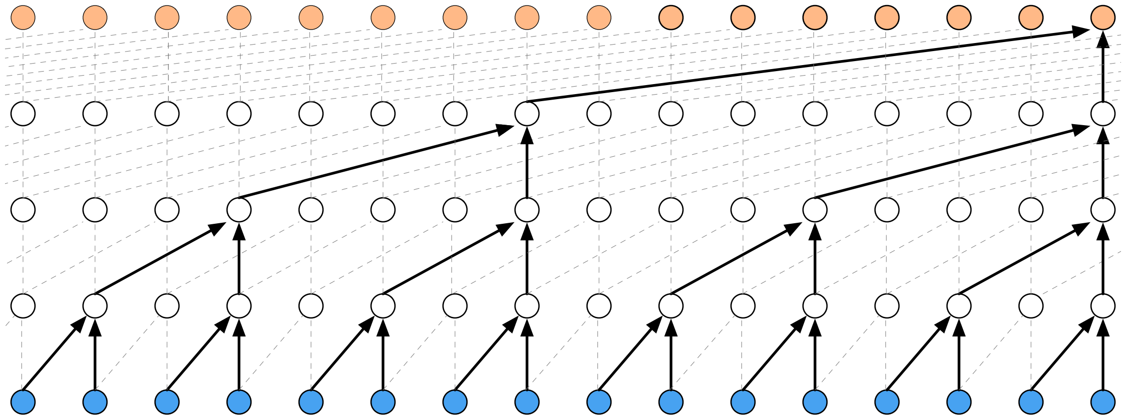


WaveNet: A generative model for raw audio (Google DeepMind)

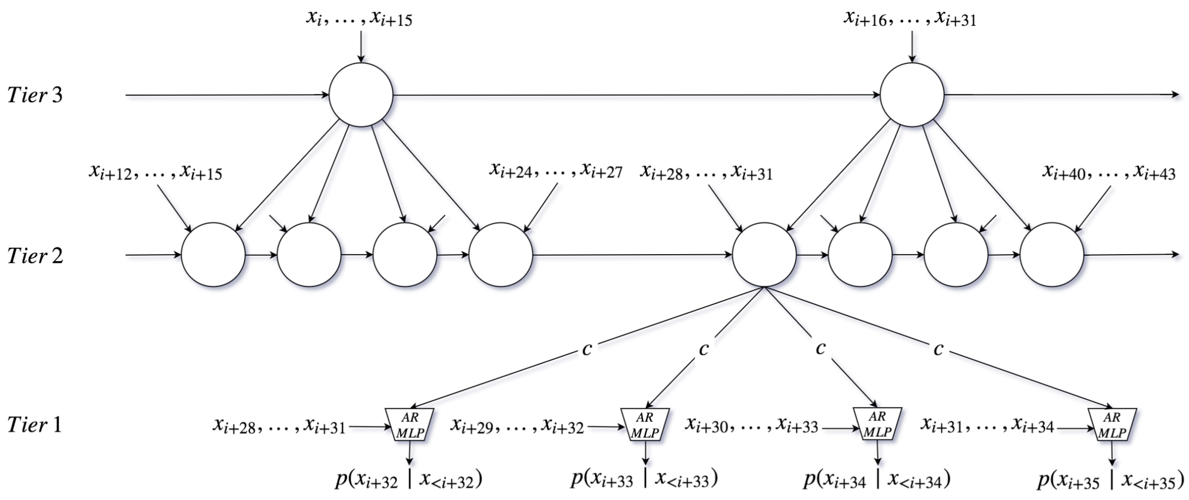

$x_i, \ldots, x_{i+15}$
$x_{i+16}, \ldots, x_{i+31}$
Tier 3
$x_{i+12}, \ldots, x_{i+15}$
$x_{i+24}, \ldots, x_{i+27}$
$x_{i+28}, \ldots, x_{i+31}$
$x_{i+40}, \ldots, x_{i+43}$
Tier 2
c
c
c
c
Tier 1
$x_{i+28}, \ldots, x_{i+31}$
AR MLP
$x_{i+29}, \ldots, x_{i+32}$
AR MLP
$x_{i+30}, \ldots, x_{i+33}$
AR MLP
$x_{i+31}, \ldots, x_{i+34}$
AR MLP
$p(x_{i+32} \mid x_{<i+32})$
$p(x_{i+33} \mid x_{<i+33})$
$p(x_{i+34} \mid x_{<i+34})$
$p(x_{i+35} \mid x_{<i+35})$

