i-synth

Kuan Yu	kuanyu@uni-potsdam.de
Simon Untergasser	untergasser@uni-potsdam.de
Jörg Schwartz	jschwartz@uni-potsdam.de

July 18, 2018

outline

- data
- model
- status
- plan

LJSpeech-1.0[1]

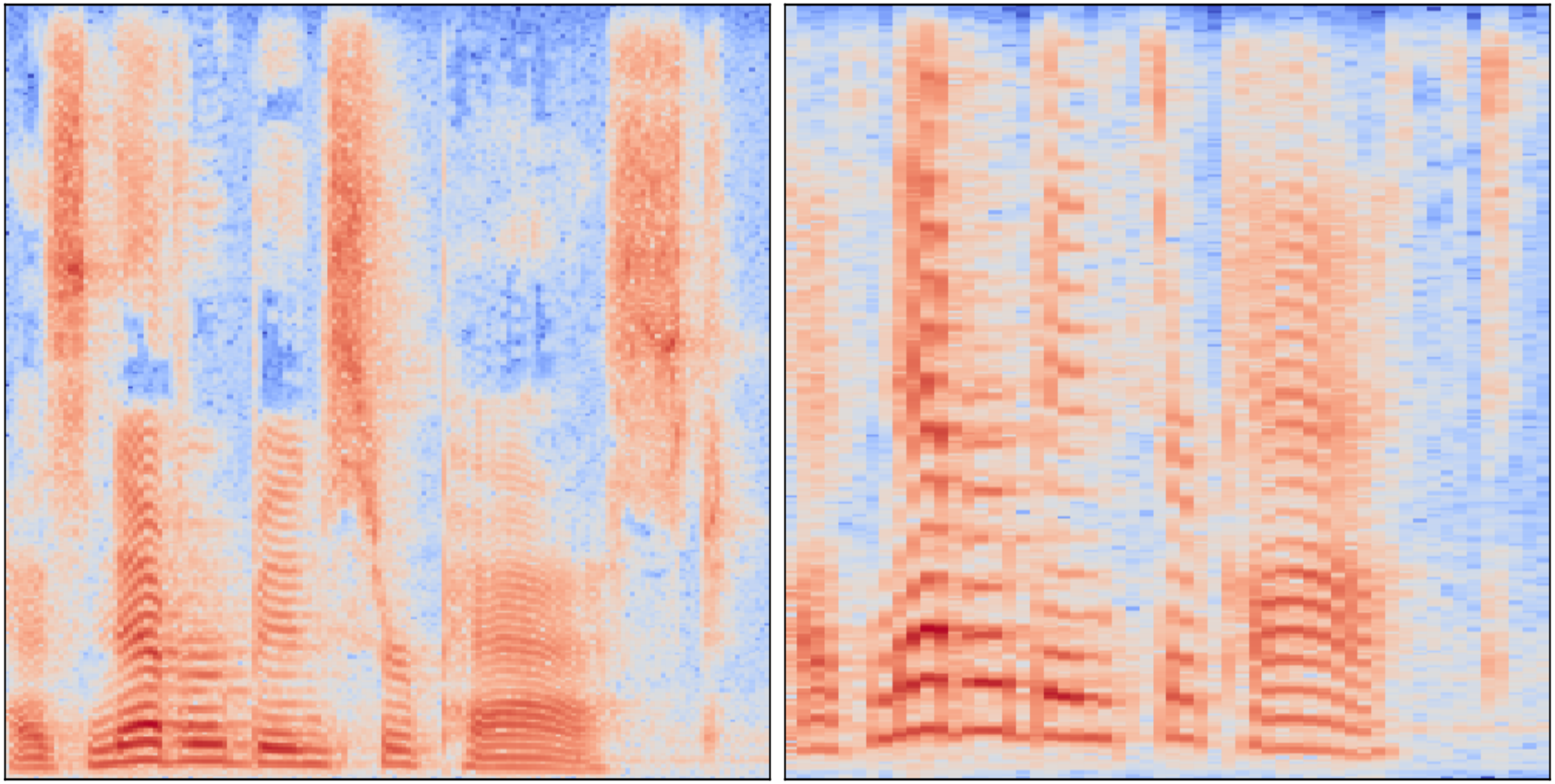
- 13,100 audio clips
- 1.11 sec to 10.10 sec
- single speaker
- with normalized transcription
- first 12 clips for validation

