Podcasts Deepfake Audio with Multilingual Representation

Sai Teja Gudipati (20825009)

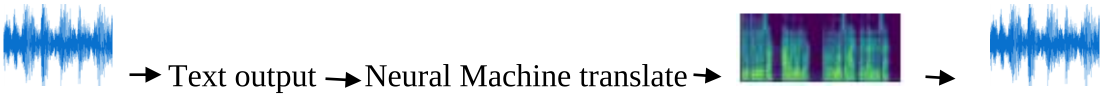
Shadman Raihan (20858688)

Somesh Kumar Gupta (20817245)

Abstract

In this realm of media content consumption, podcasts play a prominent role in its exponential growth. So, we thought that the same podcasts content should be reachable to all diversities of people with their favorite “artistic voice and language”, irrespective of which voice the original content is delivered. To achieve this with robust and reliability, Convolutional Neural Networks, NLP and GAN’s are key building blocks of our model. There are three key stages involved, first the raw audio to text conversion, second language modelling and third inversion back to audio. As this case is speech, mel-spectogram and aligned linguistic features are the intermediate representations with phoneme boundaries.

In this project, we will be considering the publicly available podcasts dataset for initial process to generate the text data and perform sequence to sequence modelling with 4 different languages. The obtained data will be given to a feed-forward convolutional architecture to generate the audio waveform in GAN setup. For speech synthesis we will be initially training the CNN with open voice datasets of any famous person and evaluate the metric with Mean Opinion Score (MOS). We will consider the traditional speech synthesis architectures such as Google Wavenet, DeepVoice2 for our architecture design and parameters.



→ Text output → Neural Machine translate → →

Raw audio

mel-spectogram

deep fake voice

References:

[1] Jeffdonahue, Sedielem, Binek, Eriche, Simonyan. End-to-End Adversarial Text-to-Speech. arXiv: 2006.03575, 2020

[2] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. arXiv:1910.06711v3, 2019

[3] Melvin Johnson, Mike Schuster , Quoc V. Le, Maxim Krikun, Yonghui Wu, Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Mitpressjournals 10.1162, 2017

[4] Sercan Ö. Arıka, Gregory Diamos Andrew Gibiansky. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. arXiv:1705.08947v2, 2017

[5] Sean Vasquez and Mike Lewis. MelNet: A generative model for audio in the frequency domain. arXiv:1906.01083, 2019