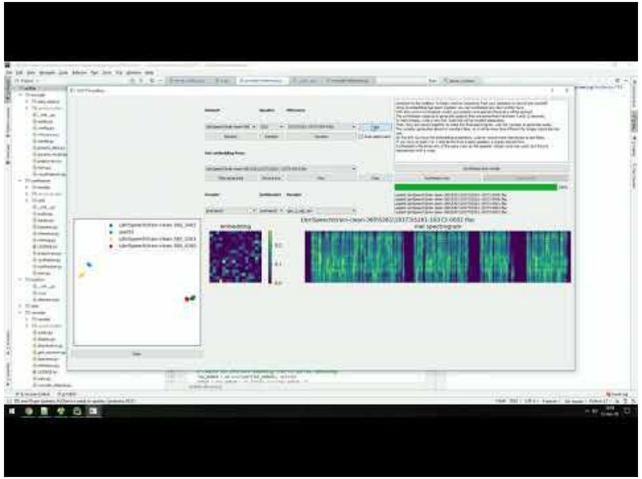
Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

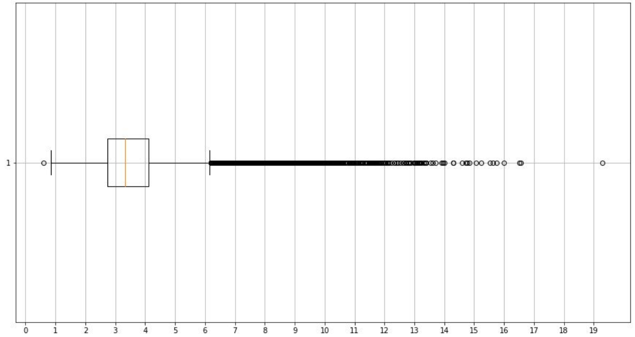
Voice Cloning

Given a reference utterance (target speaker) and a text prompt, our task is to capture the voice characteristics of the speaker to perform text-to-speech (TTS).



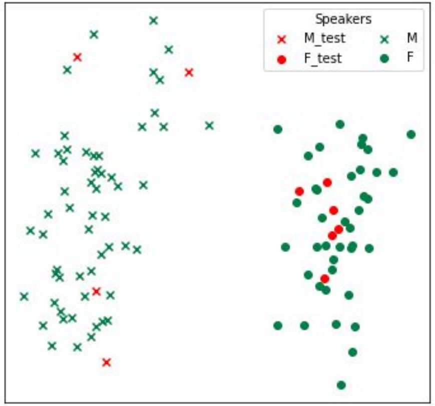
VCTK Dataset

- English
- 109 speakers (107 with transcribed utterances; 46 - male, 61 - female)
- 24bit, 96kHz → 16bit, 48kHz → 16bit, 22.05kHz
- about 400 utterances for each speaker; 44 hours of speech in total; mean length ~3.6s



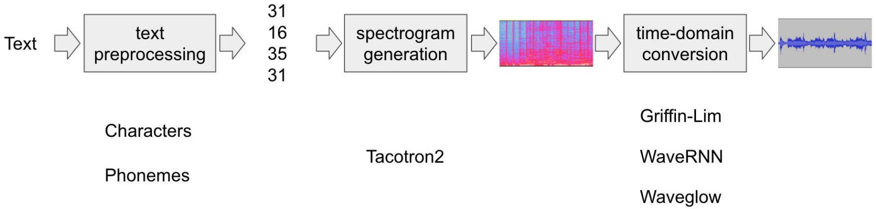
VCTK Dataset

- 109 speakers (107 with transcribed utterances; 46 - male, 61 - female)
- 11 speakers in the test set, the utterances from other speakers are split 80:20 into the training and validation sets



Usual TTS pipeline

- Synthesizer: Tacotron 2, FastSpeech 2, TransformerTTS
- Vocoder: Wavenet, Waveglow, MelGAN, HiFiGAN



