Brainwave: A Multidisciplinary Journal (ISSN: 2582-659X), Vol. 2, Special Issue No. 1, May 2021,
pp. 115-118, © Brainware University

Voice Analysis and Replication with Transfer Learning

Krishnendu Saha[1*], Reeju Bhattacherji[1], RishavBhattacherjee[1], Partha Pratim Dasgupta[2]
[1]3rd year BCA student, Department of CSS, Brainware University
[2]Department of CSS, Brainware University
**saha.krishnendu1998@gmail.com*

**Corresponding Author*

Abstract

In this cyber era, Artificial Intelligence (AI) is the ultimate thing. We all want AI as our assistance. And if we are talking about AI, then artificially synthesize voice comes first. So, we are here with a design for artificially cloned or synthesize human voice. In this design, we have three methods (1) encryption of speaker's voice, after taking a noisy speech, to generate a fixed-dimensional embedding vector from only a few seconds of reference speech from a target speaker and store it in a database in an encrypted form as like a module to use it for generating a human-like mechanical voice (2) speech to the text module, through this module we are going to synthesizes the speaker's voice and analysed the dataset to give it a textual form (3) text to speech, it is one of the valuable modules, here we are going to transform the textual state into a synthesize vocal or wave format which depends on the dimensional embedding vector form of the speaker's voice. Throughout these modules, with the learning and training process, the system is going to learn a high-quality replication of the human-voice.

Keywords: Transfer learning, Voice replication, Human voice analysis, Machine learning, Artificial neuron, Speech to Text, Text to speech, voice analysis

1. Introduction

The goal of our work is to take the STT (speech to text) and TTS (text to speech) to the next level towards the replication of human voice and take a step forward to make artificial intelligence more humanistic. We specifically address a zero-shot learning setting, where a few seconds of transcribed reference audio from a target speaker is used to synthesize new speech in that speaker's voice, without updating any model parameters.

Tech-Stack:

- HTML5
- CSS3
- Python (3.7)
- Pip package manager
- Django (3)
- venv
- MongoDB atlas

2. Methodology:

Initially the audio given as an input by the user would be processed to check for noises occurring in the background and later on remove them and modify the wave in a dynamic vector format, so that the Machine Learning model is able to use those data for training purpose to get the better results.

We are going to convert the audio file to numpy array form which would come handy for performing different scientific operations on the audio data.

For example, the audio handed over by the user is actually a summation of two sine waves of frequencies 50hz and 120hz respectively. But since we are showing how the noise is going to be

removed from the audio we would add white noise with a magnitude of 2.5 in that audio so that if we plot the audio data it would seem something like this:

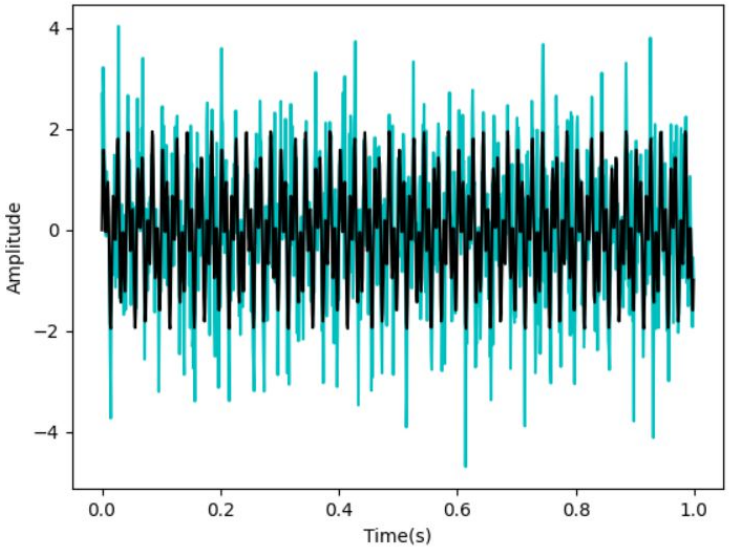

Figure 1: Amplitude Graph

The blue wavelength is the audio after the addition of noise and the black wavelength is the audio before the addition of noise. We are going to act like we only got the noisy data and show how to remove it.

First of all what we would do is perform Fast Fourier Transform on the numpy array data, what we would get after that is a vector consisting of the fourier coefficients. These fourier coefficients are complex values with a magnitude and a phase, the magnitude tells us how important that frequency is and the phase tells us if it is more sine or cosine and with what mixture.

Up next we would have to acquire the Power Spectral Density of every frequency present in the vector.

To understand what we are doing please observe following equations

a complex numberisλ
$\bar{\lambda}$ is the conjugate value of λ then,
$$\lambda\bar{\lambda} = |\lambda|^2 \text{ (Magnitude of } \lambda^2)$$
$$\lambda = a + ib$$
$$\therefore \bar{\lambda} = a - ib$$
$$\Rightarrow \lambda\bar{\lambda} = (a + ib)(a - ib)$$
$$= a^2 + b^2$$
$$= |\lambda^2|$$

We are trying to do something similar here as well. We are computing the magnitude of the square of all the fourier coefficients and we are going to get a vector of powers of each frequency. If 'f' is a

function of time and each frequency is in unit of per second that means the elements in the vector f_ is in unit of Hertz then we can actually convince ourselves that if we take f times its conjugate this will create units of power and so it is Power Spectral Density that we are computing in the variable PSD.