SampleRNN: Multirate RNN based generative model (MILA)

Generating Audio Waveforms



WaveNet: A generative model for raw audio (Google DeepMind)

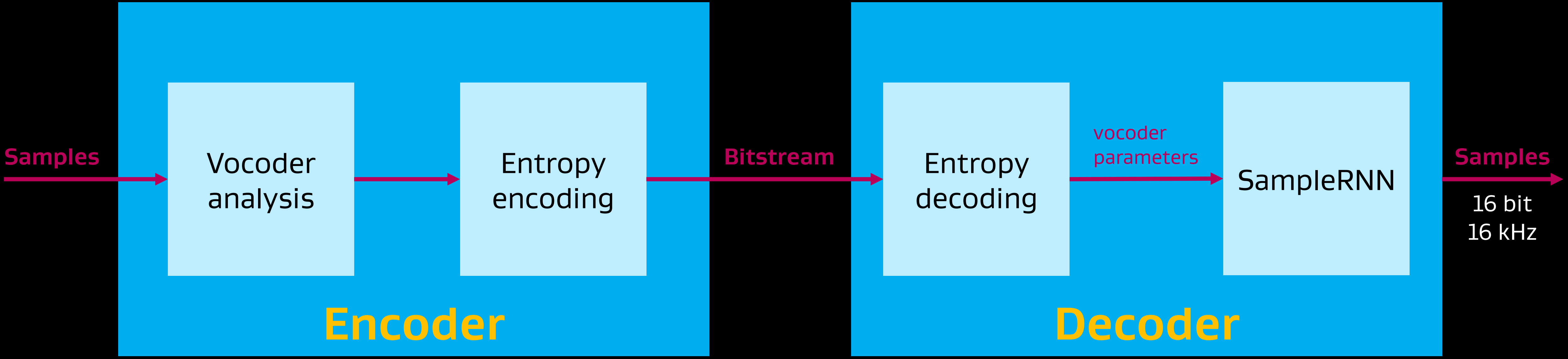

$x_i, \ldots, x_{i+15}$
$x_{i+16}, \ldots, x_{i+31}$
Tier 3
$x_{i+12}, \ldots, x_{i+15}$
$x_{i+24}, \ldots, x_{i+27}$
$x_{i+28}, \ldots, x_{i+31}$
$x_{i+40}, \ldots, x_{i+43}$
Tier 2
c
c
c
c
Tier 1
$x_{i+28}, \ldots, x_{i+31}$
AR MLP
$x_{i+29}, \ldots, x_{i+32}$
AR MLP
$x_{i+30}, \ldots, x_{i+33}$
AR MLP
$x_{i+31}, \ldots, x_{i+34}$
AR MLP
$p(x_{i+32} \mid x_{<i+32})$
$p(x_{i+33} \mid x_{<i+33})$
$p(x_{i+34} \mid x_{<i+34})$
$p(x_{i+35} \mid x_{<i+35})$

SampleRNN: Multirate RNN based generative model (MILA)

Low Bitrate Speech Coding

- Wavenet Based Low Rate Speech Coding (Google) *W. Bastiaan Kleijn, Felicia S. C. Lim, Alejandro Luebs, Jan Skoglund, Florian Stimberg, Quan Wang, Thomas C. Walters*
- High-quality speech coding with SampleRNN (Dolby) *Janusz Klejsa, Per Hedelin, Cong Zhou, Roy Fejgin, Lars Villemoes*

Coding Scheme



What’s in the conditioning?

