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 appears 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, et cetera 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

Data

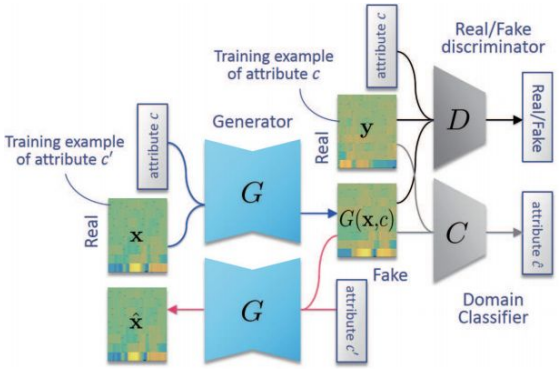
I used the CSTR VCTK Corpus dataset for this project, which includes over 10GB of speech data by 109 English speakers of various accents (http://dx.doi.org/10.7488/ds/1994). Each speaker in the dataset speaks around 400 sentences, sourced from newspapers as well as passages specifically designed to identify 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 cepstral representation of the audio sample.

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.

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.

$$\mathcal{L}_G(G) = \mathcal{L}^G_{adv}(G) + \lambda_{cls}\mathcal{L}^G_{cls}(G) + \lambda_{cyc}\mathcal{L}_{cyc}(G) + \lambda_{id}\mathcal{L}_{id}(G)$$

$$\mathcal{L}_D(D) = \mathcal{L}^D_{adv}(D)$$

$$\mathcal{L}_C(C) = \mathcal{L}^C_{cls}(C)$$

Results

I used two different implementations of the StarGAN network, one in Tensorflow and one in Pytorch. The values for cycle-consistency lambda are displayed, as well as nature of the vocal samples the task was applied to.

	Audio clarity	Linguistic content	Style transfer	Total average
TF baseline	3	2.1	6.1	3.7333
Pytorch baseline	6.4	7.1	6.9	
TF w/ English synth, λ=10	1.9	1.8	6.5	
TF w/ English synth, λ=100	1.1	1.1	2.3	
TF w/ English synth, λ=20	2.4	4.2	4.5	
TF w/ German synth, λ=10	2.3	3.2	4.3	3.2666
Pytorch w/ English synth, λ=20	7.1	7.3	6.8	7.0666

Once the baseline was established, I then trained models with synthetic data as source and natural voices as target to see if this would improve the performance of the model on synthetic voices. I then determined the optimal value for lambda. Finally, I applied the model to synthetic WaveNet vocal samples in German with target vocal samples in English.

Discussion

- The Pytorch implementation generally outperformed the Tensorflow implementation. I believe this is a result of the training for each: Pytorch trained on 10 different voices from VCTK, while Tensorflow only trained on four. There were some minor architectural differences between the models, but the difference in dataset was the primary differentiating factor.
- The models performance improved with an increased value for lambda. The resulting vocal samples were more coherent while retaining similar levels of style transfer.
- Alternatively, increasing the lambda value to 100 drastically decreased the performance of the model. The samples were incoherent and not particularly well styled either.
- The model performed surprisingly well with the German vocal samples. Generally speaking, the content of the samples 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.

Future Work

- Applying this technique to more disparate languages. German and English share many phonemes in common - applying this technique to, say, Mandarin might result in different performance
- Swapping out the attribute c for a voice embedding vector to make this model generalizable (any-to-any)
- Deeper network with more units in the generator/discriminator trained with more of the VCTK dataset
- Further experimentation with loss weighting
- Use new TTS models such as Tacotron to achieve style cross-lingual style transfer

References

[1] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. 2018. High-quality nonparallel voice conversion based on cycle-consistent adversarial network.

[2] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks.