Using Discriminators to Transfer Vocal Style on end-to-end TTS Systems

Justin Rose
justrose@stanford.edu

Motivation

Text to speech (TTS) systems have shown enormous advances recently due to the use of deep end-to-end models. However, most of the models require large amounts of paired text and audio data to train on. Given the difficulty in generating these labels and the vast amount of unlabeled audio available, we propose a discriminative addition to current TTS models to transfer vocal style to new speakers.

Dataset and Features

For the paired text and speech data, we used the LJ Speech dataset [1], which consists of 13,100 short audio clips and their corresponding transcription. For the target vocal style without text data, we chose 6,229 clips of British male speakers from the VCTK [2] dataset. The characters were input to the model as embeddings, and magnitude and mel spectrograms were generated from the audio files, which was used as the targets by the model.

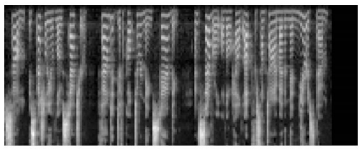


Fig: a sample mel spectrogram

Models

Our baseline TTS model was one designed by Tachibana et. al [3], which uses only convolutions and attention, speeding up training time. Text and audio encoding are fed to an attention module which produces mel spectrograms. We added a discriminator network to the model, which used a series of 1-D convolutions to try to distinguish between generated spectrograms and target speaker. During training, the discriminator influenced the TTS model to produce outputs more similar to the British male distribution, while still having similar contextual information.

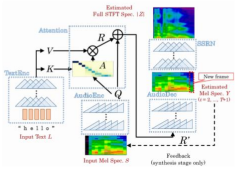


Fig. TTS Model Overview

Objective Functions

We trained the TTS model and discriminator individually before combining. The TTS model had context loss equal to L1 loss, binary divergence loss, and guided attention loss (described in [3]). We use standard GAN loss.

$$L_{con}(Y,S) = L_{att}(A) + \frac{1}{FT}\sum_{f,t}|Y_{ft} - S_{ft}| - (S_{ft}\log Y_{ft} + (1-S_{ft})\log(1-Y_{ft}))$$

$$L_{att}(A) = \frac{1}{nt}\sum_{n,t} A_{nt}W_{nt}, \text{where } W_{nt} = 1 - \exp\left(-(n/N - t/T)^2/2g^2\right)$$

$$L_d(Y_{tar}, Y_{gen}) = -\frac{1}{m}\sum_i (\log D(Y_{tar}^{(i)}) + \log(1 - D(Y_{gen}^{(i)})))$$

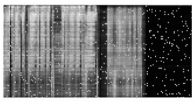
$$L_g(Y_{gen}) = -\frac{1}{m}\sum_i \log D(Y_{gen}^{(i)})$$

$$L_{tot}(Y_{gen}) = \beta L_{con}(Y_{gen}, S) + (1-\beta)L_g(Y_{gen})$$

Results

Our model's results proved to be quite unsatisfactory. Although the TTS model produced relatively high quality samples after being trained on just the female LJ Speech, when the discriminator and generator loss were added, the spectrograms diverged to essentially unintelligible speech. Due to the fact that GANs are highly nonconvex, we believe that could have been a source of our problems. We report the MOS on the pretrained TTS and transferred speech.

	MOS	Fem.	Brit.
Pre-20k	3.1	4.7	1
Disc Model	1.2	1	1.5



Discussion and Future Work

The model didn't converge to any meaningful state. Although it originally produced adequate audio samples in the distribution for the female LJ Speech speaker, when trying to transfer to the British male speakers, the spectrograms got noisy and produced unintelligible speech. We believe that the conflicting goals of the generator and context loss can be blamed for this, the fact that the generator loss has no cycle consistency, and the non-convex nature of GAN. We believe an approach similar to CycleGANs [4] might prove successful in future work.

References

[1] Keith Ito, "The LJ Speech Dataset", 2017. https://keithito.com/LJ-Speech-Dataset/
[2] Christophe Veaux et. Al, "CSTR VCTK Corpus," 2010.
[3] H. Tachibana, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention." arXiv:1710.08969, 2017.
[4] J. Zhu et. al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." arXiv:1703.10593, 2017.