[1] https://keithito.com/LJ-Speech-Dataset/

clean up transcription

- inconsistent use of quotes and brackets
- non-ascii characters
- case

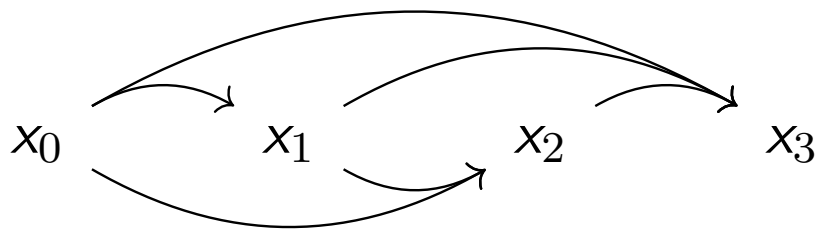
downsampling



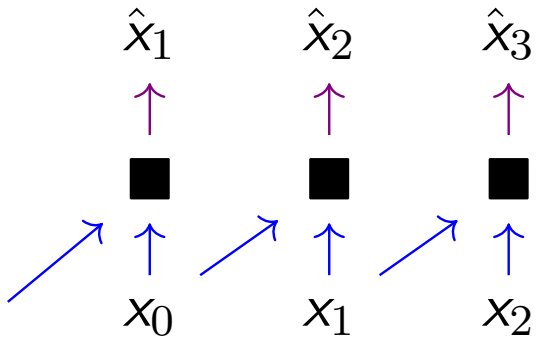
22050 8000

autoregressive model

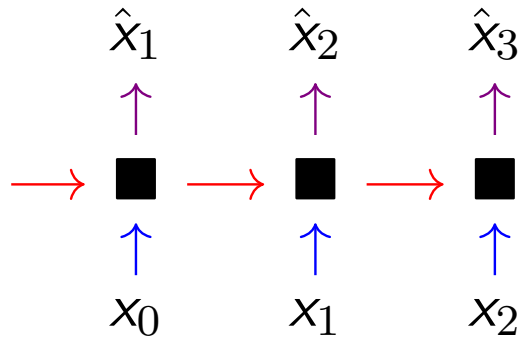
- number of frequency bins $f = 256$
- frame $x_i : \mathbb{C}^f \simeq \mathbb{R}^{2f}$
- model $x_{0\ldots i} \to x_{i+1}$