Text

text preprocessing

31
16
35
31

spectrogram generation

time-domain conversion

Characters

Phonemes

Tacotron2

Griffin-Lim

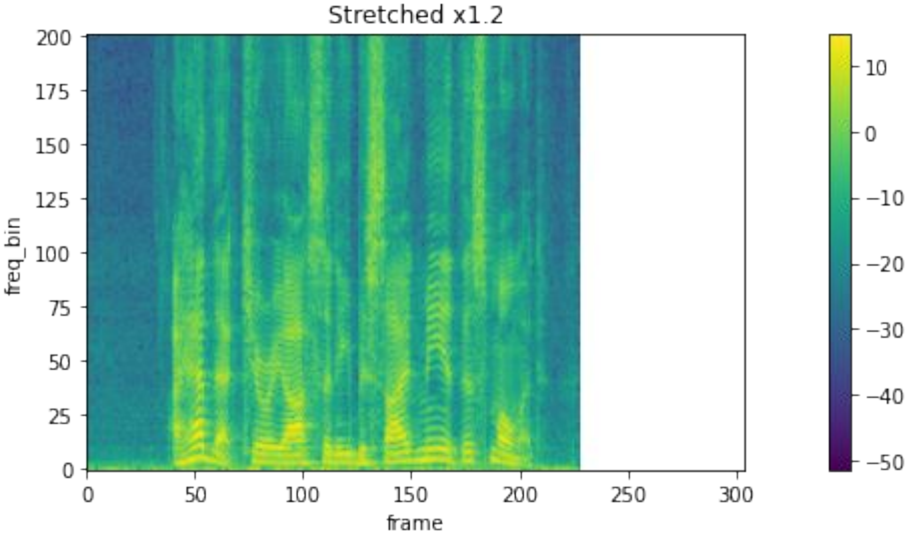
WaveRNN

Waveglow

Mel-spectrograms

- Spectrogram:
  - x-axis - time
  - y-axis - frequency in Hz
  - colour - amplitude in dBs
- Mel-spectrogram:
  - y-axis - frequency in mels. E.g.:

    $$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$
- No phase information



Multispeaker TTS architecture

- Speaker Encoder - Resemblyzer package (pretrained)
- Synthesizer - modified Tacotron 2
- Vocoder - Waveglow (pretrained)

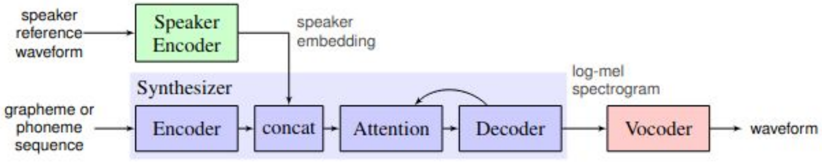


Figure 1: Model overview. Each of the three components are trained independently.

Speaker encoder

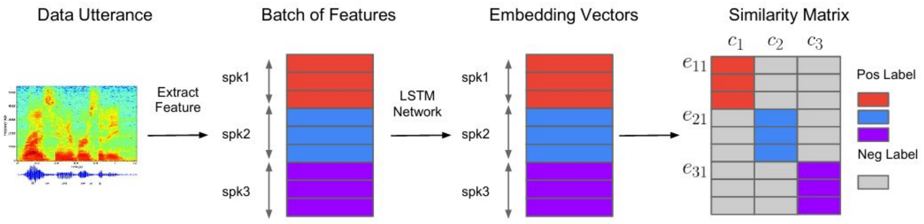


Fig. 1. System overview. Different colors indicate utterances/embeddings from different speakers.

$$\mathbf{e}_{ji} = \frac{f(\mathbf{x}_{ji}; \mathbf{w})}{||f(\mathbf{x}_{ji}; \mathbf{w})||_2}. \quad (4)$$

$$\mathbf{c}_j^{(-i)} = \frac{1}{M-1} \sum_{\substack{m=1 \\ m \neq i}}^{M} \mathbf{e}_{jm}, \quad (8)$$

$$\mathbf{S}_{ji,k} = \begin{cases} w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_j^{(-i)}) + b & \text{if} \quad k = j; \\ w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b & \text{otherwise.} \end{cases} \quad (9)$$

$$L(\mathbf{e}_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \leq k \leq N \\ k \neq j}} \sigma(\mathbf{S}_{ji,k}), \quad (7)$$

$$L(\mathbf{e}_{ji}) = -\mathbf{S}_{ji,j} + \log \sum_{k=1}^{N} \exp(\mathbf{S}_{ji,k}). \quad (6)$$

$$L_G(\mathbf{x}; \mathbf{w}) = L_G(\mathbf{S}) = \sum_{j,i} L(\mathbf{e}_{ji}). \quad (10)$$

