Non-Parallel Many-to-Many Cross-Lingual Voice Conversion
Michael Vobejda (mvobejda@stanford.edu), Mentor: Annie Hu

Problem and Motivation
Voice conversion (VC) is the task of modifying one speaker’s words so that they appear to have been uttered by different speaker. In its final converted form, known as the resulting voice, the speech signal from the first speaker, known as the source speaker, should retain its linguistic content, but it should also be maximally altered in terms of vocal timbre, range, inflection, etcetera in order to match the voice of the second speaker, known as the target speaker. This task has a wide array of applications in synthesizing voices, and it could serve as a key component in producing human-sounding artificial voices for machines. One recent technique in this space is CycleGAN [1], a GAN that uses a notion of cycle-consistency to retain the linguistic content of the source sample. StarGAN [2] is the newest, state-of-the-art technique for this task, based on a modification of CycleGAN that allows for many-to-many mappings using a one-hot speaker identity encoding.

Method
I used the CSTR VCTK Corpus dataset for this project, which includes over 10GB of speech data by 109 English speakers of various accents (http://dx.doi.org/10.7488/ds/1994). Each speaker in the dataset speaks around 400 sentences, sourced from newspapers as well as passages specifically designed to identify the speaker’s accent. Each vocal snippet lasts a few seconds on average. The samples are preprocessed by finding the mel-frequency cepstrum coefficients (MFCC), which are a stable representation of the audio sample.

Data
There are three loss functions that correspond to each component, each weighted by a value lambda. After my initial experiments resulted in audio samples that were well-styled but hard to understand, I experimented with increasing the cycle-consistency lambda value, which is responsible for retention of linguistic content.

Task and Evaluation
In this project, I conducted several experiments aimed at determining the limits of current VC techniques as well as improving upon the current state-of-the-art models. I primarily focused on applying VC to synthetic voices, with the goal of creating a synthetic voice that has the high audio quality and natural pacing of the state-of-the-art in the speech synthesis (WaveNet) while having the vocal style and timbre of the target speaker. Finally, I attempted an application of current techniques to cross-language examples, wherein the source and target vocal sample are in different languages. I evaluated the results of each experiment on three criteria: audio clarity, linguistic content retention, and amount of style transfer. I surveyed 8 colleagues as well as myself and my collaborator.

Model
I use a StarGAN model to perform voice conversion. This model consists of three separate components: a generator, a discriminator, and a domain classifier. The generator is tasked with creating realistic vocal samples, the discriminator is tasked with determining real from fake samples, and the classifier determines which speaker the voice belongs to.

$$L_G(D) = L_G^D(D) = L_{D}(D) - \lambda C \left( G(D) \right)$$

$$L_G(C) = L_G^C(D)$$

There is no loss function for retention of linguistic content.

Results
I used two different implementations of the StarGAN network, one in Tensorflow and one in Pytorch. The values for cycle-consistency lambda and style transfer were still comprehensible and there was clear vocal style transfer from the target sample to the source. This shows that current techniques in VC are capable of handling cross-lingual applications with little modification.

Task: I evaluated the samples using three criteria: audio clarity, linguistic content retention, and amount of style transfer. I surveyed 8 colleagues as well as myself and my collaborator.

Method: I used the CSTR VCTK Corpus dataset for this project, which includes over 10GB of speech data by 109 English speakers of various accents. Each speaker in the dataset speaks around 400 sentences, sourced from newspapers as well as passages specifically designed to identify the speaker’s accent.

Data: There are three loss functions that correspond to each component, each weighted by a value lambda. After my initial experiments resulted in audio samples that were well-styled but hard to understand, I experimented with increasing the cycle-consistency lambda value, which is responsible for retention of linguistic content.

Task and Evaluation: In this project, I conducted several experiments aimed at determining the limits of current VC techniques as well as improving upon the current state-of-the-art models. I primarily focused on applying VC to synthetic voices, with the goal of creating a synthetic voice that has the high audio quality and natural pacing of the state-of-the-art in the speech synthesis (WaveNet) while having the vocal style and timbre of the target speaker. Finally, I attempted an application of current techniques to cross-language examples, wherein the source and target vocal sample are in different languages. I evaluated the results of each experiment on three criteria: audio clarity, linguistic content retention, and amount of style transfer. I surveyed 8 colleagues as well as myself and my collaborator.

Model: I use a StarGAN model to perform voice conversion. This model consists of three separate components: a generator, a discriminator, and a domain classifier. The generator is tasked with creating realistic vocal samples, the discriminator is tasked with determining real from fake samples, and the classifier determines which speaker the voice belongs to.

Results: I used two different implementations of the StarGAN network, one in Tensorflow and one in Pytorch. The values for cycle-consistency lambda and style transfer were still comprehensible and there was clear vocal style transfer from the target sample to the source. This shows that current techniques in VC are capable of handling cross-lingual applications with little modification.

References