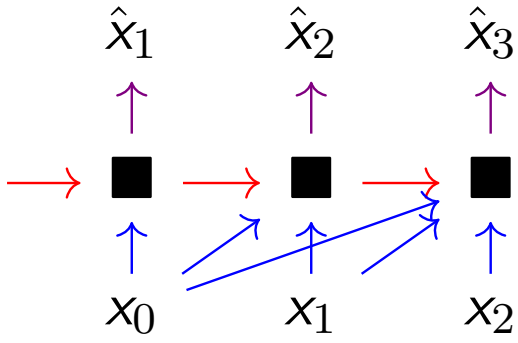
choices



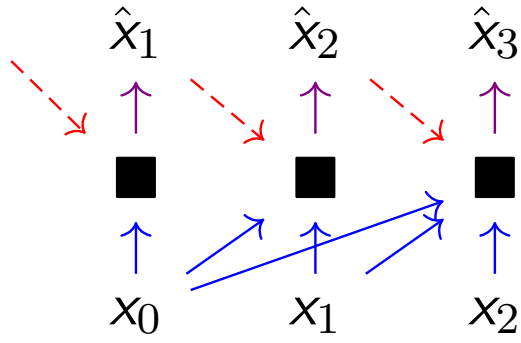
convolution



recurrent

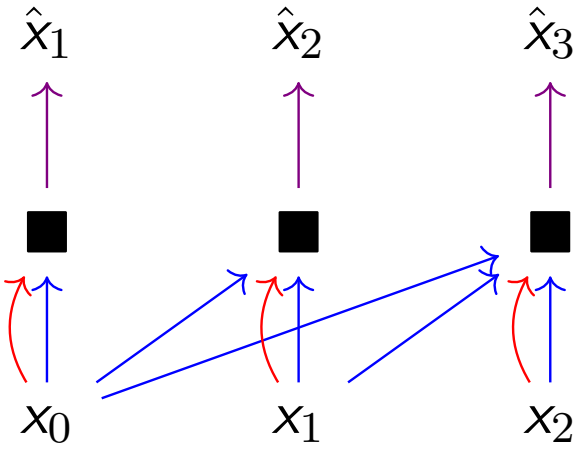


recurrent with attention

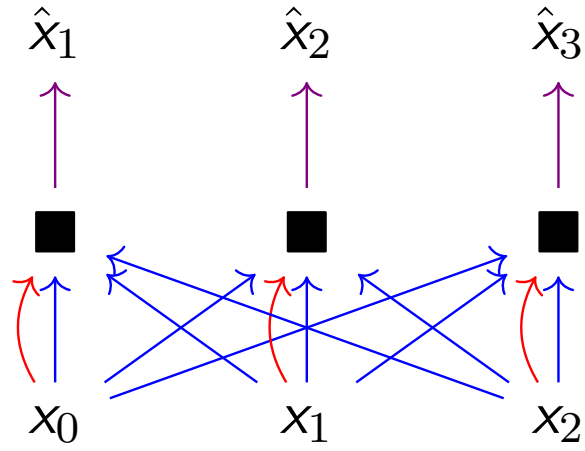


redirect recurrent connection

self-attention[2]



decoder

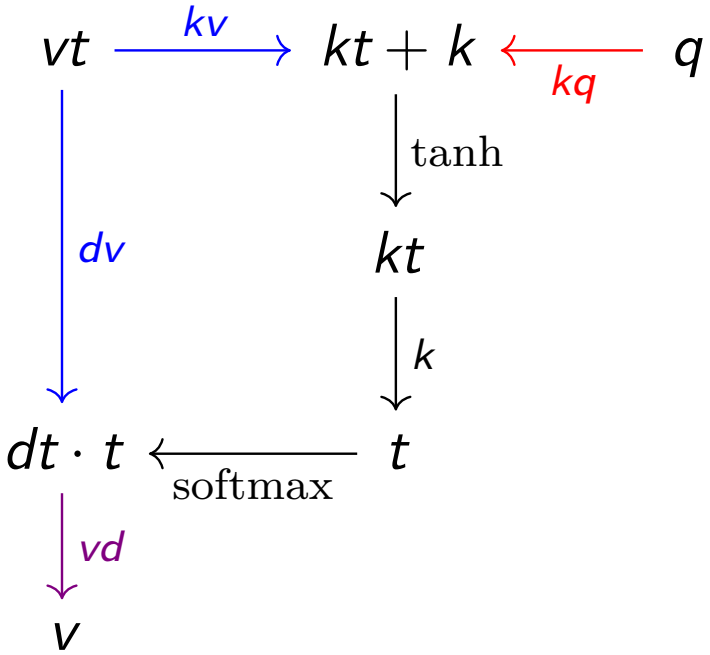


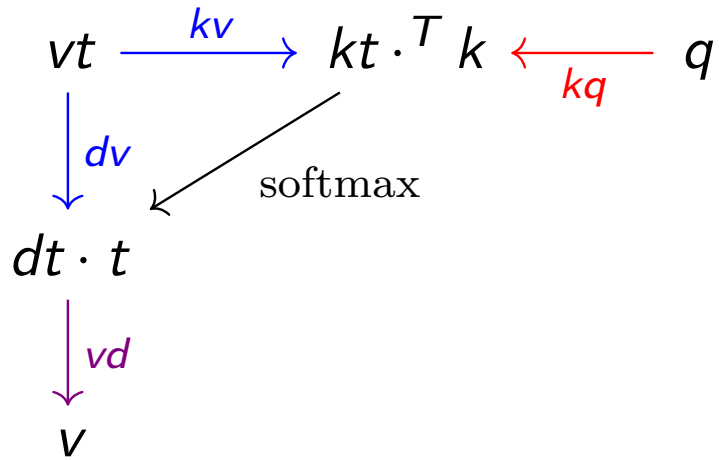
encoder

[2] https://arxiv.org/abs/1703.03130

attention: additive[3], dot-product[4], key-value[5]

time steps t, value dim v, key dim k, query dim q, model dim d

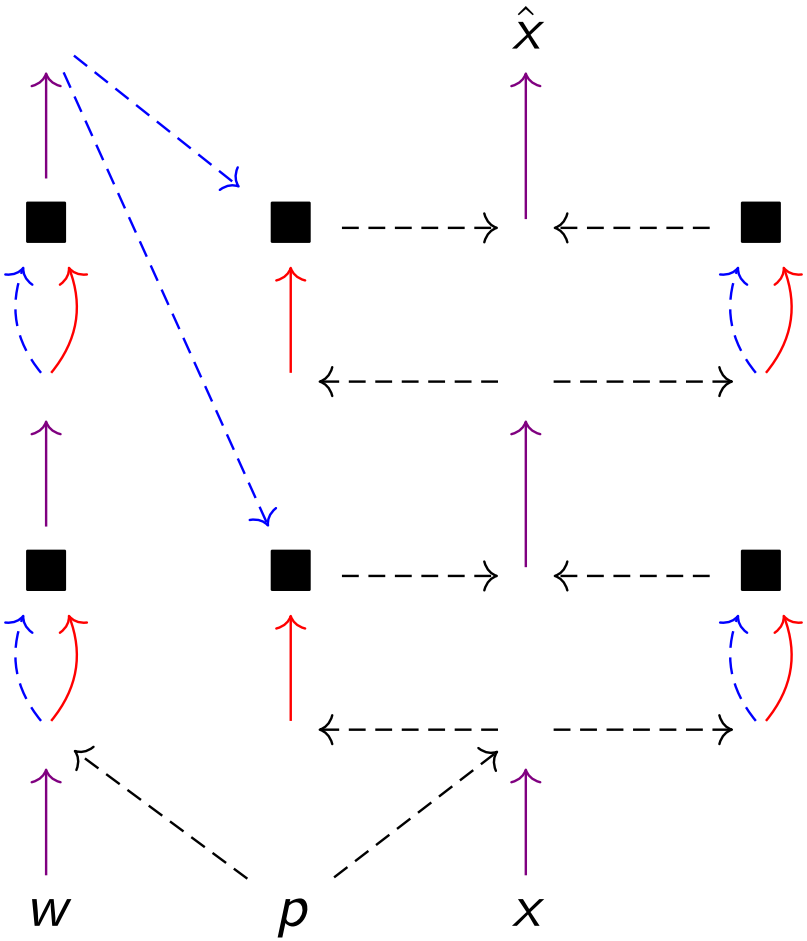




[3] https://arxiv.org/abs/1409.0473
[4] https://arxiv.org/abs/1508.04025
[5] https://arxiv.org/abs/1702.04521

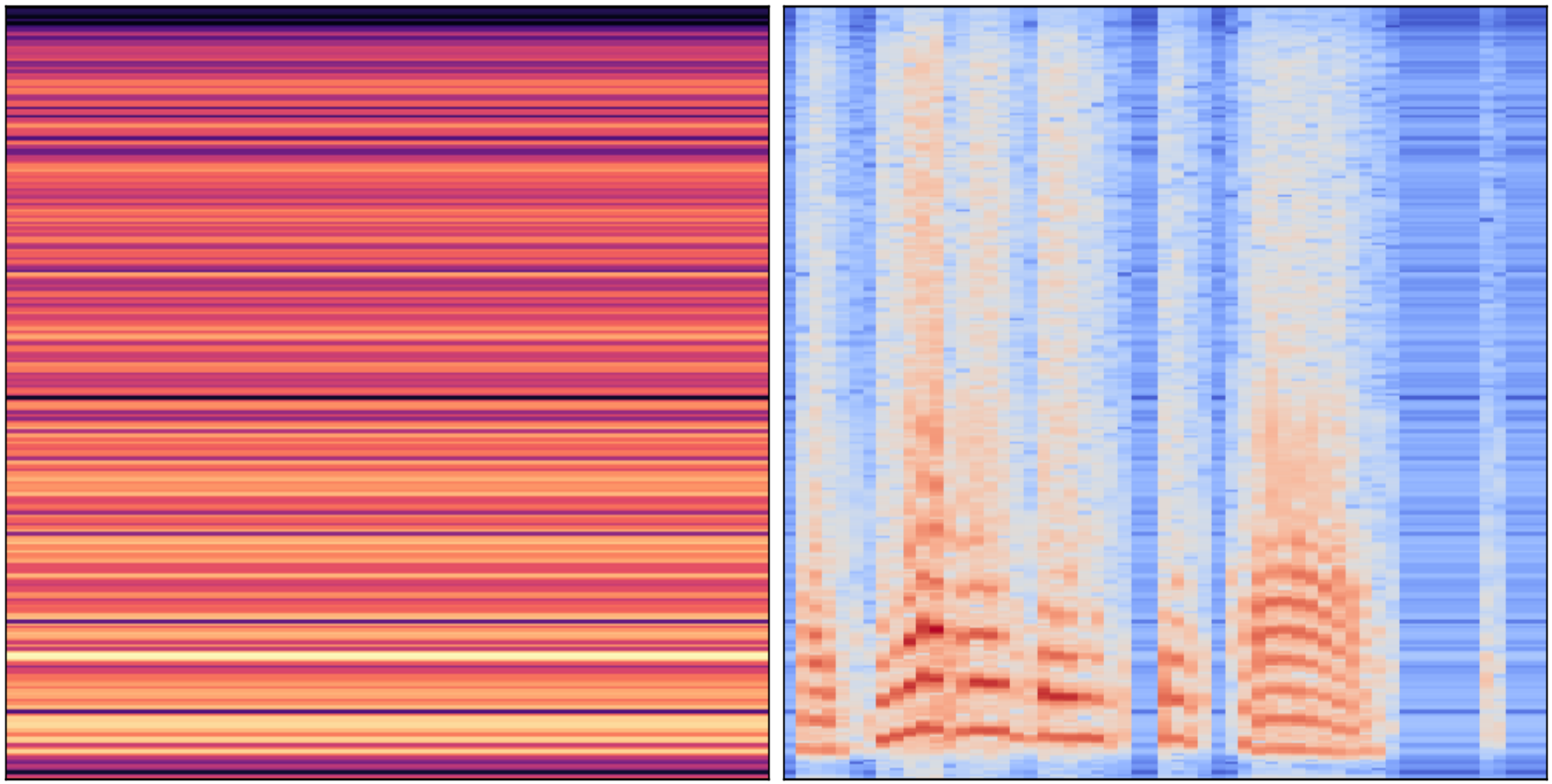
architecture based on transformer[6]

- w: transcript (characters)
- p: position encoding (sinusoids)



[6] https://arxiv.org/abs/1706.03762

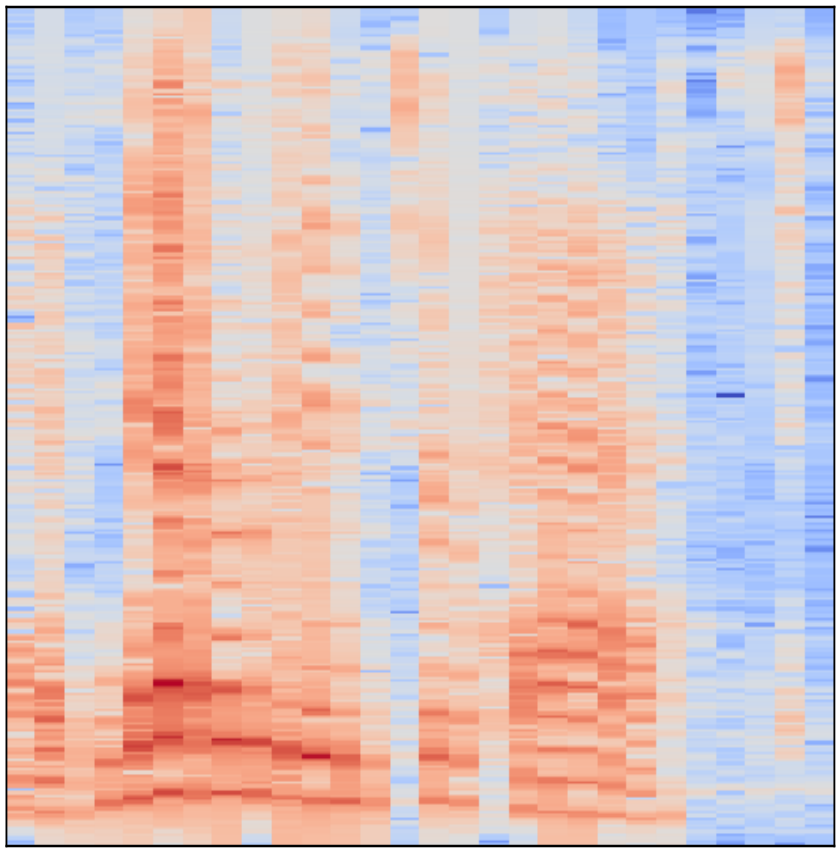
results



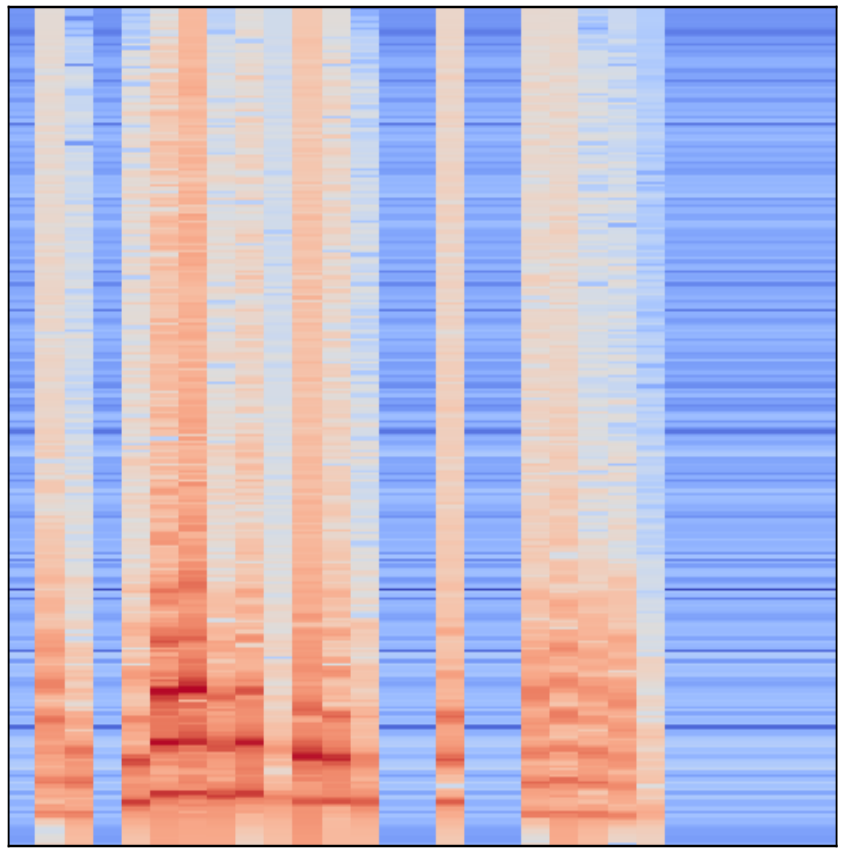
autoregressive

teacher forcing

boxcar

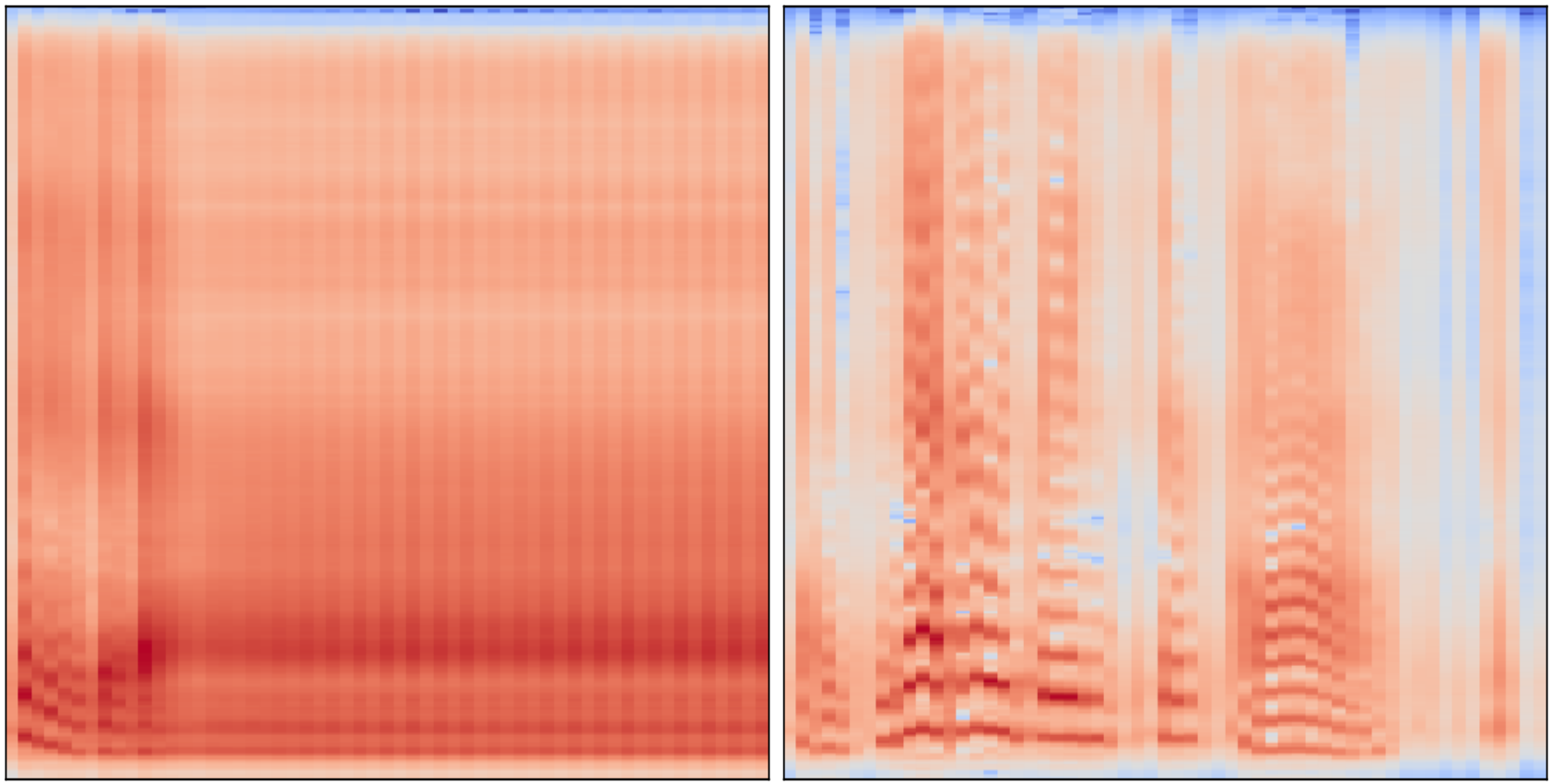


boxcar



boxcar teacher forcing

magnitude only



autoregressive

teacher forcing

plan

- steal from tacotron[7]
- complex arithmetics

[7] https://github.com/keithito/tacotron