Vocoder

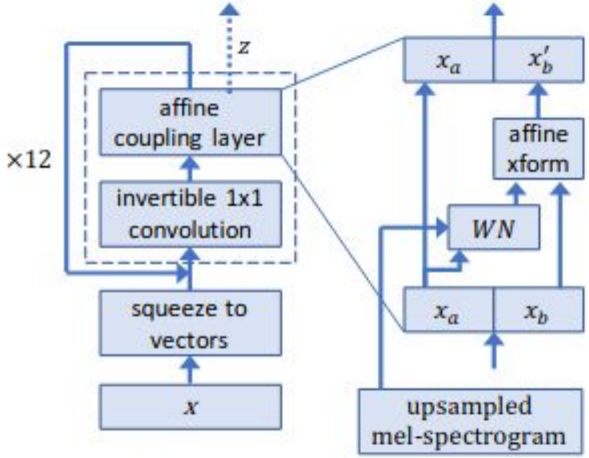
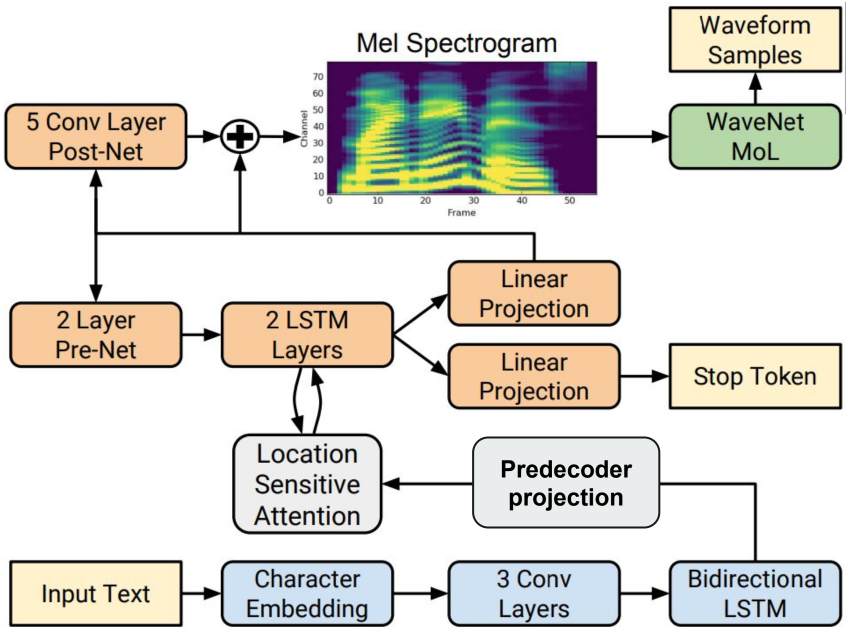


Fig. 1: WaveGlow network

Synthesizer

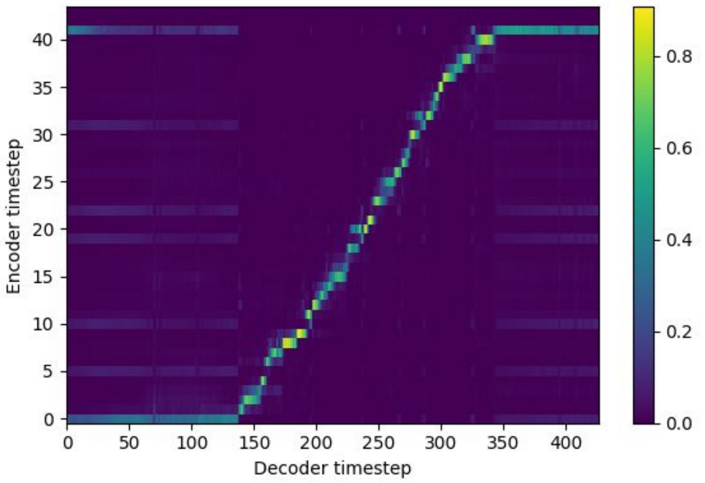


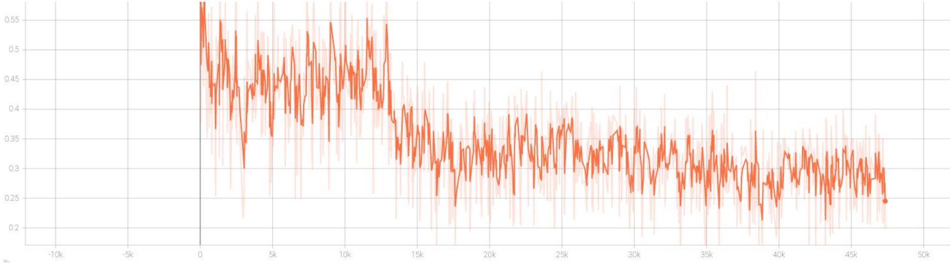
Experiment 1

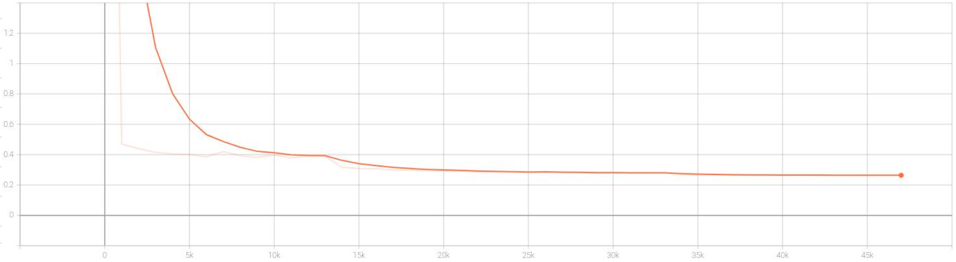
- Warm start from pretrained Tacotron weights
- Train only the layers after the encoder
- Use utterance embeddings
- Ground-Truth Aligned

