**SPEECH TO TEXT CONVERSION**

**B.Tech2020.R.CCE.1.19CCE281 - SIGNAL PROCESSING**

**SEMESTER 3, COMPUTER AND COMMUNICATION ENGINEERING**

**GROUP 6:**

|  |  |
| --- | --- |
| **NATESH KUMAAR V** | **[CB.EN.U4CCE20038]** |
| **PRASATH V S** | **[CB.EN.U4CCE20044]** |
| **PRIYANGA R** | **[CB.EN.U4CCE20045]** |
| **R RANGASHREE DHANVANTH** | **[CB.EN.U4CCE20048]** |
| **RIYA KALLANKATTIL PRAMODH KUMAR** | **[CB.EN.U4CCE20049]** |

**ABSTRACT:**

We design a code to convert audio signals to digital signals by using the principles of Signal Processing. In a speech recognition problem, the input is in the form of an audio signal and the final output will be in the form of a text message. However, we can’t take the raw audio signals as they will have a lot of noise in it. Also, analog signals are more memory occupying since they have an infinite number of samples hence it’s important, we convert them to digital signals. For performing the conversion, we do the technique of sampling of signal i.e., basically selecting a few samples per second from the analog signal. By doing so we can capture the properties of the signal that are required for signal processing. The spectrogram gives us the spectral information or representation of the signal, which is obtained by using an algorithm called the Fast Fourier Transfer (FFT) and we can also characterize the signals by filter banks. Speech to Text conversion used in Google, Alexa, Siri to name a few applications uses a combination of deep learning and natural language processing (NLP).

**SYSTEM DESCRIPTION AND APPLICATION:**

Speech to text is the technique by which a device automatically recognizes the audio or words of an individual on the basis of the information in the speech wave and then converts it into a text output. The system is designed to take a speech signal or audio signal as the input and convert it into a text message, which is the output. That is, for an audio input “Hey, there” we get a text output hey there. It has major applications in a lot of modern-day devices like apple Siri, amazon’s Alexa, Windows Speech Recognition, and even normal voice-to-text conversion used in search engines like Google.

**METHODOLOGY:**

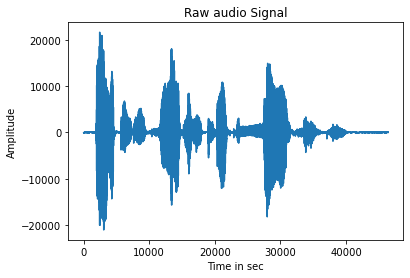
In speech recognition, the input will be the audio signal and the output will be the text form of that audio signal. The raw audio signals should be first converted from analog to digital signal which can be done by the A/D converter (analog to digital converter). To improve the signal quality, we do pre-emphasis which is done by the first order high pass filter which increases the magnitude of the energy at higher frequencies. For vowels it is observed that the energy in the higher frequency is lesser than the energy in the lower frequency, so by increasing the energy in higher frequencies we can improve the quality of the signal. Hamming window, used as a traditional computation of MFCC is used to reduce the spectral leakage from the spectrogram before the step of processing. Then the signal should be converted from the time domain to the frequency domain by applying DFT (Discrete Fourier Transform) because as the input is an audio signal it will be easier to perform the operations in the frequency domain than in the time domain. The way how humans perceive sound is very different from machines, human ears have higher resolution at a lower frequency and lower resolution at a higher frequency, whereas for machines the resolution is the same for all the frequencies. So, we use Mel-filter bank which can give us high resolution at a lower frequency and low resolution at a higher frequency. Hence, we use the log function on Mel frequency, thereby representing the human hearing system. Under the operation of IDFT, we get to know that the fundamental frequency in the frequency domain is considered as the lowest frequency, whereas in the time domain it is considered as the highest frequency. After the application of the IDFT operation, the first twelve coefficients of the signal is chosen by the MFCC model and also the energy of the signal sample. Next the MFCC technique performs first order and second order derivative of the features which again consist of twenty-six other features. Thereby in total the MFCC technique produces thirty-nine features from the audio signal samples, which serves as the input for the given speech to text conversion problem.

