



Boulder

# Speech & Music; Synthesis



[YouTube Playlist](#)

**Maziar Raissi**

**Assistant Professor**

Department of Applied Mathematics

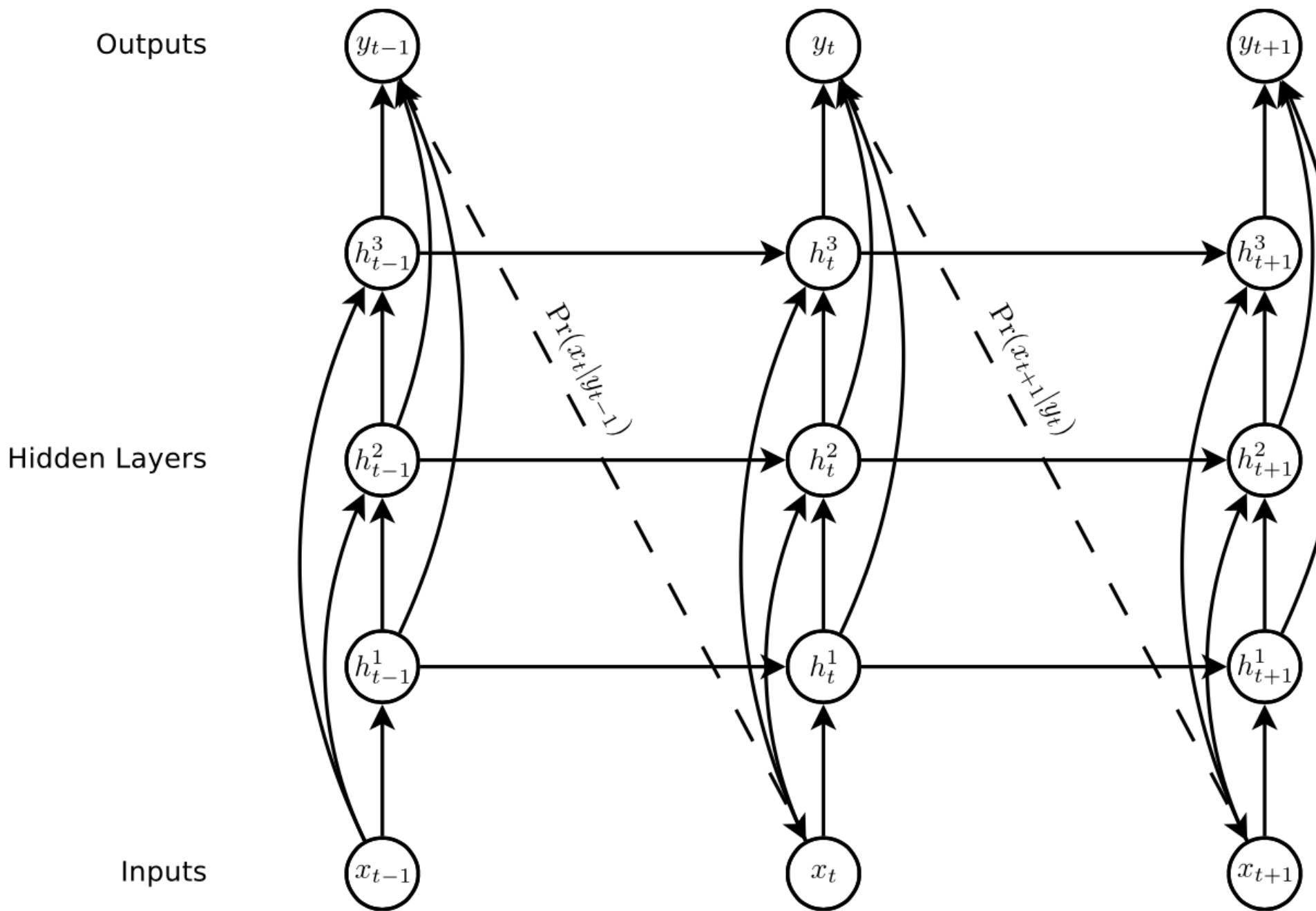
University of Colorado Boulder

[maziar.raissi@colorado.edu](mailto:maziar.raissi@colorado.edu)



