Abstract

Title: Vocal Translation using Pix2Pix Generative Adversarial Networks

Text (1250 Character Maximum):

Voice synthesis has been a popular topic in the machine learning community, due to its growing adoption in modern technology. In the past few years, there has been monumental progress with popular implementations such as Tacotron or Wavenet dominating the field of voice synthesis. However, much of the vocal reproduction research has been focused on pure synthesis, while voice translation on the other hand is a relatively untapped research topic. As opposed to voice synthesis, which generates a completely new voice, voice translation transposes the audio and features of an input onto a target voice. This process mimics the characteristics and audio style of the input voice while maintaining the semantics of the target audio. Using a Pix2Pix generative adversarial network (GAN), audio samples in the form of spectrograms can be paired together to train a neural network that can generate translated audio. Once applied to audio samples, the Pix2Pix network was able to generate audio that mimicked style and semantics albeit with significant noise. However, the results are promising and hopefully will serve as a baseline for future work on vocal translation using a Pix2Pix GAN.