Quantized vocoder parameters:

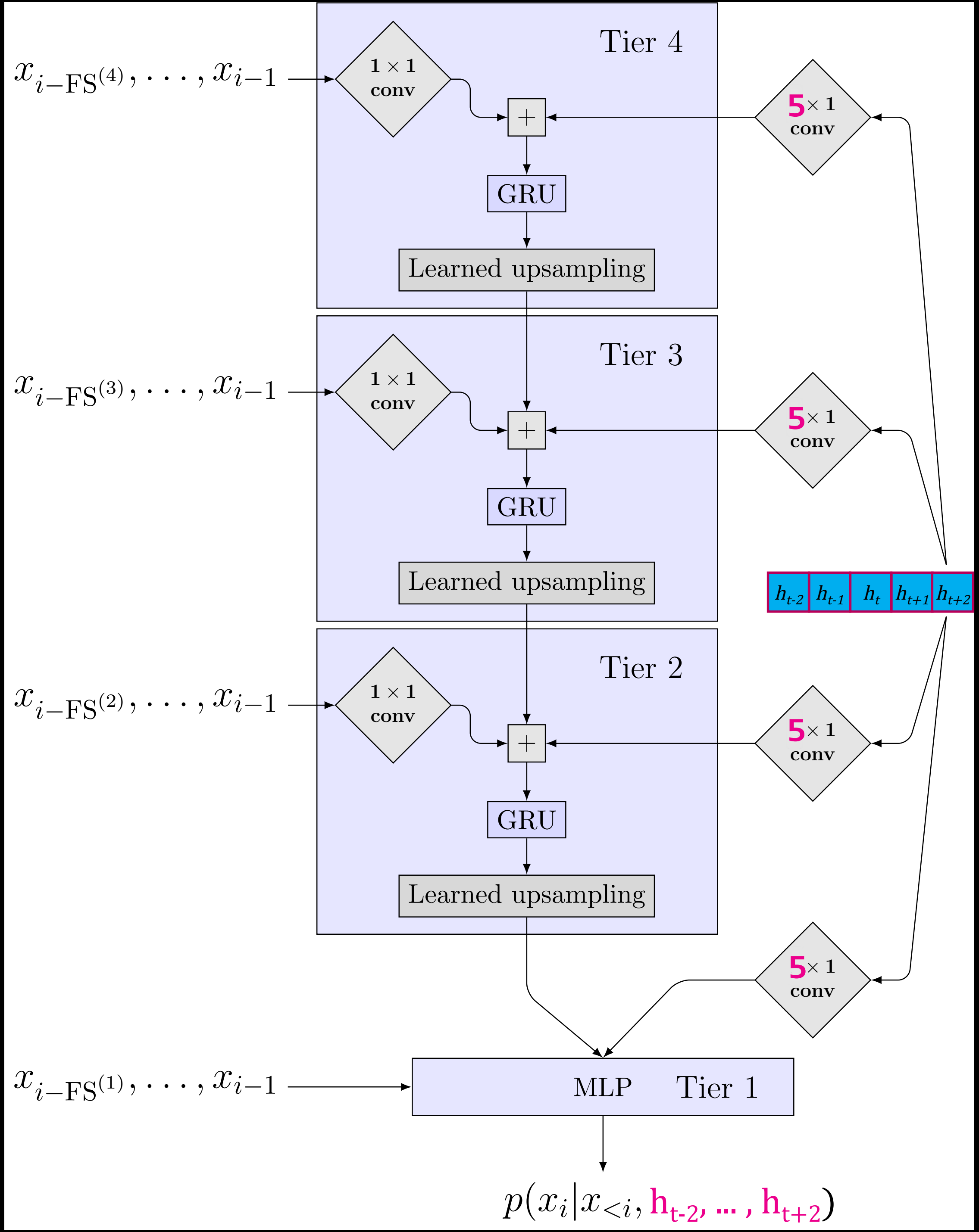
- Pitch
- LPC filter coefficients
- RMS level of residual
- Voicing level per band (6 bands)

The vocoder is based on:

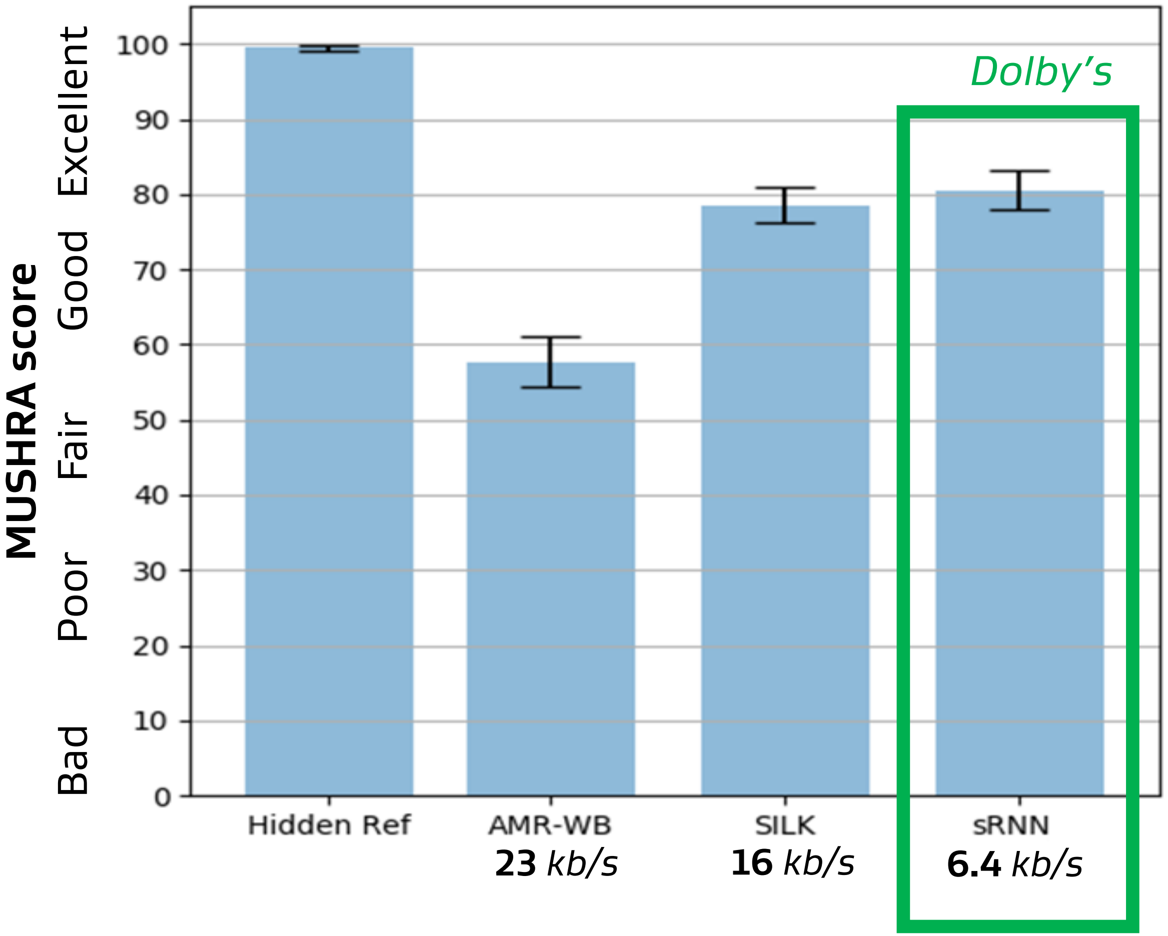
Per Hedelin, “A sinusoidal LPC vocoder,” in 2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421), Sept 2000, pp. 2–4

Conditional SampleRNN

- By itself, SampleRNN can only ‘babble’
  ⟶ we need conditioning
- 4-tier configuration
- Conditioning with lookahead

