# Tacotron 2

Neural network for speech synthesis directly from text.

References: Tacotron2 Location-sensitive attention

## Load Data

- Input: Text
- Output: Audio Recording -> Time-Domain waveforms -> Mel-frequency spectrogram

## Preprocessing

### Input

1. Create word index
2. Convert string to list of numbers based on index
3. Pad all to be same width

### Output

- Short-time Fourier transform (STFT)

  - Frame size: 50 ms
  - Frame hop: 12.5 ms
  - Window function: Hann

- Mel Scale Transforms

  - 80 channel mel filterbank spanning 125 to 7.6 Hz
  - Clip filterbank output to min of 0.01
  - Log dynamic range compression

## Model Components

### Encoder

- Embedding layer
  - Input size: Size of text
  - Output size: 512
- 3 Conv1D-BatchNorm-Relu-Dropout
  - Kernel: 5
  - Stride: 1
  - Output: 512
  - Padding: same
  - p = 0.5 (training only)

- Bi-directional LSTM-Zoneout
    - p = 0.1
    - Units: 256 in each direction

## Location-sensitive attention

possibly use a different/newer attention?

## Decoder

- Prenet - 2 Linear-Dropout
    - Outputs: 256
    - Activation: relu
    - p = 0.5
- Attention network
- Uni-directional LSTM-Zoneout
    - Units: 1024
    - p = 0.1
- Outputs:
    - Linear transform - predicts target spectrogram frame
    - Postnet - 5 Conv-BatchNorm-Tanh(except last)-Dropout
        - Kernel: 5
        - Stride: 1
        - Output: 512
        - Padding: same
        - p = 0.5 (training only)
    - Linear transform/projection down to scalar
    - Activation: Sigmoid

## Model Architecture

## Training

- Feed correct output instead of predicted output to decoder
- Batch size: 64
- Loss Function: Mean Squared Error (MSE)
- Optimizer
    - Adam
    - $\beta_1 = 0.9$
    - $\beta_2 = 0.999$
    - $\epsilon = 10^{-6}$
    - $lr = 0.001$
    - Decay: exponential to $0.00001$ starting at 50,000 iterations
        - Iterations (steps) per epoch $\frac{input size}{batch size}$
        - Iterations to epochs: $Epoch = iterations \times \frac{batchsize}{inputsize}$