Generative Modelling of Sequential Data

M.Sc. thesis in collaboration with Corti ApS

Magnus Berg Sletfjerding

February 9th, 2022

Contents

Introduction

Supervised and Unsupervised learning

Supervised Learning

Learns a mapping from data $\mathbf{x}$ to labels $\mathbf{y}$:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{N} p(y_i|x_i)$$

Unsupervised Learning

Learns the structure of the data $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{i=1}^{N} p(x_t)$$

Why study hierarchies of information in sequences?

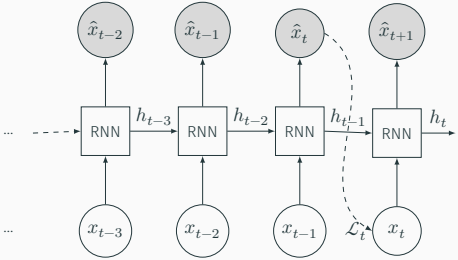
- Most data we work with has some hierarchical structure
    - Text
    - Video
    - Proteins/DNA
- Human brains process hierarchies of information natively
    - Human-like AI requires hierarchical processing
- All real-world data has a sequential dimension - time!

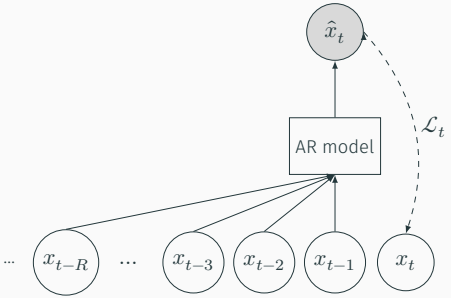
Unsupervised sequence modeling

Unsupervised sequence modeling optimize the likelihood $p(\cdot)$ of the data $\mathbf{x}$, calculated by conditioning the likelihood of x_t on previous timesteps:

$$p(\mathbf{x}) = \prod_{t=1}^{N} p(x_t | x_{<t}), \qquad \mathbf{x} \in \mathbb{R}^N$$

Recurrent vs. Convolutional Autoregressive models





Recurrent architectures

Condition $p(x_t|x_{<t})$ through one or more hidden states h_t passed between timesteps:

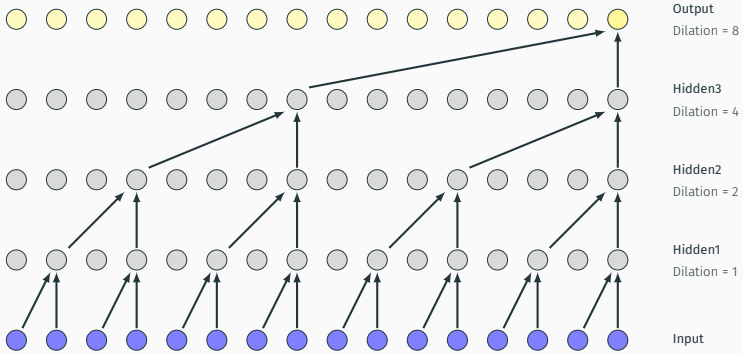
$$p(x_t, h_t|x_{<t}) = p(x_t|x_{t-1}, h_{t-1})$$

Autoregressive Architectures

Condition $p(x_t|x_{<t})$ by viewing a receptive field of size R of the input sequence.

$$p(\mathbf{x}) = \prod_{t=R+1}^{N} p(x_t|x_{\geq t-R+1,<t})$$

WaveNet - Convolutional Autoregressive Sequence Modelling



- Common vocoder in Text to Speech production systems
- Makes use of dilated convolution to inflate receptive field
- No “hidden state” for representing earlier timesteps
- Constrained to look back within receptive field

Problem and Hypotheses

Main Problem with WaveNet ...

