Team 18 Proposal: **Transfer Learning from Speaker Verification to Zero-Shot Multispeaker Text-To-Speech Synthesis in Korean**

20150144 김상우 / 20193138 김재윤 / 20160171 김진우 / 20193649 하진철

1. **Project goal:** <Option 2> we are solving our own problem.

   Problem definition

   - Current **Korean text-to-speech (TTS) systems** (e.g. one from Clova) rely on minutes to hours of training voice data from target speaker, and therefore are not capable of synthesizing voice from limited data (e.g. of those who passed away) or of unseen people. To address the problem, we aim to build a TTS system that generates a Korean dialogue from text in the voice of unseen speaker, given only few seconds of speech audio. For this task, we will directly apply the model in the following paper:

     Ye Jia et al., 'Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis', NIPS 2018.

   - **Implementation**

     There exists a public implementation of the paper in TensorFlow, so we will refer to that. Considering our task and dataset, we will modify the data pipeline, architecture of encoder / synthesizer / vocoder modules, and possibly other model components if needed.s

   - **Dataset**

     We will train all three model components (encoder / synthesizer / vocoder) on AI-Hub Korean speech dataset.

   - **Evaluation**

     Unlike most tasks, evaluation of a synthesized speech relies on subjective crowdsourced scores (e.g. Mean Opinion Scores on similarity / naturalness). We'll discuss more on how to evaluate our trained network in a reasonable way (e.g. ask other students?).

     More objective evaluations uses external HMM models, and are limited to testing similarity. So we are likely using MOS measures from a small group.

2. Model description

By transfer learning from speaker verification task, the authors aimed to build a TTS system that can generate speech audio in the voice of different speakers, even those unseen during training. The system consists of three components that are *independently* trained:

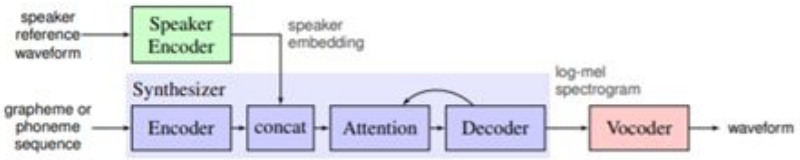

Figure 1: Model overview. Each of the three components are trained independently.

After training, a few seconds of audio from a possibly novel target speaker is used to synthesize new speech in the speaker's voice (zero-shot learning setting).

- ***Speaker encoder network*:** Generates embedding vector from reference speech from a target speaker.

  Model: 3*[768cell LSTM > FC to 256dim] > L2 norm. Trained on **speaker verification task** (with softmax classification layer at the end)

  Training data: Pairs of (speech audio segments (1.6s), speaker identity labels)

- **Seq2seq *synthesis* network:** Generates a mel spectrogram from text, conditioned on the speaker embedding.

  Model: Based on Tacotron 2

  Training data: Pairs of (text transcript, target audio)

  Text preprocessing: input should be given as a sequence of phonemes (음소). In Korean, we can simply tokenize the target text to each phonemes ('가', '나', 다'…), or alternatively use open-source morpheme (형태소) analyzers ('하늘', '은', '스스로'…). We will first use phonemes, but will experiment out whether the second approach could be helpful.

  Transfer learning: embedding vector for the target speaker is concatenated with encoder output at each time step.

- **Autoregressive *vocoder network*:** Converts mel spectrogram to time-domain waveform sample.

  Model: Based on WaveRNN

  Training data: Speech audio from many different speakers

3. References

[1] Clova Speech Synthesis (CSS), https://developers.naver.com/products/clova/tts/

[2] Ye Jia et al., 'Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis', NIPS 2018.

[3] AI-Hub Korean Speech Dataset, http://www.aihub.or.kr/

[4] Jonathan Shen et al., 'Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions', ICASSP 2018.

[5] Nal Kalchbrenner et al., 'Efficient Neural Audio Synthesis', arxiv.