**SIMULATION RESULTS:**

**Results:**

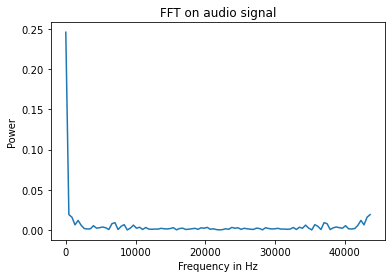
**The Audio file(.wav) is read and the signal is plotted in time domain**

**Time vs Amplitude Audio signal**

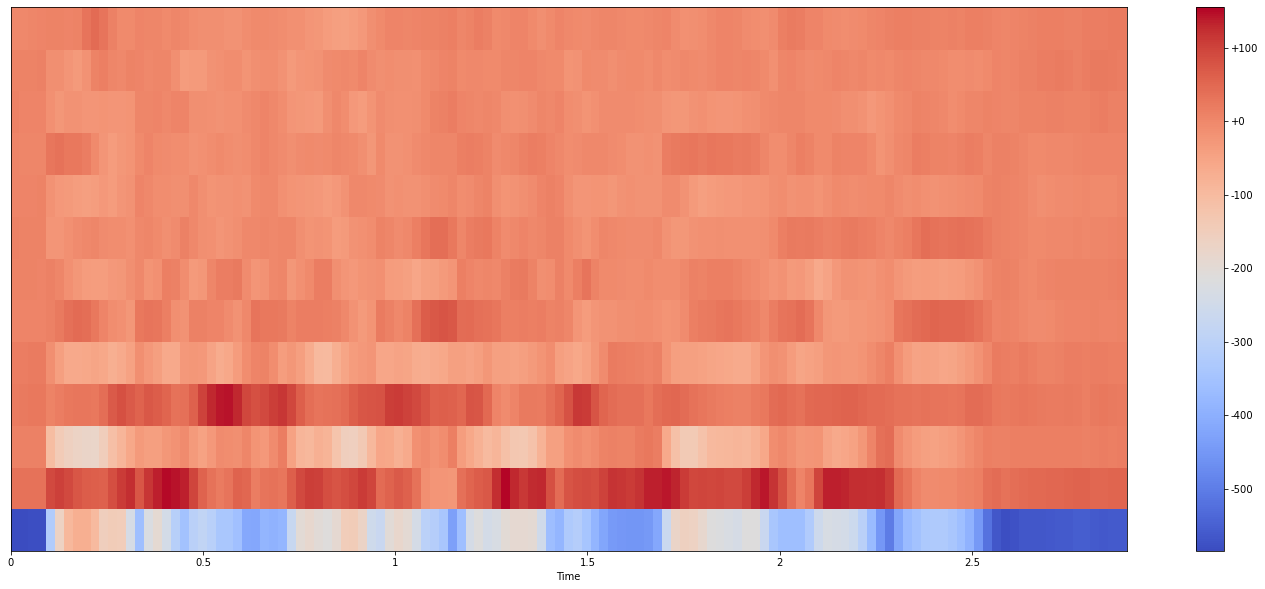


**The audio signal is transformed to frequency domain by performing Fast Fourier transform (FFT)**

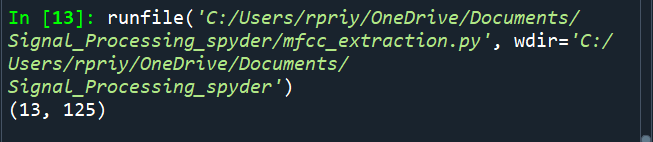
**Frequency vs Power graph of the audio signal**



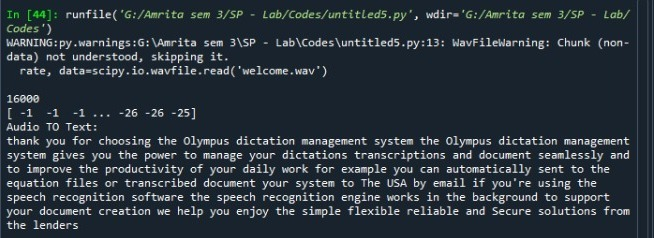
**MFCC is performed on the .wav file**



**There are 13 rows**



**The Final Speech to text conversion:**



**CONCLUSION:**

This report explains briefly the mechanism of speech to text conversion by using principles of signal processing. We get the spectral form of information from the segment of the speech or audio signal by performing Fast Fourier Transform. So initially we begin by converting the given analog signal (speech signal) to a digital signal (text message). We characterize the signals using filter bank and apply the processes like pre-emphasis, windowing, Discrete Fourier Transform (DFT), Mel Filter bank, applying log, and so on to get a complete code that performs a perfect speech to text conversion. Speech to text builds up the gap of communication between a man and machine, which will make it possible for us to have a natural conversation with our smartphones or any other smart electronic devices. Speech-to-text conversion follows many algorithms but the one we have chosen to work upon is the Mel-frequency cepstral coefficient (MFCC) algorithm.

**REFERENCES:**

<https://www.analyticsvidhya.com/blog/2019/07/learn-build-first-speech-to-text-model-python/>

<https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/>

<https://www.leftrightmind.com/blogs/the-future-of-voice-recognition>

<https://www.cs.rochester.edu/u/james/CSC248/Lec13>.

<https://www.ijert.org/zigbee-based-wireless-voice-to-text-translator>