- Local signal structure
- Missing long-range correlations
- Low receptive field (300ms)
- Generated audio sounds like babbling if not conditioned on phoneme or text representations

Hypotheses Investigated

1. WaveNet's receptive field is the main limiting factor for modeling long-range dependencies.
2. WaveNet's stacked convolutional layers learn good representations of speech.
3. WaveNet's hierarchical structure makes it suitable to learn priors over representations of speech such as text.
4. A large WaveNet architecture trained on speech can generate coherent words and sentence fragments

Experiments

Experiment overview

1. Expanding Receptive Field by Stacking
2. Latent Space of Stacked WaveNets
3. WaveNet as an ASR preprocessor
4. WaveNet as a Language Model

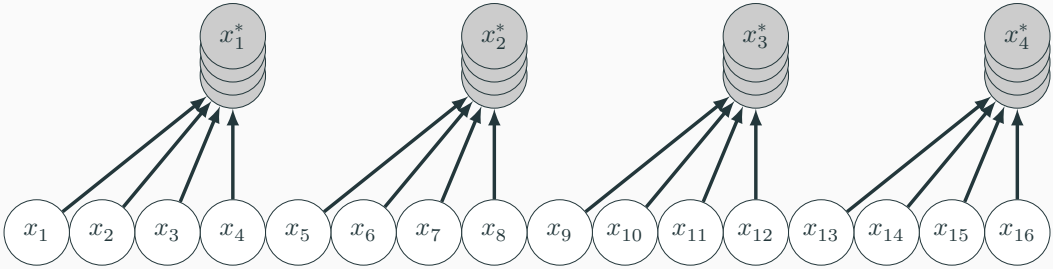
Expanding Receptive Field by Stacking - Setup

Hypothesis tested

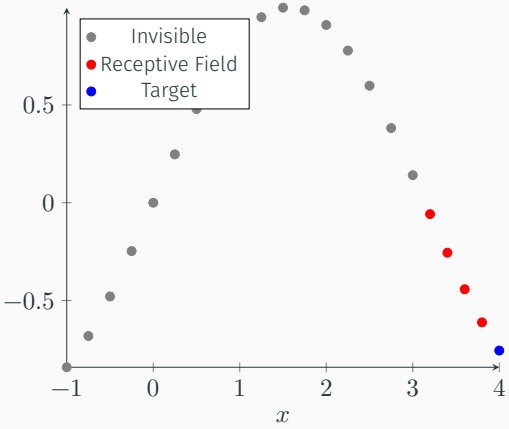
1. WaveNet's receptive field is the main limiting factor for modeling long-range dependencies.

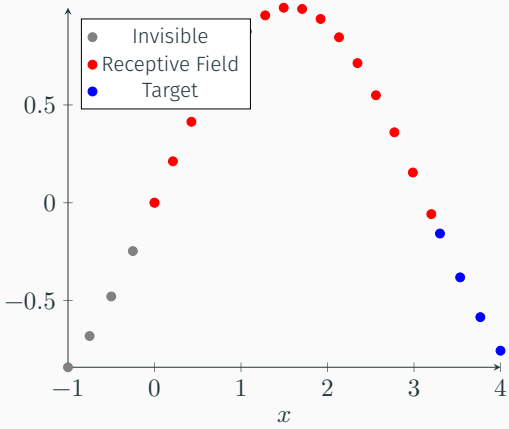
Setup

Transform x as:

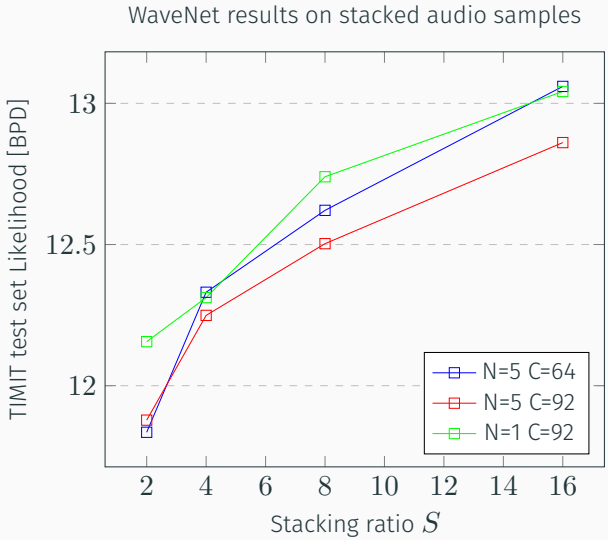


Visualization of stacking on Sin curve





Expanding Receptive Field by Stacking - Results



Expanding Receptive Field by Stacking - Conclusions

1. Stacking does not improve likelihoods significantly.
2. Increasing residual channels increases evaluation likelihoods.

Does this mean that the WaveNet does not extract any semantic information at all?

Is this a failure to measure the output correctly?

