EATS

Ondrej Pudiš
End-to-end
- No stages
- No expert-supervision

adversarial
- Uses a generative network
- Inspired by GAN-TTS

text-to-speech
- Text or phonemes input
- Waveform output at 24kHz
Model

Feed-forward convolutional network
latent $z$ is speaker embedding $s$.

- **ALIGNER** aligned with output, learnt, abstract 200 Hz
- **DECODER** 24 kHz
ALIGNER

- Input and output not aligned
- Maps unaligned input sequence to a feature space aligned with the output
- Sample rate of 200Hz
- Speaker-conditioned by an embedded speaker-id
- Replaces linguistic methods and supervision
output for time step $t$: $a_t = \sum_{n=1}^{N} x_{nt} h_n$

interpolation weights for each $t$ in output time steps.

token representation $h$

a batch of input texts padded to the same length

uniform random training window
Input is the 200Hz audio-aligned sequence
- Outputs audio at 24kHz
  - Linguistic and pitch features
- Generator
- Ensemble of random window discriminators
  - Audio fragments of different lengths (240, 480, 960, 1920, 3600)
- Additional spectrogram discriminator
Loss
\[ L_G = L_{G, \text{adv}} + \lambda_{\text{pred}} L_{\text{pred}} + \lambda_{\text{length}} L_{\text{length}} \]

- adversarial (discriminator) loss
- spectrogram-dynamic-time-warping loss
- audio-aligned sequence length loss
Spectrogram loss

- Minimize $L1$ loss between the log-scaled mel-spectrograms
- Spectrograms generated from the generator output and ground truth audio
- Dynamic time warping relaxes the expectation that the spectrograms are exactly aligned

$$\frac{1}{F} \sum_{t=1}^{T} \sum_{f=1}^{F} \left| S_{gen}[t,f] - S_{gt}[t,f] \right|$$
Sequence length loss

- Ensures realistic token lengths predictions
- Makes the predicted lengths close to the ground truth
- Simple squared difference between actual time steps in the sequence $l_{gt}$ and the sum of predicted token lengths

$$L_{\text{length}} = \frac{1}{2} \left( l_{gt} - \sum_{n=1}^{N} l_n \right)^2$$
Training data

- A private dataset of texts and audio recordings
- Performed by professional narrators
- Multi-speaker
- Phonemization

"Marley was dead, to begin with. There is no doubt whatever about that."

"mɑːɹli wæz dɛd tə bɪg ʌɪn ʍɪd. ɛɹ iz nʊ dɔːt ʍɪnɪŋ əbət dæt."
Evaluation

- Metric: **Mean Opinion Score**
  - Given by humans
  - On a scale 1-5 (natural speech 4.55)
- Various combinations of features used
- Did not train without the additional losses
- Reported best MOS of **4.08**
- Other exiting models are better
  - WaveNet -- 4.41
  - Tacotron 2 -- 4.52
  - GAN-TTS -- 4.21
In this work, we take on the challenging task of learning to synthesise speech from normalised text or phonemes in an end-to-end manner, resulting in models which operate directly on character or phoneme input sequences and produce raw speech audio outputs.
THANKS

Does anyone have any questions?
RESOURCES