Boulder

# Generating Sequences With Recurrent Neural Networks



## Text Prediction

$x_t \rightarrow$  one-hot-vector (word-level or character-level)

$$\Pr(x_{t+1} = k | y_t) = y_t^k = \frac{\exp(\hat{y}_t^k)}{\sum_{k'=1}^K \exp(\hat{y}_t^{k'})}$$

$y_t \rightarrow$  output of the network (Stacked RNNs with LSTM cells)

$$\mathcal{L}(\mathbf{x}) = - \sum_{t=1}^T \log y_t^{x_{t+1}} \rightarrow \text{loss function}$$

Penn Treebank & Wikipedia Experiments

**Evaluation:** bits-per-character (BPC)  $\rightarrow$  average value of  $\log_2 \Pr(x_{t+1}|y_t)$  over the whole test set  
perplexity  $\rightarrow 2^{5.6\text{BPC}}$  where 5.6 is the average word length (in characters) on the test test

## Handwriting Prediction

$$x_t \in \mathbb{R}^2 \times \{0, 1\}$$

$x$  &  $y$  coordinates of the pen offset from the previous inout  
end of stroke (yes:1 or no:0)

$$y_t = (e_t, \{\pi_t^j, \mu_t^j, \sigma_t^j, \rho_t^j\}_{j=1}^M) \rightarrow \text{output of the network}$$

$M \rightarrow$  number of mixture components

$e_t \rightarrow$  end of stroke probability

$\pi_t \rightarrow$  mixture weights

$\mu_t, \sigma_t \rightarrow$  mean and standard deviation (two dimensional)

$\rho_t \rightarrow$  correlation

$$\Pr(x_{t+1}|y_t) = \sum_{j=1}^M \pi_t^j \mathcal{N}(x_{t+1} | \mu_t^j, \sigma_t^j, \rho_t^j) \begin{cases} e_t & \text{if } (x_{t+1})_3 = 1 \\ 1 - e_t & \text{otherwise} \end{cases}$$