Latent space of stacked WaveNet - Setup

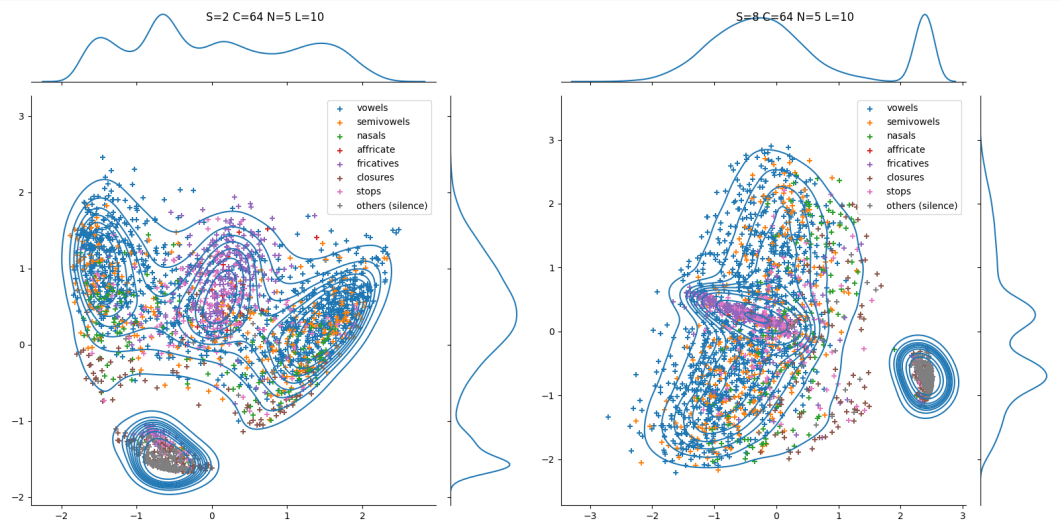
Hypothesis tested

2. WaveNet's stacked convolutional layers learn good representations of speech.

Setup

1. Extract the hidden states of WaveNet for the TIMIT test set.
2. Reduce dimensionality from C to 2 using Principal Component Analysis.
3. Plot density plot of all samples
4. Overlay 2500 samples with phoneme labels to observe clusters

Latent space of stacked WaveNet - Results



WaveNet as an ASR preprocessor - setup

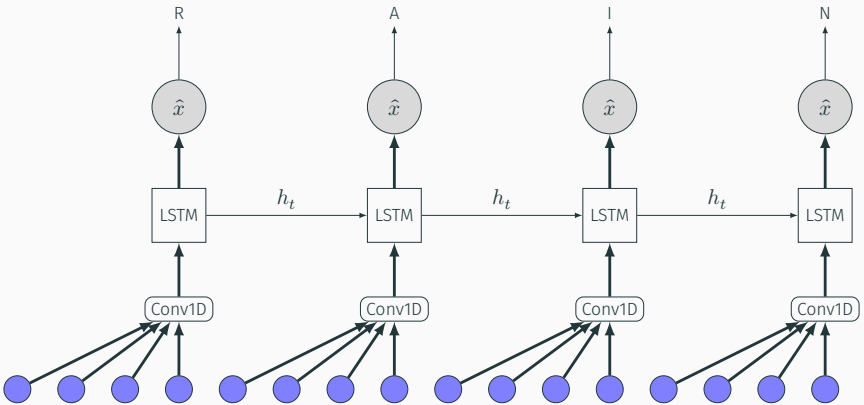
Hypothesis tested

2. WaveNet's stacked convolutional layers learn good representations of speech.

Idea

Are WaveNet's unsupervised representations more useful for Speech Recognition models than raw audio?

WaveNet as an ASR preprocessor - setup (0 layers)



