TTS (Text to Speech) 기술의 Audio 음질 개선 프로젝트

권순용, 박기윤, 한세희, 이강만
동국대학교 멀티미디어공학과

e-mail : story_gardener@naver.com

A Project of Audio Quality Development from TTS

Soonyong-Kwon, Giyoon-Park, Saehee-Han & Gangman Yi
Dept. of Multimedia Engineering, Dongguk University, Seoul, 04620, Korea

요 약

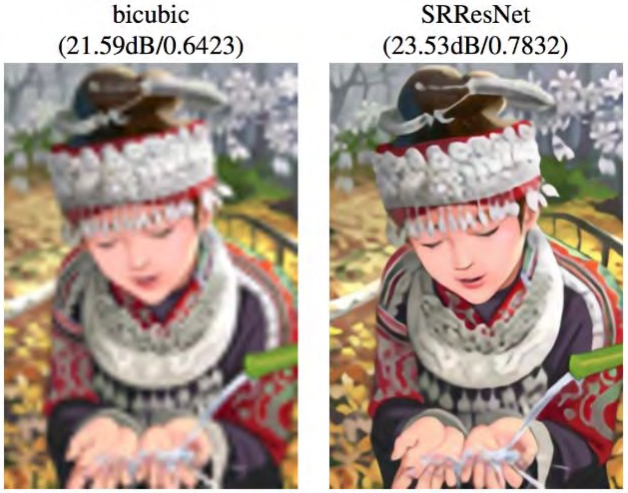
기존의 TTS (Text to Speech) 기술이 가지고 있는 음질의 한계를 GAN - Super Resolution 기술을 이용하여 개선시킨다.

1. 서론

최근 다양한 산업분야에서 TTS(Text to Speech)의 수요가 증가하고 있고, 이러한 요구에 맞추어 다양한 TTS 기술들이 개발되고 있다. 하지만 기존의 TTS기술에서의 음성은 잡음이나, 기계음 과 같이 자연스럽지 못한 소리가 섞여 나와 사용자가 이용하는데 있어 어색함을 느낄 수 있다. 따라서 기존의 TTS 기술에 GAN - Super Resolution 기술을 이용하여 더 나은 음질을 사용자에게 제공할 수 있도록 한다.

2. 본론 (1)

기존의 TTS(Text to Speech) 기술인 Tacotron를 이용하여 Audio의 음질을 개선하고자 한다. 여기서 우리는 GAN (Generative Adversarial Network) - Super Resolution 기술을 응용하여 음질을 개선시킨다. 원래 GAN-Super Resolution (이하 SRGAN) 기술은 음질이 아닌 화질개선을 위한 비지도 학습 머신러닝을 의미한다.



(그림 1) 실제 SRGAN을 활용한 복원 이미지

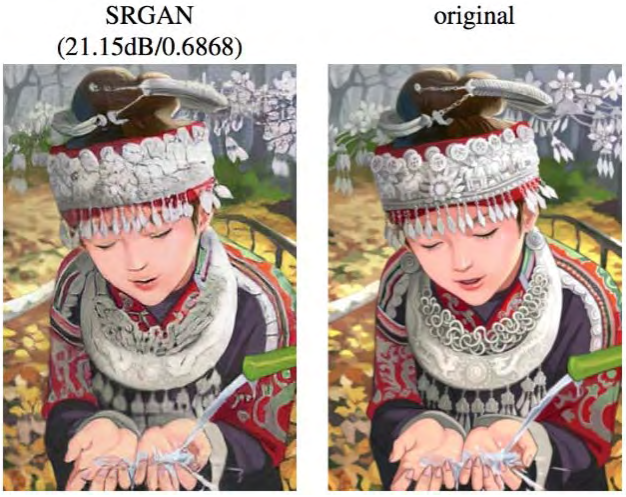


그림 1과 같이 SRGAN 기술은 저화질의 이미지를 고화질의 이미지로 화질을 개선 시켜주는 것이다.

이러한 SRGAN 기술을 이미지가 아닌 Audio 에 적용하여 저음질의 소리를 고음질의 소리로 개선해주는 것이 주된 목적이다.

3. 본론 (2)

기존의 GAN 기술을 통해 기계를 학습시키기 위해서는 '원본의 고화질 이미지'와 '저화질 이미지'의 한 쌍 (Pair)이 필요하다. 이러한 한 쌍의 이미지를 통해 저화질 이미지를 어떻게 고화질 이미지로 바꿔줄 수 있는지를 기계가 학습하게 된다. 이러한 한 쌍의 데이터를 반복 학습시키게 되면 최종적으로 임의의 저화질 이미지가 들어왔을 때 해당 이미지를 가장 적절한 고화질 이미지로 만들어 주는 것이다.

이미지와 마찬가지로 한 쌍의 음원이 필요하다. 원본이 되는 고음질의 음성과 저음질의 음성을 학습시켜주어야 한다. 이때 고음질로 분류되는 음성은 실제 사람의 목소리를 녹음한 원본 음성이 되고, 저음질로 분류되는 음성은 고음질로 분류되었던 실제 사람의 음성을 학습시켜 TTS를 통해 나오는 Output이 된다 .

(그림 2) GAN을 이용한 음질개선의 Flow Diagram

그림 2 에서와 같이 먼저 Tacotron에 직접 녹음한 음성파일과 (나중 GAN에서의 고음질 원본 음성) 그에 상응하는 Text파일을 Pair데이터로 넣어 학습을 시킨다. 학습되어 있는 Tacotron에 원하는 문장 Text를 입력하게 된다면 TTS 기술을 통해 audio가 나오게 된다. (나중 GAN에서의 저음질의 음성데이터)

직접 녹음한 고음질의 원본 음성데이터와 Tacotron의 TTS기술을 통해 나온 저음질의 음성데이터를 한 쌍으로 묶어 GAN Audio Super Resolution (이하 A-GAN)기술에 학습 시켜준다. 학습이 완료된 A-GAN 시스템이 임의의 저음질 음성을 고음질 음성으로 바꾸어 준다.

(그림 3) 각각의 언어에 따른 다른 데이터 셋

앞서 설명했듯이 Tacotron을 학습시키기 위해서는 Text와 Aduio의 Pair데이터가 있어야한다. 즉 언어가 달라지면 각각의 언어에 따른 다른 데이터를 학습시켜야 한다. 이에 대해 해당기술에 한국어와 영어, 이 두 가지 언어를 가르쳐 결과물을 도출해 낸다.

4. 결론

따라서 개발하고자 하는 A-GAN 기술의 목적은 기존의 TTS 기술을 보다 깨끗하고 잡음과 기계음이 없는 원본에 가까운 음성을 바꾸는 것이다.

저음질의 음성을 고음질의 음성으로 바꿔주는데 있어서 기존의 노이즈 제거 기술과는 아예 다른 기술이 사용된다. 프로그래밍 되어있는 기술이 모든 음성 파일에 일정하게 적용되는 방식이 아니라 A-GAN 기술을 통해 기계가 스스로 저음질과 고음질을 학습하여 더 나은 결과물을 도출해 낸다. 향후 이 기술이 더욱 발전한다면 모든 TTS 기술에 적용될 수 있을 것이고, 사용자들은 Text를 입력했을 때 더 자연스러운 Audio 음성 파일을 제공 받을 수 있게 될 것이다.

Acknowledgments

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 "(2016-0-00017)

자료 출처

[그림 1 - 논문]

Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial

Network

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, ´

Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi

Twitter