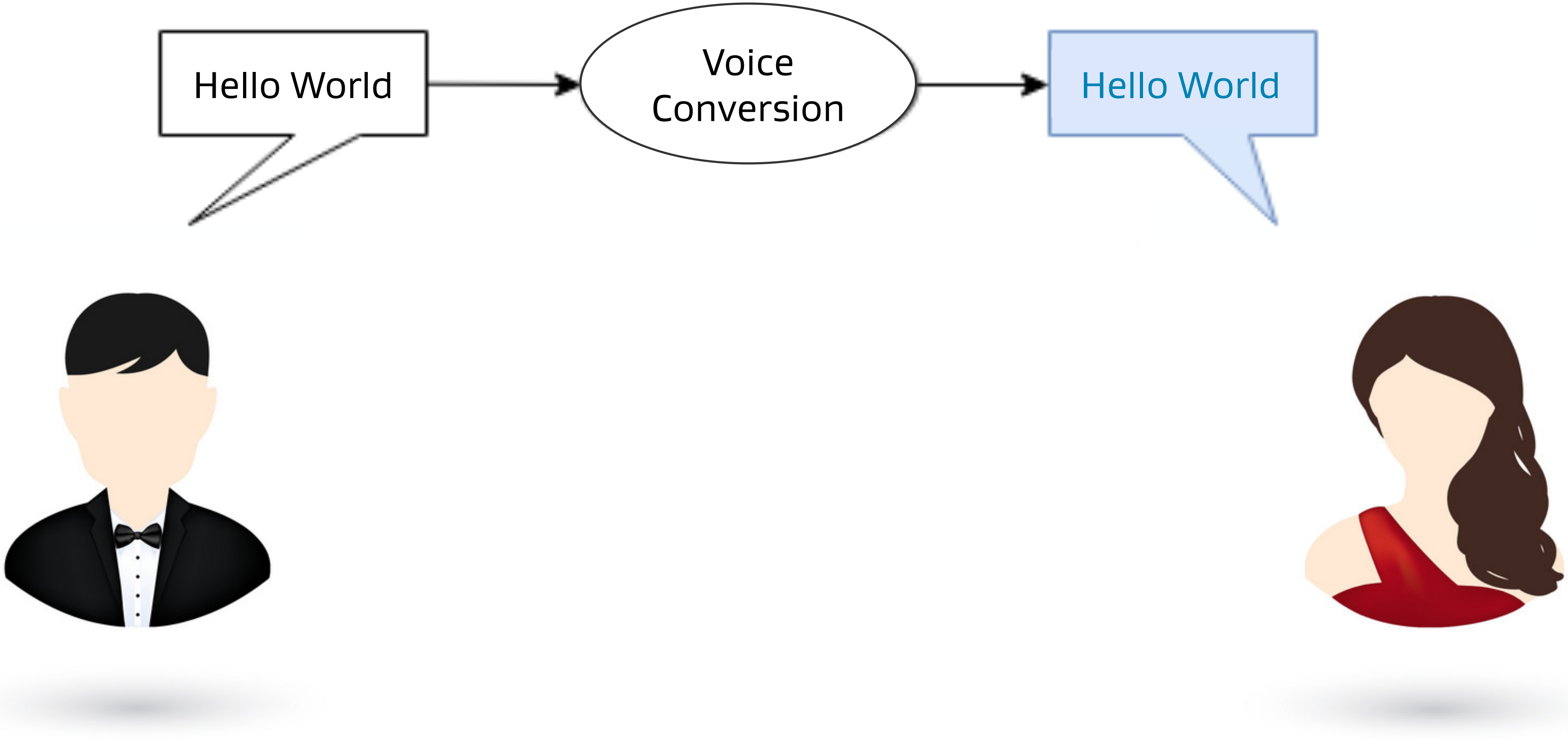

Tier 4
$x_{i-\mathrm{FS}^{(4)}}, \ldots, x_{i-1}$
1 × 1 conv
+
5 × 1 conv
GRU
Learned upsampling
Tier 3
$x_{i-\mathrm{FS}^{(3)}}, \ldots, x_{i-1}$
1 × 1 conv
+
5 × 1 conv
GRU
Learned upsampling
h_{t-2} h_{t-1} h_t h_{t+1} h_{t+2}
Tier 2
$x_{i-\mathrm{FS}^{(2)}}, \ldots, x_{i-1}$
1 × 1 conv
+
5 × 1 conv
GRU
Learned upsampling
5 × 1 conv
$x_{i-\mathrm{FS}^{(1)}}, \ldots, x_{i-1}$
MLP Tier 1
$p(x_i | x_{<i}, h_{t-2}, \ldots, h_{t+2})$