Decoded Output

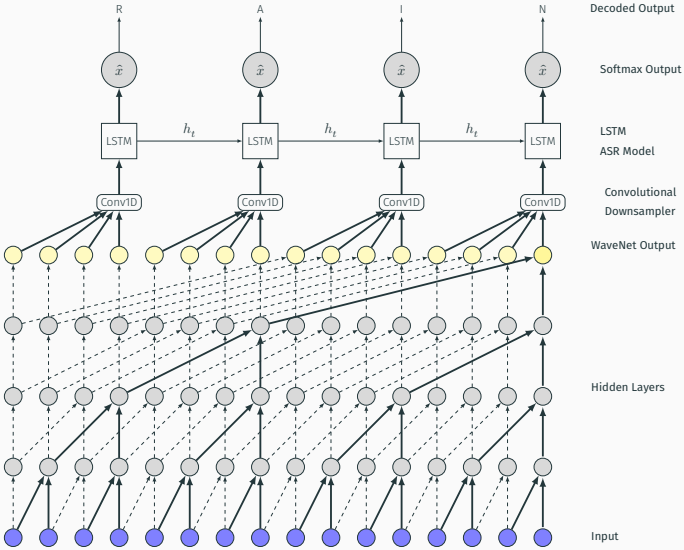
Softmax Output

LSTM
ASR Model

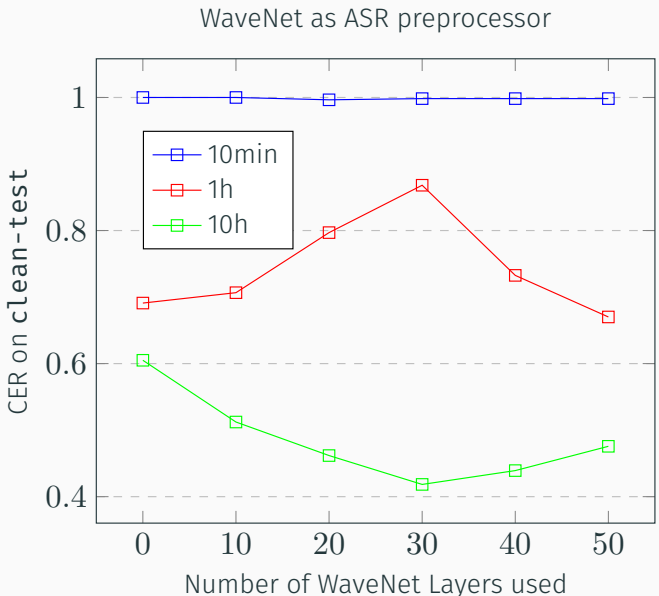
Convolutional
Downsampler

Input

WaveNet as an ASR preprocessor - setup (3 layers)



WaveNet as an ASR preprocessor - Results



WaveNet as an ASR preprocessor - Conclusions

- Using WaveNet as a preprocessor decreases the loss when trained on the 1 hour and 10-hour training subsets.
- The best performance occurs when using 30 layers of the WaveNet trained on 10 hours of training data.
- Notably, WaveNet's use as a preprocessor grows more competitive when increasing the training data size.

WaveNet as a Language Model - Setup

Hypothesis tested

3. WaveNet's hierarchical structure makes it suitable to learn priors over representations of speech such as text.

Setup

- WaveNet implemented as a character-level language model
- Categorical output distribution over alphabet
- Receptive field of 126 characters to match typical sentence length.

WaveNet as a Language Model - Results

Model	Dataset	BPD (test)
Mogrifier LSTM [?]	PTB	1.083
Temporal Convolutional Network [?]	PTB	1.31
WaveNet N=5 L=4 R=24 [RF 126]	PTB	1.835
WaveNet N=5 L=4 R=32 [RF 126]	PTB	**1.666**
WaveNet N=5 L=4 R=48 [RF 126]	PTB	1.678

WaveNet as a Language Model - Conclusions

- WaveNet remains below the bar compared to state-of-the-art models for Language Modelling, contradicting the hypothesis
- This may contribute to WaveNet's limited performance in speech synthesis

Conclusions

Conclusions

Hypothesis	Support?
WaveNet's receptive field is the main limiting factor for modeling long-range dependencies.	No
WaveNet's stacked convolutional layers learn good representations of speech.	Yes
WaveNet's hierarchical structure makes it suitable to learn priors over representations of speech such as text.	No
A large WaveNet architecture trained on speech can generate coherent words and sentence fragments	No

References

Appendix Slides

Experiment: WaveNet Gradient analysis over input space - 1

Explained

Run gradient evaluation over a trained WaveNet model and visualize the outputs.

