**Summary**
- Multi-speaker Tacotron 2 TTS network conditioned on speaker embedding computed from reference utterance.
- Similar to [1, 2] except focus on transfer from pre-trained speaker encoder.
- Make efficient use of available training data.
  - untranscribed and noisy audio to train speaker encoder.
  - smaller dataset of clean speech to train synthesizer.
- Generalizes better than joint training only using clean speech dataset.
- Allows zero-shot adaptation from ~5 second reference utterance.
  - although result is still distinguishable from real speech from that speaker.
- Performance improves with number of speaker encoder training speakers.

**System architecture**
- Speaker encoder computes speaker embedding from spectrogram.
  - stacked LSTM with 3 layers, embedding taken from output at final frame.
  - discriminatively trained on speaker verification task [4].
- Synthesizer generates mel spectrogram from input phoneme sequence.
  - sequence-to-sequence model with attention, based on Tacotron 2 [3].
- Vocoder inverts spectrogram to time-domain waveform.
  - conditional WaveNet [4], 30 dilated convolution layers.

**Experiments**
- Train speaker encoder on internal corpus of 39K hours from 18K speakers.
  - noisy and reverberant speech without transcripts.
- Train synthesizers and vocoders on clean, read speech from LibriSpeech (clean subset) 436 hours from ~1.2K speakers.
  - VCTK: 44 hours from 108 mostly British speakers.
  - hold out 10 speakers from training to evaluate adaptation to unseen speakers.
- Metrics:
  - Subjective mean opinion score ratings of speech naturalness (MOS-nat) and speaker similarity (MOS-sim).
  - Speaker verification equal error rate (SV-EER), measured using eval-only speaker encoder trained on separate dataset.

**Results**
- Speaker verification equal error rate (SV-EER) of 2.9% after enrolling 10 real LibriSpeech speakers and 10 synthetic versions.
  - but real and synthetic utterances consistently form distinct clusters (right).
  - Speaker embeddings: Real vs Synthetic speakers.
  - Female vs Male.
  - PCA, t-SNE.
  - Human, Synthetic.

**Synthesis examples**
- Speaker reference utterance:
  - "and his brothers and sisters stood round and listened with their mouths open".
- Synthesized mel spectrogram:
  - "this is a big red apple".

**Transfer from speaker encoder**
- Matched (LS-Clean) vs Mismatched.
  - MOS.
  - SV-EER.
  - MOS-nat MOS-sim SV-EER MOS-nat MOS-sim SV-EER.
  - 1.2K joint training speakers.
  - 1.2K joint training loss.
  - 2.4K LS-Other.
  - 2.4K LS-Other + VC.
  - 8.4K LS-Other + VC + VQ.

**Fictitious speakers**
- Synthesize text conditioned on randomly sampled speaker embeddings.
  - All samples consist of phonetic content, but varied fundamental frequency and speaking rate.
  - Fictitious speakers are distinct from training speakers.

**References**