## Synthesis Network

$c \rightarrow$  character sequence of length  $U$

$x \rightarrow$  data sequence of length  $T > U$

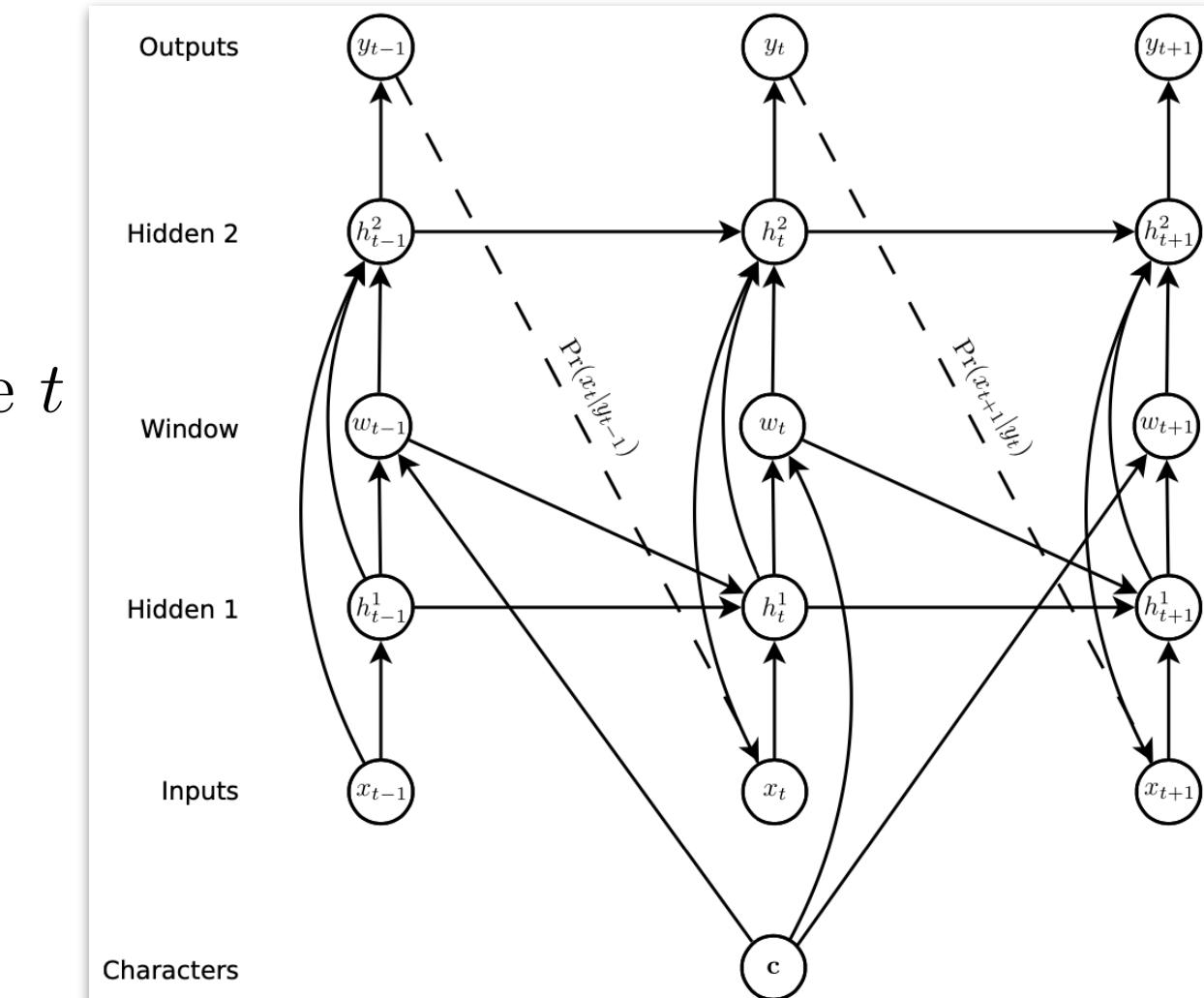
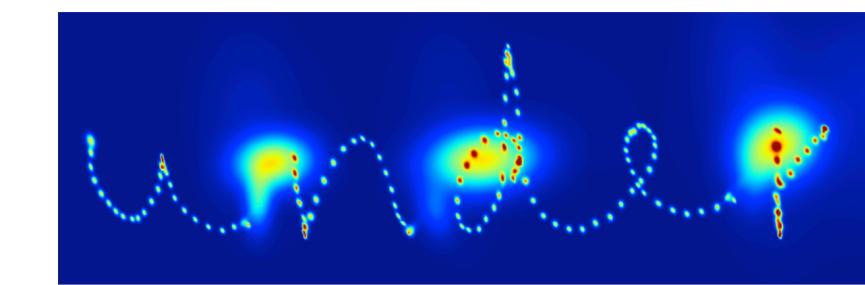
$$w_t = \sum_{u=1}^U \phi(t, u) c_u \rightarrow \text{soft window into } c \text{ at time } t$$

$$\phi(t, u) = \sum_{k=1}^K \alpha_t^k \exp(-\beta_t^k (\kappa_t^k - u)^2)$$

window weight of  $c_u$

IAM online handwriting database

would find the bus safe and sound  
As for Clark, unless it were a  
cousin at the age of fifty-five  
Editorial. Dilemma of  
the the tides in the affairs of men;





Boulder

# Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling



[YouTube Video](#)

- Polyphonic music modeling
- speech signal modeling

RNN

$$x = (x_1, x_2, \dots, x_T)$$

$$h_0 = 0$$

$$h_t = \phi(h_{t-1}, x_t), t = 1, \dots, T$$

$$y = (y_1, y_2, \dots, y_T)$$

$$h_t = g(Wx_t + Uh_{t-1})$$

Generative RNN

$$p(x_1, \dots, x_T) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_T | x_1, \dots, x_{T-1})$$

$$p(x_t | x_1, \dots, x_{t-1}) = g(h_t)$$

LSTM

$$h_t = o_t \tanh(c_t)$$

$c_t$  → memory cell

$o_t$  → output gate

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)$$

$V_o$  → diagonal

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t$$

$f_t$  → forget gate

$i_t$  → input gate

$\tilde{c}_t$  → new memory

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1})$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})$$

$V_f, V_i$  → diagonal

Gated Recurrent Unit (GRU)

$$h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \rightarrow \text{update gate}$$

$$\tilde{h}_t = \tanh(Wx_t + U(r_t \odot h_{t-1}))$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \rightarrow \text{reset gate}$$

$$\text{or } \tilde{h}_t = \tanh(Wx_t + r_t \odot (Uh_{t-1}))$$

$$\text{If } z_t = r_t = 1 \text{ then } h_t = \tilde{h}_t = \tanh(Wx_t + Uh_{t-1}).$$

$$\text{If } z_t = 0 \text{ then } h_t = h_{t-1}.$$

$$\text{If } z_t = 1 \text{ and } r_t = 0 \text{ then } h_t = \tilde{h}_t = \tanh(Wx_t).$$

Sequence modeling

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{T_n} \log p(x_t^n | x_1^n, \dots, x_{t-1}^n; \theta)$$

Polyphonic music modeling datasets

Nottingham, JSB Chorales, MuseData and Piano-midi.