Hypothesis tested

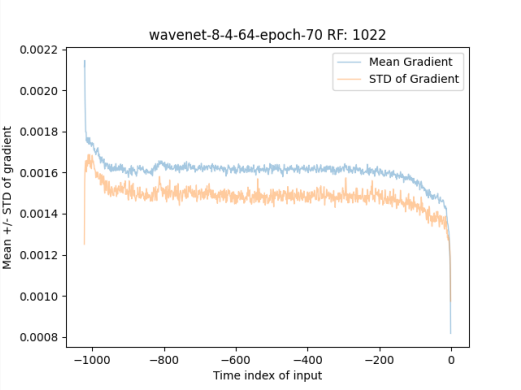
WaveNet uses the entirety of its receptive field for next-step prediction.

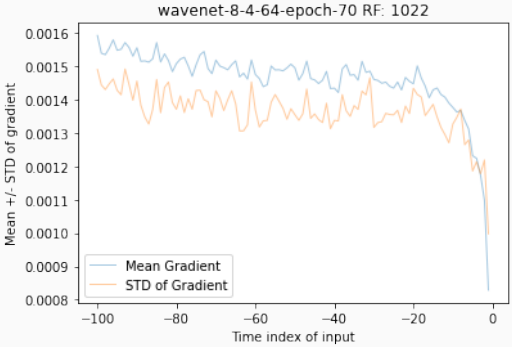
- Gradients in the end of the RF (close to output) are larger than the gradients in the rest of the RF.
- Gradients do NOT collapse to 0 around the beginning of the RF (furthest away from output).

Method

1. Calculate vector-Jacobian product with `torch.autograd`
2. Calculate norm with `torch.linalg.norm`

Experiment: WaveNet Gradient analysis over input space - 2





Notation

Symbol	Explanation
x_i,x_t	The ith index of $\mathbf{x}$, of size N. $x_i \in \mathbb{R}^N$. x_t is used when data is time-resolved.
$\mathbf{x}$	The data x, composed of vectors x_i. $\mathbf{x} \in \mathbb{R}^{T \times N}$
$p_\theta(\cdot), p(\cdot)$	Likelihood function over model parameters θ. Denoted $p(\cdot)$ for brevity
$\hat{x}_i$	Model prediction for x_i.
$\mathcal{L}_i$	Loss function for ith index.
R	Receptive field size.
S	Size of stack size used in stack transformations
d_i	Dilation of ith layer in a WaveNet architecture
C	Number of residual channels

Overview of Codebase and Experiments

Codebase

- https://github.com/JakobHavtorn/vseq/tree/wavenet-exps
- Collaborative codebase with Jakob Havtorn and Lasse Borgholt (PhDs at Corti)
- Includes custom implementations of likelihoods, data processing steps

Work timeline

Experiment	Time
Single Timestep WaveNet	May
WaveNet gradient evaluation check	May-June
Audio LSTM control	May-June
Stacked WaveNet (softmax distribution)	June-Aug
Stacked WaveNet (DMoL)	Aug-Nov
Latent space of stacked WaveNet	Oct-Nov
WaveNet as an ASR preprocessor	Nov-Dec
WaveNet as a Language Model (Text)	Nov-Dec
WaveNet as a Language Model (Nanobody)	Dec

Overview of extra experiments

Model	Dataset(s)	Notes
Single-Timestep WaveNet (softmax output)	TIMIT	Slow convergence compared to later DMoL
Stacked WaveNet (softmax output)	TIMIT, Librispeech	Collapses to predict silence for all timesteps
Single Timestep WaveNet	Generated Sinusoids with periodically modulated pitch.	Fails to follow modulation in pitch

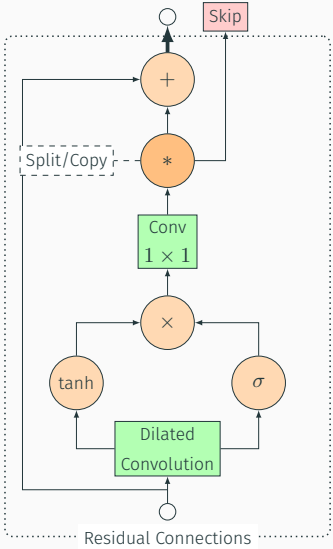
Different tested embeddings for stacked WaveNet input

Embedding type	Dim	Number	Note
Lookup table embedding with input dimensionality $S \times C$	128	1024	Outputs collapsed to silence (suspect too sparse embeddings)
S embeddings convolved together	128	$S \cdot 256$	White noise output.
2 Layer perceptron with input size S and output size R	R	Continuous	Final used embedding