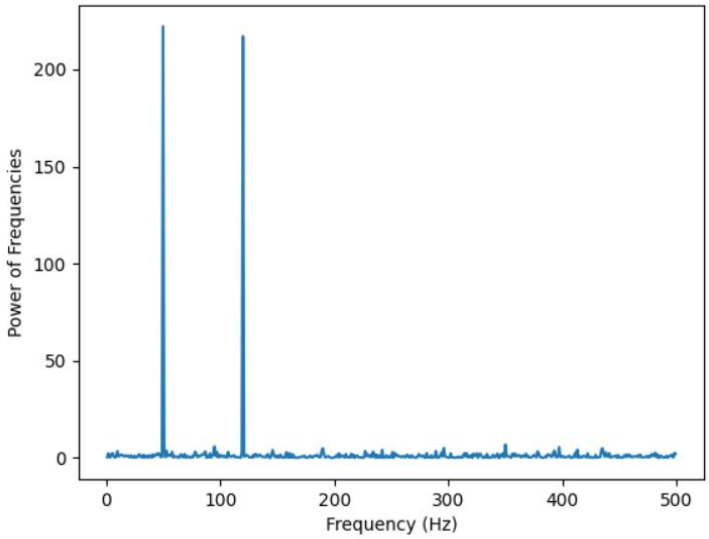

Figure 2: Frequency Graph

From the above graph of Frequency vs PSD our job is to eliminate all the fourier coefficients with lesser Power Spectral Density and in this case, closer to 0 PSD. The resulting Frequency vs PSD graph after eradicating all the fourier coefficients would look like this:

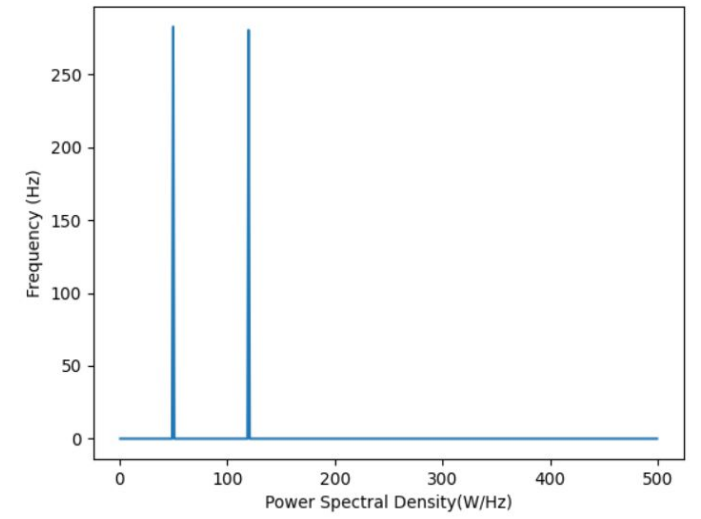

Figure 3: Power of Frequency graph

Well, our job should be almost done by now but, we cannot handover this vector of coefficients to the Machine Learning model so, we need to perform Inverse Fourier Transform to convert the vector consisting of fourier coefficients back to a numpy array form.

After Noise Cancellation the graph plotted, would look like the graph below:

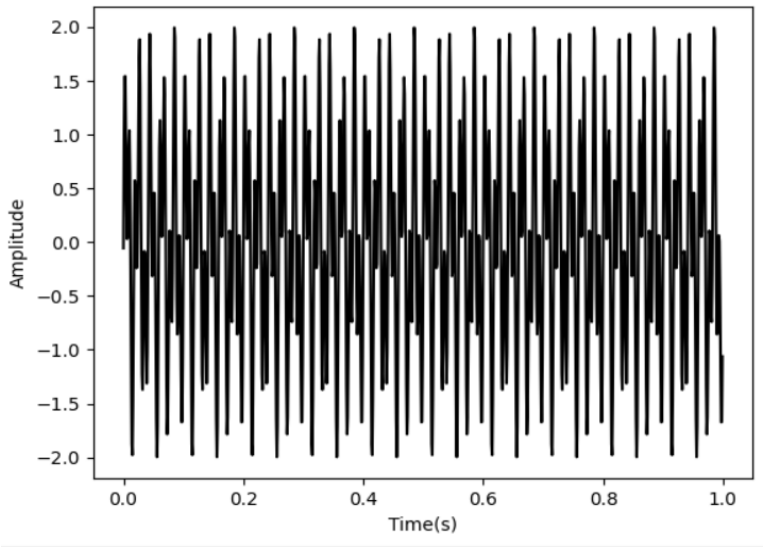

Figure 4: Amplitude Graph

After all the Noise Cancellation the filtered audio would be handed over to the Transfer Learning model which has the task of extracting the voice of the speaker from the audio and later on use that voice to create another speech which is the main ambition behind this project.

3. Conclusion

The documentation of this regarding project contains very few information, but the research is ongoing, with the help of various thesis papers and documentations, to make it as perfect as we can.

Reference

1. H Aldrich, C Zimmer, T Jones - The Sociological Review, 1986 - *journals.sagepub.com*
2. B Mehler, B Reimer, J Dobres, H McAnulty- MIT AgeLab Technical , 2014 - *academia.edu*
3. J Van Borsel, J Janssens, M De Bodt - Journal of **Voice**, 2009 – *Elsevier*
4. DM Saunders, BH Sheppard, V Knight - Responsibilities and Rights , 1992 – *Springer*
5. J Abitbol, P Abitbol, B Abitbol - Journal of **voice**, 1999 – *Elsevier*
6. G Flis, A Sikorski, A Szarkowska - Journal of Specialised Translation, 2020 - *academia.edu*
7. A Gutierrez Jr, KD Bennett, LS McDowell - Behavioral , 2016 - *Wiley Online Library*
8. EB Holmberg, RE Hillman, B Hammarberg - Journal of voice, 2001 – *Elsevier*
9. DR Ladd, KEA Silverman, F Tolkmitt - The Journal of the , 1985 - *asa.scitation.org*