Each symbol in these datasets is respectively a 93-, 96-, 105-, and 108-dimensional binary vector.

Logistic sigmoid function as output units

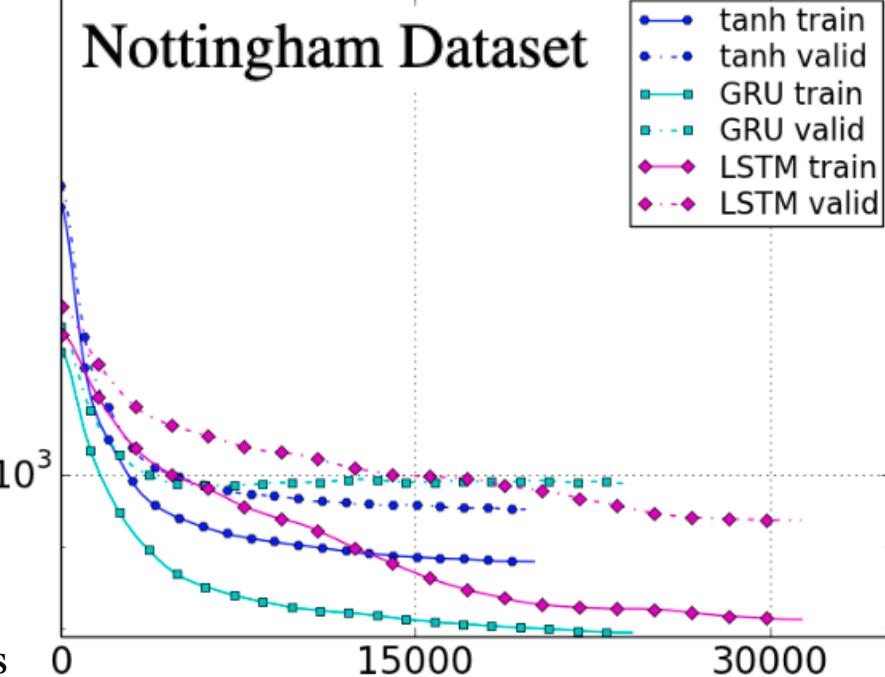
Speech signal modeling

Look at 20 consecutive samples to predict the following 10 consecutive samples in a one-dimensional raw audio signal.