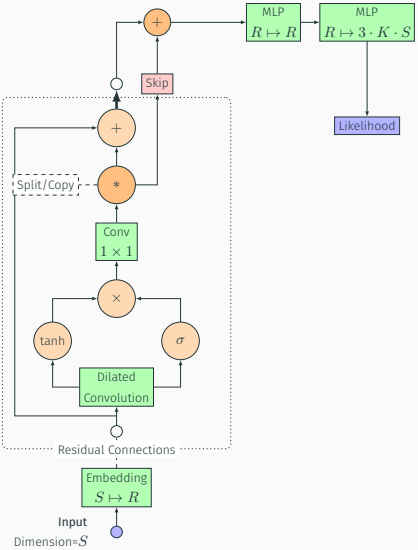
Stacking Transformation

$$x_t^* = \begin{pmatrix} x_t \\ \vdots \\ x_{t+S} \end{pmatrix}, \qquad t \in \{1, S+1, \dots, T-S\}, \mathbf{x}^* \in \mathbb{R}^{N/S \times S} \mathbf{x} \in \mathbb{R}^N$$

Residual Block of WaveNet



Full WaveNet architecture



Discretized Mixture of Logistics

With a mixture of K logistic distributions, for all discrete values of x except edge cases:

$$P(x|\pi,\mu,s) = CDF(x-0.5, x+0.5) = \sum_{i=1}^{K} \pi_i [\sigma(\frac{x+0.5-\mu_i}{s_i}) - \sigma(\frac{x-0.5-\mu_i}{s_i})]$$

Where $\sigma(\cdot)$ is the logistic sigmoid: $\sigma(x) = \frac{1}{1+e^x}$, π is the relative weight vector, μ is the location vector and s is the scale vector.

Softmax distribution

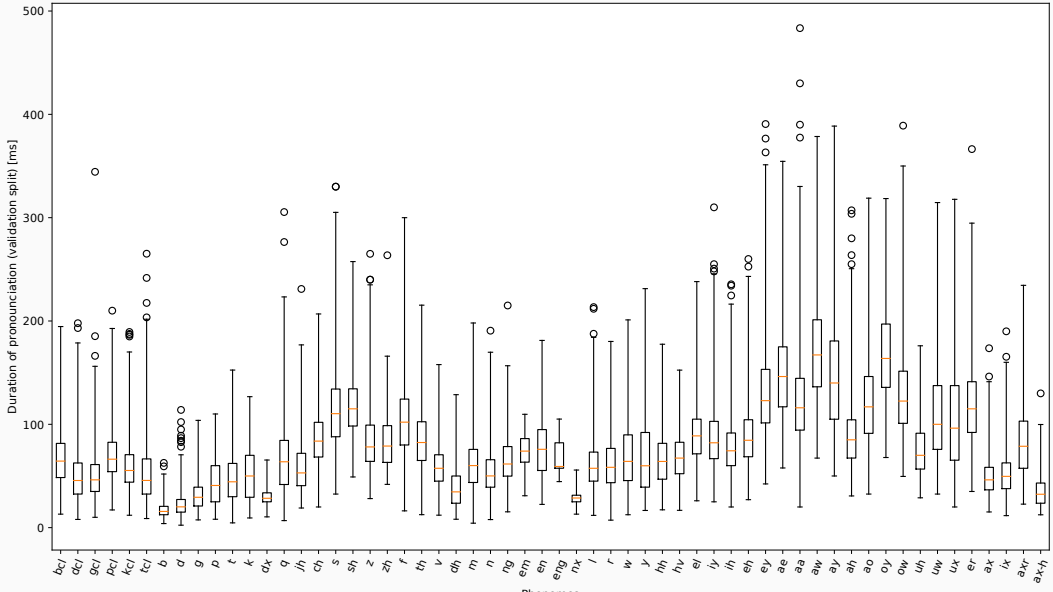
In a softmax distribution, the probability of the ith out of N discrete values is defined by:

$$\sigma(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{N} \exp(x_j)}$$

Group	Phonemes
Vowels	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h
Stops	b, d, g, p, t, k, dx, q
Closures	bcl, dcl, gcl, pcl, tck, kcl, tcl
Affricates	jh, ch
Fricatives	s, sh, z, zh, f, th, v, dh
Nasals	m, n, ng, em, en, eng, nx
Semivowels and Glides	l, r, w, y, hh, hv, el
Others	pau, epi, h#, 1, 2

Table 6: TIMIT phoneme groupings

Phoneme lengths in TIMIT



Mu Law Distribution Illustrated

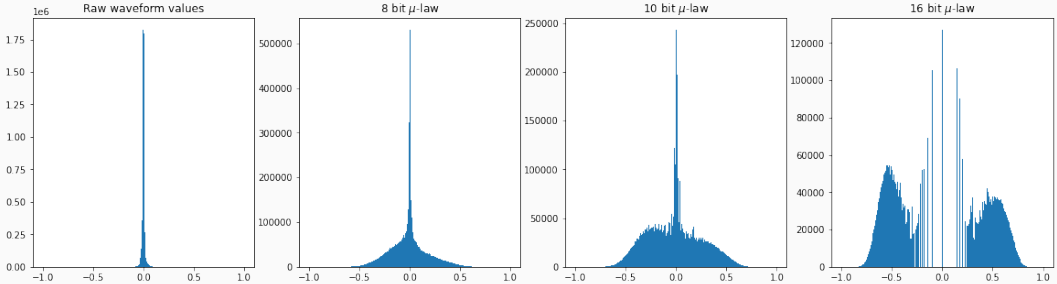
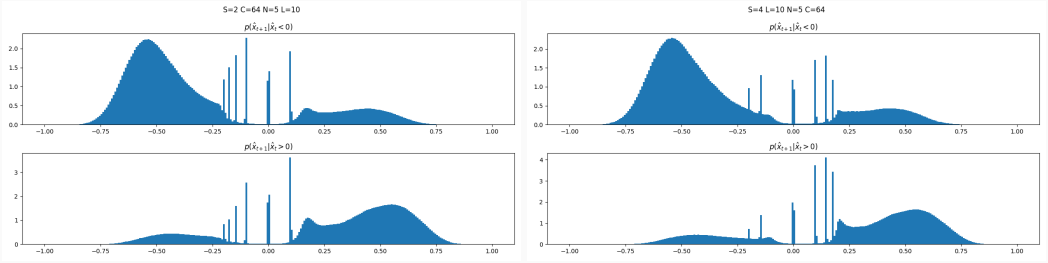


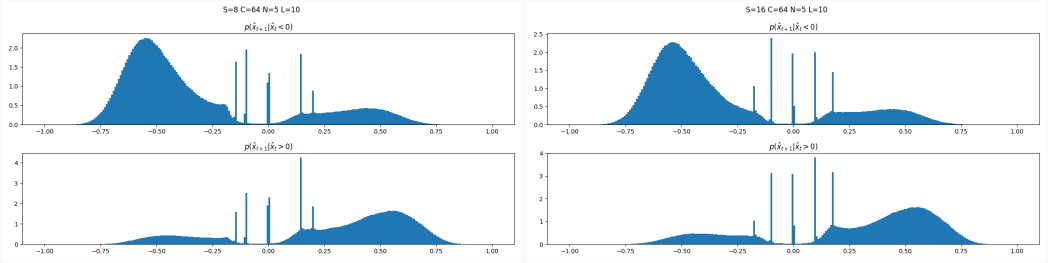
Figure 3: Distribution of raw PCM values from the TIMIT test set. Far left: PCM (16-bit integers). Others: corresponding distribution after μ-law encoding the waveform values with $\mu \in 8, 10, 16$.

Output distribution of WaveNet from Librispeech clean-100h - 1



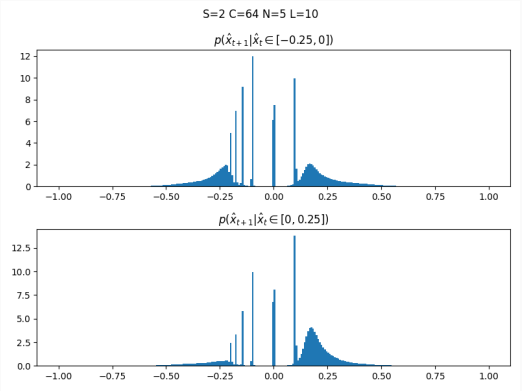
Sampled Output Distributions for WaveNet models trained on the Librispeech `clean-100h` subset. Distributions are in 16 bit μ-law space and binned into 256 bins from -1 to 1.