Experiment 2

- Train the Tacotron encoder too, with the learning rate α/2
- Use utterance embeddings
- Ground-Truth Aligned

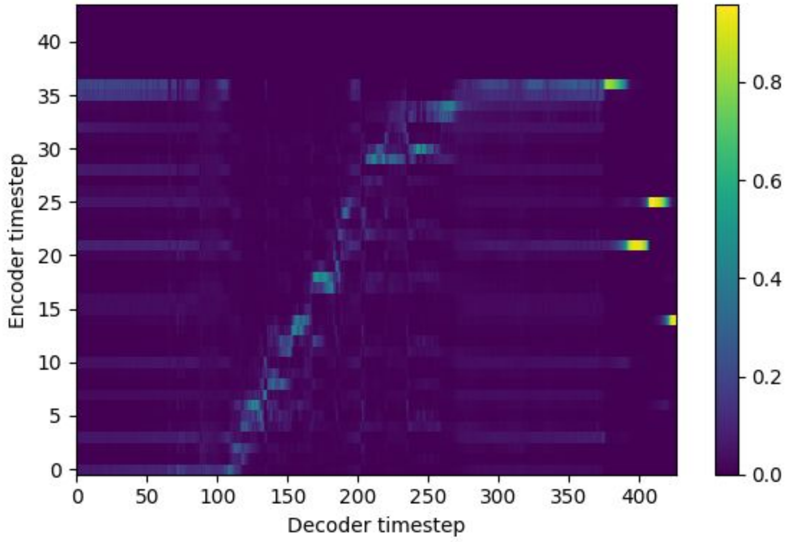


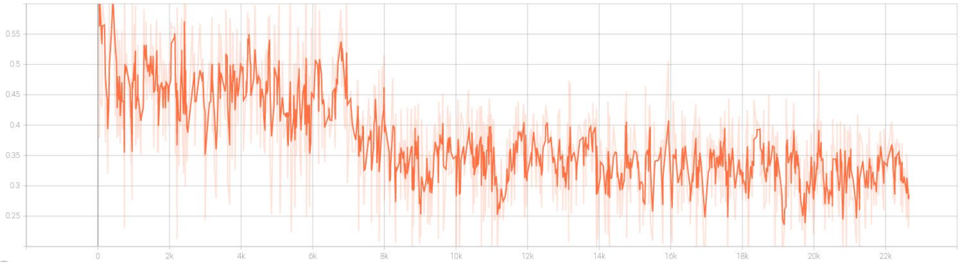


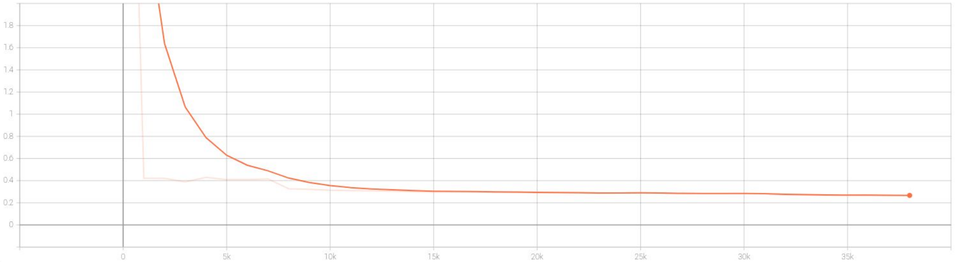


Experiment 3

- Same as previously
- Use speaker embeddings (normalized mean of the speaker utterance embeddings)







Some results

- p345 - reference 🔊
- p361 - reference 🔊

- p345 - “The board is currently resolving its differences” 🔊
- p361 - “This is a test of my trained network” 🔊

Sources

- Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis; Jia, Zhang, et al. - https://arxiv.org/abs/1806.04558
- Resemblyzer repository - https://github.com/resemble-ai/Resemblyzer
- NVIDIA implementation of Tacotron2 - https://github.com/NVIDIA/tacotron2
- Automatic Multispeaker Voice Cloning; Jemine, Corentin - https://matheo.uliege.be/handle/2268.2/6801
- Some images from PyTorch docs - https://pytorch.org/audio/stable/transforms.html
- Generalized End-to-End Loss for Speaker Verification; Wan, Wang, et al. - https://arxiv.org/abs/1710.10467
- Waveglow: A Flow-based Generative Network For Speech Synthesis - https://arxiv.org/abs/1811.00002