uRobo: Speech Synthesis with Limited Data

- Speaker specific speech synthesis is an important area across domains
  - Medicine
  - Media Production
  - Chat bots
- How to produce speech that sounds like one person when you have limited audio of that person’s voice?
- How well does a speaker independent model perform given many hours spken by multiple and diverse voices?

A 4-Layer System

1) ASR/Alignment

- Uses Kaldi to train alignment models on LibriSpeech data (collected and processed by Kaldi)
- Final Forced Alignments used in Preprocessing Step

3) Target Feature Prediction

- Train a Neural Model predict the features extracted in step 2

4)Concatenative Synthesis

- Build a data set of audio units from preprocessed data
- Select units by minimizing target and concatenation cost

2)Preprocessing

- Break each utterance into n-phones (triphones with diphone and monophone backoff and single phone overlap)

HH_B, EH0_I, L_I, OW1_E, M_B, AY1_E/

HH_B, EH0_I, L_I, [OW1_E, M_B], [M_B, AY1_E]...

- Extract Features from each nphone (Duration, Initial Phone F_0, Final Phone F_0, Energy
- Features scaled for speaker independences

3)Target Feature Prediction Training

- N-phone indexes are fed into a 3-layer bidirectional LSTM neural network and trained on features
- Learns 32-dimensional n-phone embeddings
- Contextual information forwards and backwards

4) Concatenative Synthesis

- Uses Viterbi algorithm to minimize target and concatenation cost

$$C^t(t_i, u_i) = \sum_{j=1}^{p} |t_{i_j} - u_{i_j}| \qquad C^c(u_{i-1}, u_i) \sum_{j=1}^{q} |u_{i-1_j} - u_{i_j}|$$

- Logarithmic crossfade over overlapping/adjacent join points

Experiment, Results, Conclusion

To determine the effectiveness of the architecture, three voices were tested against a control voice *c*

- *t10f*: 10 hours of audio units from different female speakers
- *ts39*: 25 minutes of audio units from a single female speaker
- *ts39_m*: Same as above but only utilizing monophones rather than n-phones

- 5 unique mechanical turk workers listened to each of 10 utterances from each voice

- Performance surprising given the low amount of base unit data used to synthesize the voice
- uRobo is a complex system and could be tuned at many different layers to show improvement
  - Different treatment of silences / sequence to sequence of words to phones
- Different voice models may also help
  - Single Speaker with 10 hours of Data
  - Multi-Speaker with pitch contour normalization