Mixture of Gaussians with 20 components as output layer

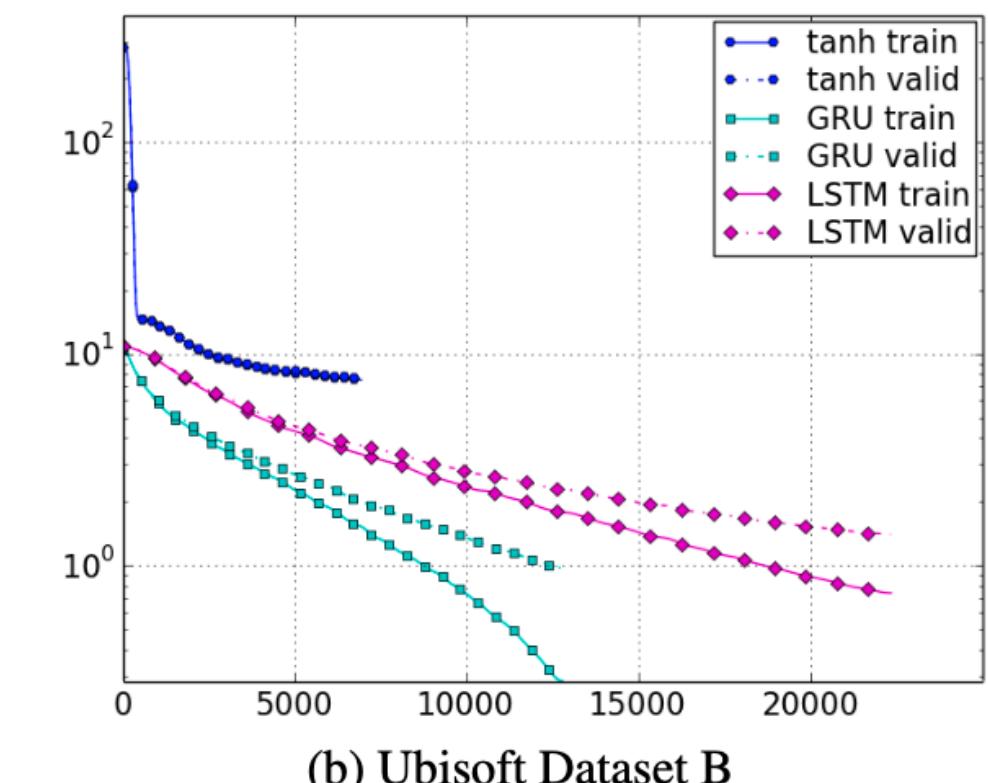
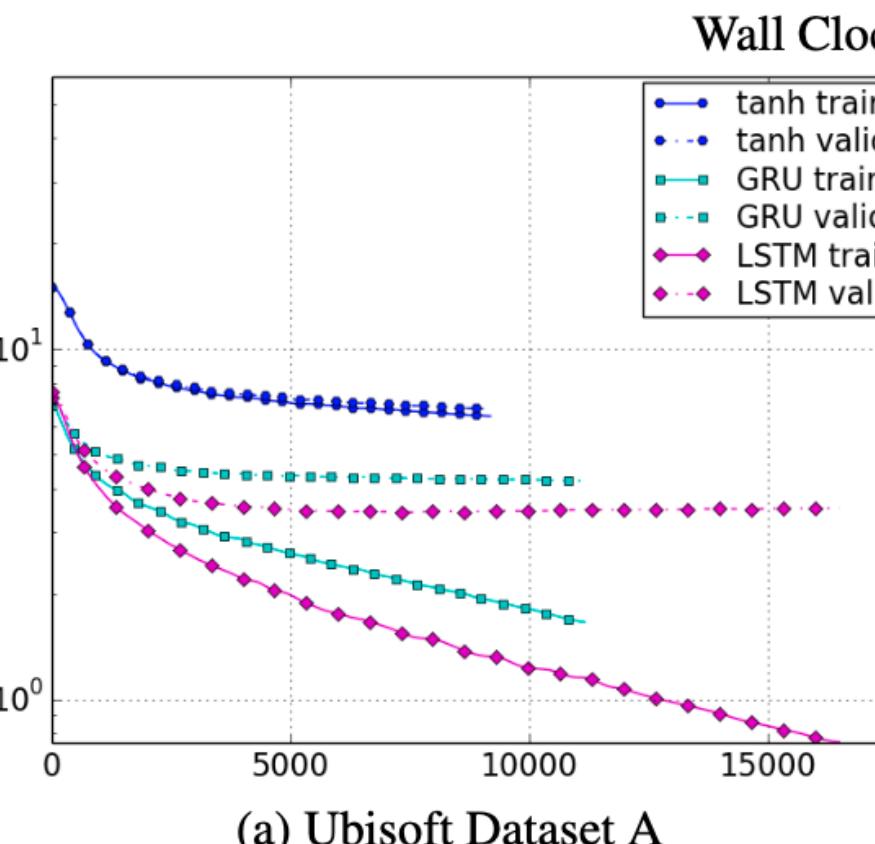
Unit	# of Units	# of Parameters
Polyphonic music modeling		
LSTM	36	$\approx 19.8 \times 10^3$
GRU	46	$\approx 20.2 \times 10^3$
tanh	100	$\approx 20.1 \times 10^3$
Speech signal modeling		
LSTM	195	$\approx 169.1 \times 10^3$
GRU	227	$\approx 168.9 \times 10^3$
tanh	400	$\approx 168.4 \times 10^3$

The average negative log-probabilities of the training and test sets



		tanh	GRU	LSTM
Music Datasets	Nottingham	train	3.22	2.79
	Nottingham	test	<b>3.13</b>	3.23
	JSB Chorales	train	8.82	6.94
	JSB Chorales	test	9.10	<b>8.54</b>
Ubisoft Datasets	MuseData	train	5.64	5.06
	MuseData	test	6.23	<b>5.99</b>
	Piano-midi	train	5.64	4.93
	Piano-midi	test	9.03	<b>8.82</b>
Ubisoft Datasets	Ubisoft dataset A	train	6.29	2.31
	Ubisoft dataset A	test	6.44	3.59
	Ubisoft dataset B	train	7.61	0.38
	Ubisoft dataset B	test	7.62	<b>0.88</b>

Results are not conclusive in comparing LSTM and GRU!







# HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis

WaveNet generates one sample at each forward operation; it is prohibitively slow in synthesizing high temporal resolution audio.

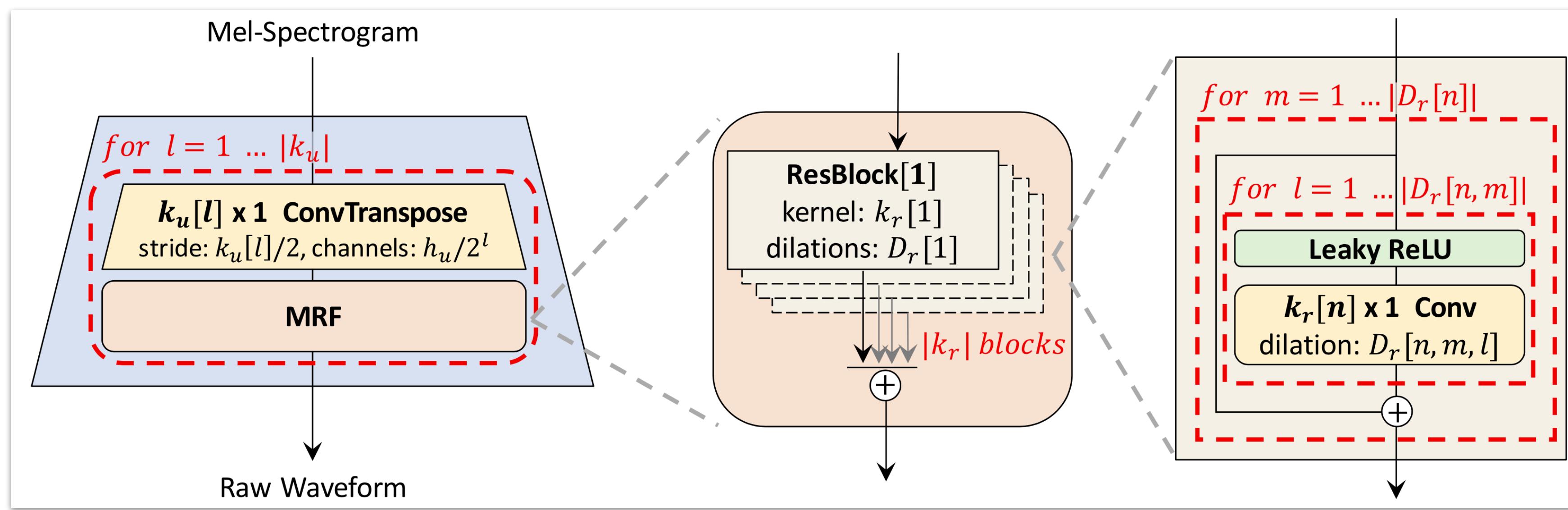
HiFi-GAN achieves both higher computational efficiency and sample quality than autoregressive (e.g., wavenet) or flow-based models

Most neural speech synthesis models use a two-stage pipeline:

- 1) predicting a low resolution intermediate representation such as mel-spectrograms or linguistic features from text, and
- 2) synthesizing raw waveform audio from the intermediate representation.

HiFi-GAN consists of one generator and two discriminators: multi-scale and multi-period discriminators.