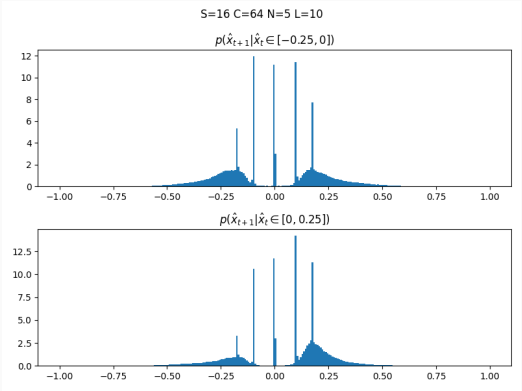
Output distribution of WaveNet from Librispeech clean-100h - 2

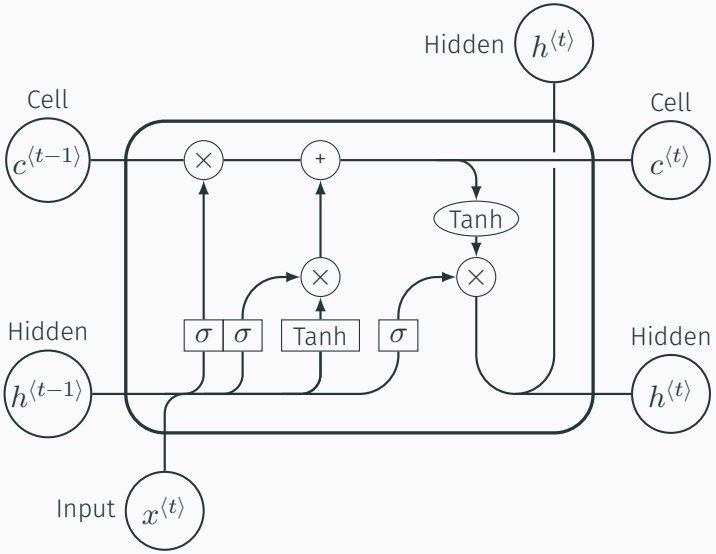


Sampled Output Distributions for WaveNet models trained on the Librispeech `clean-100h` subset. Distributions are in 16 bit μ-law space and binned into 256 bins from -1 to 1.

Output distribution of WaveNet from Librispeech clean-100h - constrained

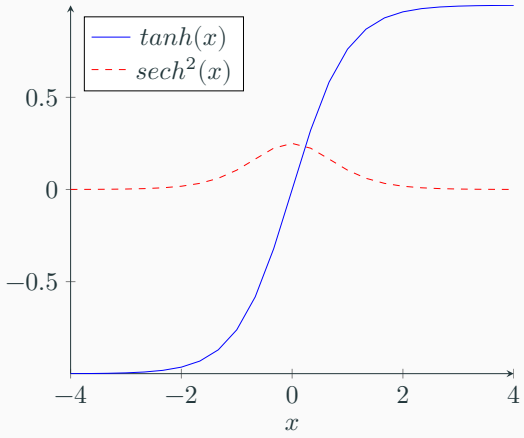




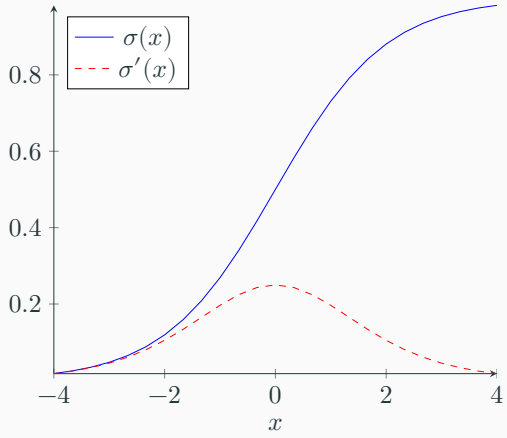

Hidden
$h^{\langle t \rangle}$
Cell
$c^{\langle t-1 \rangle}$
Cell
$c^{\langle t \rangle}$
Tanh
σ
σ
Tanh
σ
Hidden
$h^{\langle t-1 \rangle}$
Hidden
$h^{\langle t \rangle}$
Input
$x^{\langle t \rangle}$

Sigmoid and Tanh Activation functions

Tanh



Sigmoid $x \mapsto \frac{1}{1+\exp(-x)}$



ReLU activation function