Listening Tests



High quality speech at 2.5x lower bitrate than SOTA codecs

Voice Conversion

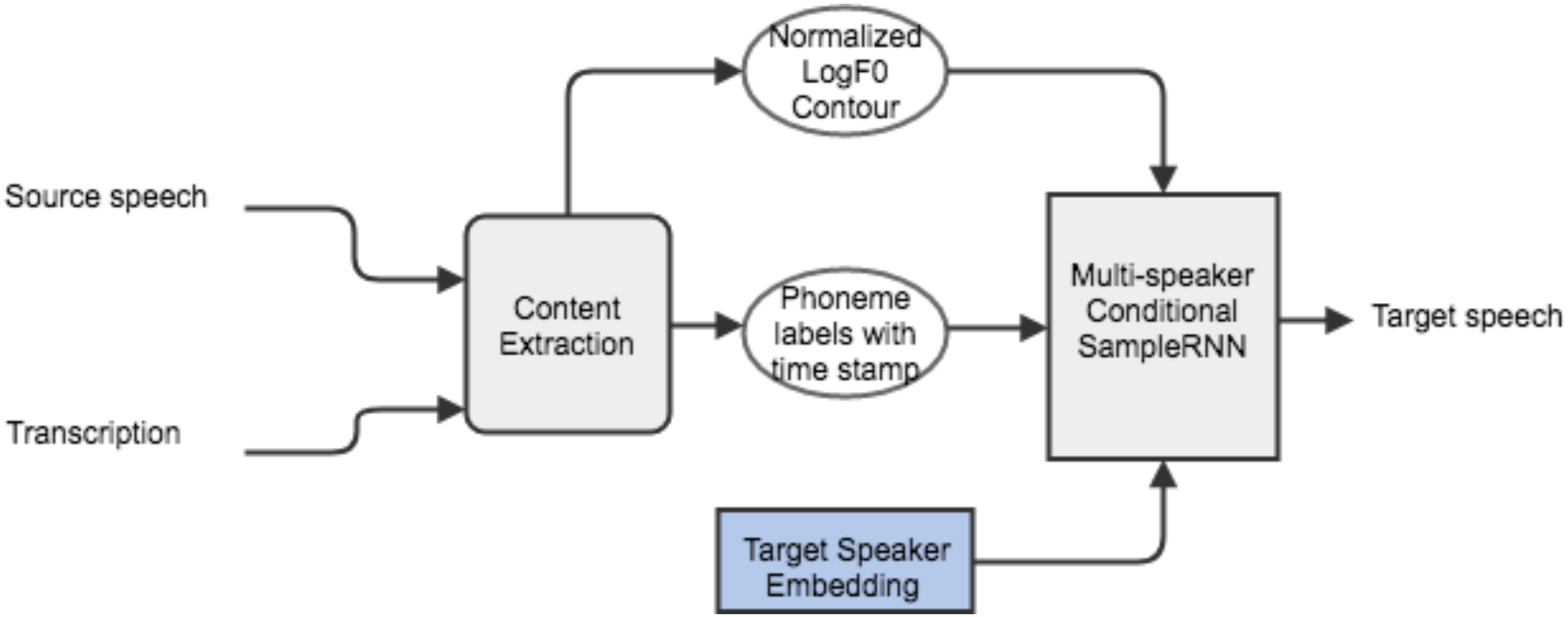


Hello World

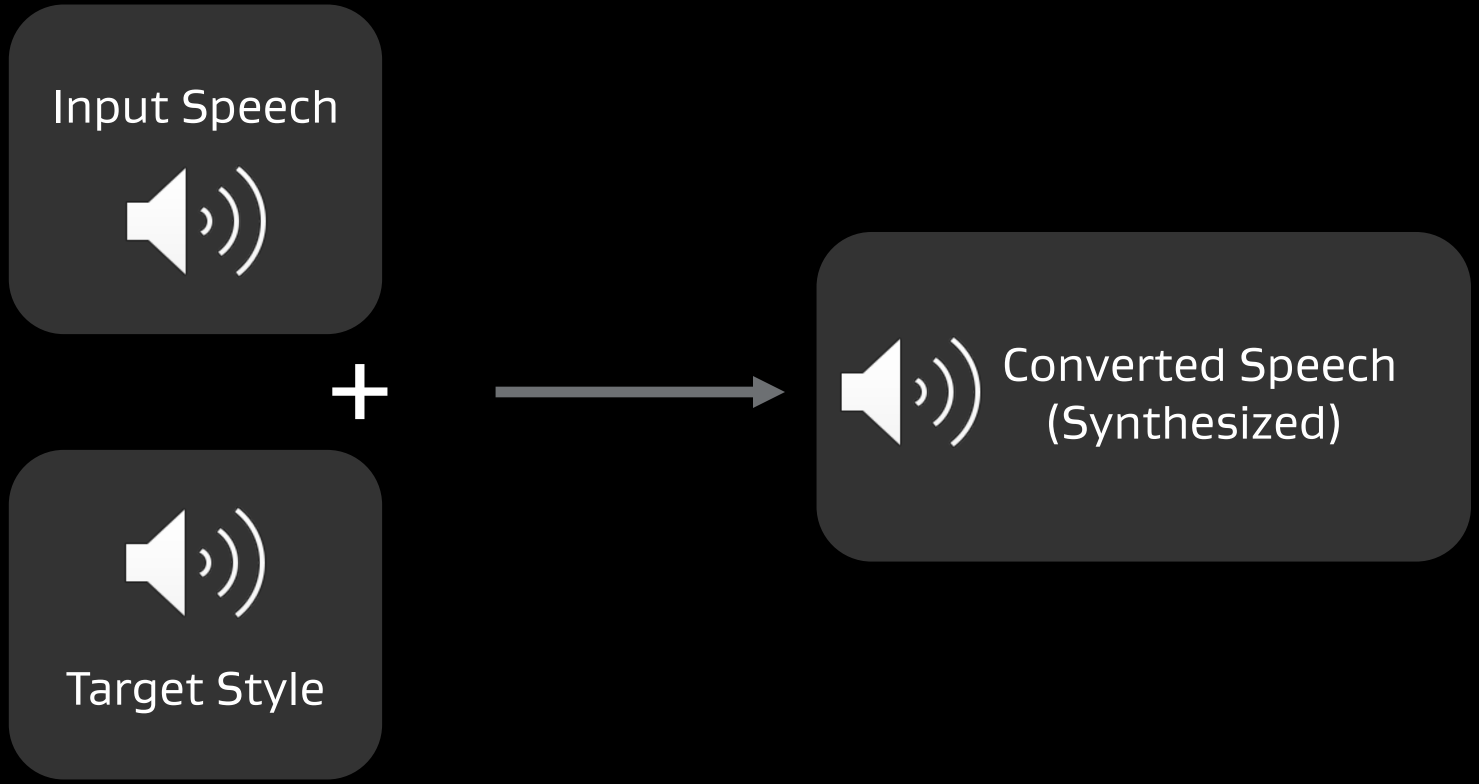
Voice Conversion

Hello World

Voice Conversion



Voice Conversion : Demo



Input Speech

+

Target Style

Converted Speech (Synthesized)

End-to-end Learning Audio Research

Voice Conversion with Conditional SampleRNN @ Interspeech 2018

Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, Dan Darcy

High-quality speech coding with SampleRNN @ ICASSP 2019

Janusz Klejsa, Per Hedelin, Cong Zhou, Roy Fejgin, Lars Villemoes

Jordi Pons (jpons@dolby.com /@jordiponsdotme / www.jordipons.me)