## Generator



$h_u \rightarrow$  hidden dimension (e.g.,  $h_u = 512$ )

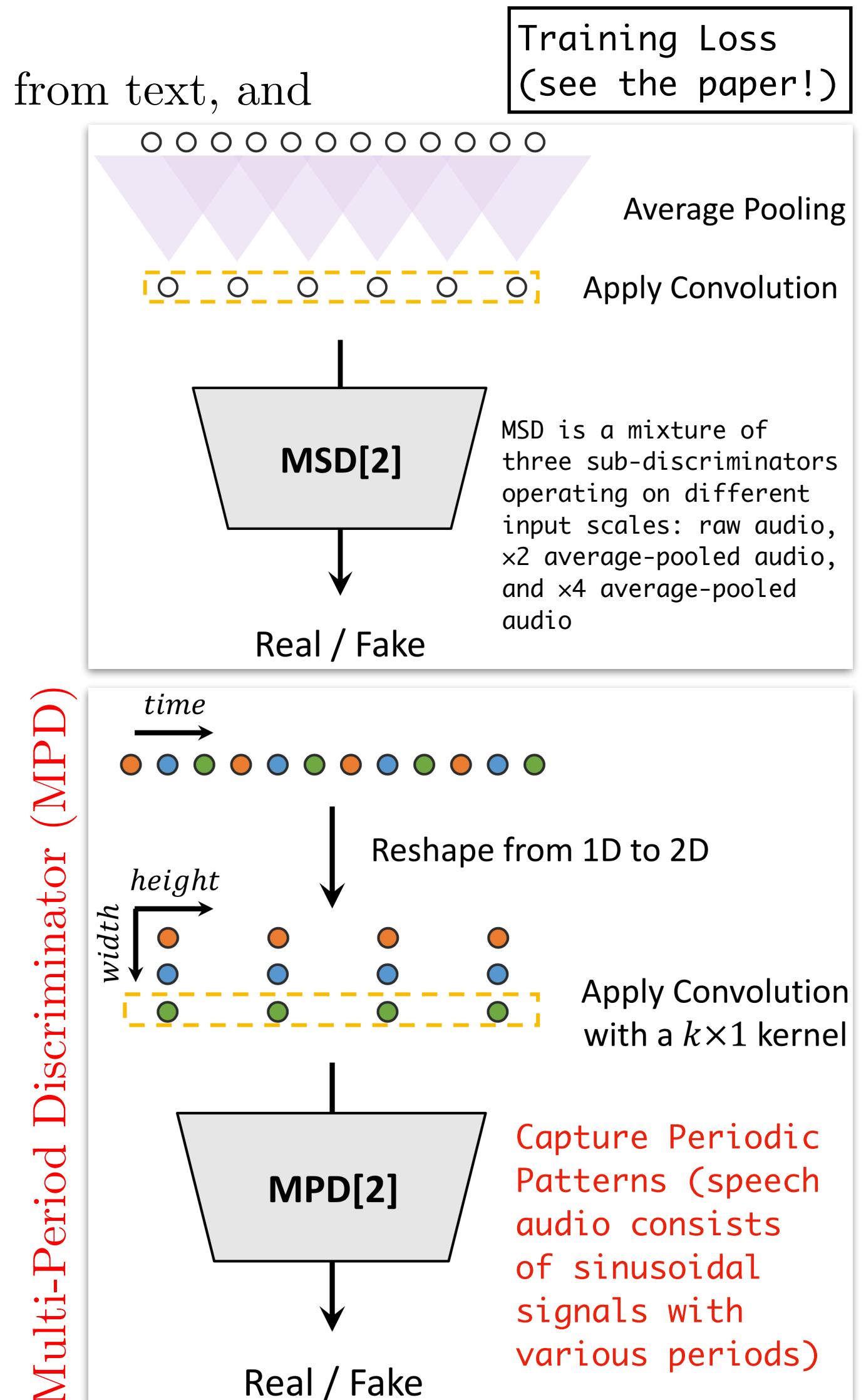
$k_u \rightarrow$  kernel sizes of the transposed convolutions (e.g.,  $k_u = [16, 16, 4, 4]$ )

$k_r \rightarrow$  kernel sizes of MRF (Multi-Receptive Field Fusion) modules (e.g.,  $k_r = [3, 7, 11]$ )

$D_r \rightarrow$  dilation rates of MRF (Multi-Receptive Field Fusion) modules (e.g.,  $D_r = [[1, 1], [3, 1], [5, 1]] \times 3$ )

**Multi-Scale Discriminator (MSD)** → capture consecutive patterns and long-term dependencies

Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis." *Advances in Neural Information Processing Systems* 33 (2020): 17022-17033.





Boulder



# Questions?

[YouTube Playlist](#)

---