Analyzing sounds from videos

amineelkhadiri

December 2020

1 Introduction

In a world where technologies and digitals became used every where and mostly all the time especially in the current world circumstances where courses has been carried out by many softwares and became included in every day life.this project aim to help deaf people understand and maintain the meeting in video calls proposed by some softwares like Skype and zoom because this layer of society is almost forgettable by the society chiefly in corona situation.one of the project goals is to let this social class benefit also from these kinds of software and be able to maintain courses.
This project tackles the study of sounds and voices extracted from videos.is has two particular insights,one comes from a scientific perspective through analyzing sound,fast Fourier transformation as well as giving some plots for further informations.the second approach emphasize on the content of this sound or voice by extracting speech from it and translate into subtitles so that we follow the speaker efficiently and have a good comprehension about his talking .at the beginning the original idea was whenever there is a video call which uses the camera and microphone,the program should normally write and translate what the interlocutor tells in different languages of choice and make subtitles with synchronization ,but soon i realized that was a really tough task because it involves 3 process working at the same time,details will be put later on the report.

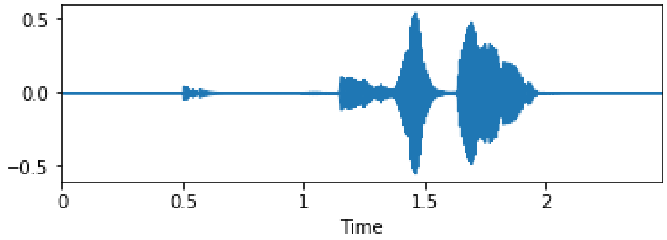
1.1 Scientifically

I tried in this section to get the signals directly from my microphone so i can analyze them and converts them.i made the machine converting the speech to a text in text file through an object called recognizer.furthermore i had plotted a couple of figures like signal waves graph(time domain),Fast Fourier transformation(FFT),spectogram and i ended with some spectrum real time animations.

1.1.1 Time domain:

The figure down below shows the evolution of signal waves of some random input voice through time.notice that the maximum magnitude can reach 0.5

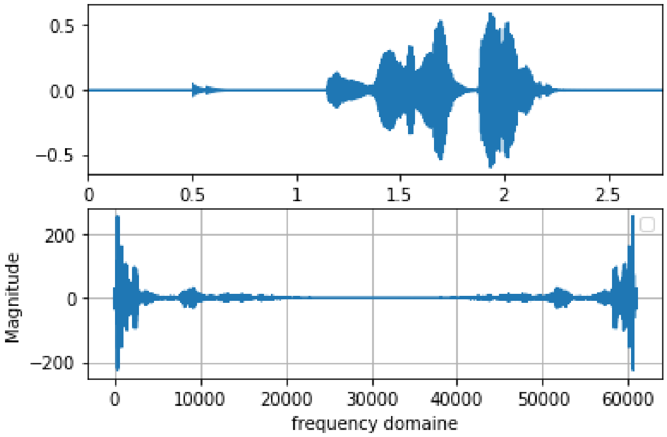
and the amplitude=0 means there is a silence. This amplitude is actually the amplitude of air particles which are oscillating because of the pressure change in the atmosphere due to sound.



The time domain in general is not that informative so we need to convert it to frequency domain in order to make things little bit clear and this is where the fast Fourier transform get in the game.thanks to it we can provide a very important amount of information showing the evolution of the magnitude by frequencies.

1.1.2 Frequency domain:Fast fourier transformation

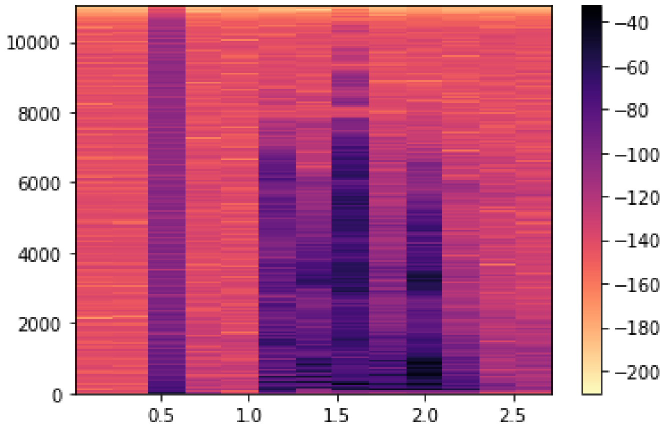
Fast Fourier algorithms calculate Discrete Fourier Transform(DFT) of a given sequence.it has made a huge progress in term of efficiency and speed compared to its sister Discret Fourier transformation that is why itis well used in many fields like digitals world,audio and image compression and satellite TV.the DFT is very expensive and much bigger when it is about audio signals or images in terms of complexity (O(n^2)) whereas the fft return the same output with order of (O($n \log n$)). the graph down below shows two domains one of time and the other of frequencies,each frequency has its own magnitude.



we observe a very high amplitude near 0 approximately around 0 and 3000 Hz and in the end around 60000Hz with amplitude=250 and there is also a small amplitude around 8000Hz.the fast fourrier algorithm returns a list of positive and negative frequencies which normal to see the symmetry.You can pick out any one half and calculate absolute values to represent the frequencies present in the signal.

1.1.3 Spectrograms

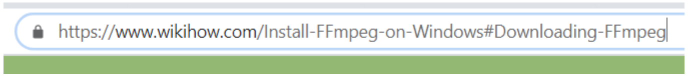
Spectrogram is the best way to visualize things where we mix both time domain and frequency domain.it is very useful for clinicians and dialectologists.so i plot the spectrogram on the same voice data as we can see down below.



on vertical axis we got the frequency,time on horizontal axis and the amplitude (amount of sound energy) as either darkness or coloration. we can see again between 1 and 2.5 second we got high concentration color in other words high amplitude between 1 to 2.5 seconds around about 8000Hz.
Notice:
I've done a real time spectrum aimation as well as the real time fft transformation.keep in mind that this code did not work on spyder so i had to move to Jupyter notebook.see the notebook file .

1.2 Speech recognition,live stream and subtitles

Before we get into details you have to download the ffmpeg library through a link that i will put right after this where it shows how to do it ,there is some simple modification that needs to be done,it is easy and wont require so much.the link:


https://www.wikihow.com/Install-FFmpeg-on-Windows#Downloading-FFmpeg

This section is special because im going to describe my progress during this project because lot of things happened here to arrive to the result of making translated subtitles on live stream for deaf people.So in the beginning i intended to overtake and call the devices on my laptop through some specific libraries like Pyaudio and open cv in order to record the audio and the image at the same time,then convert the speech on subtitles and share it on the same screen .i thought that is it was possible to do that but unfortunately it was a quiet naive approach because later on i had discovered that these to libraries woudn't give me the result i wanted that is making a sort of synchronization between the video,the audio and the subtitles which is not a simplpe thing to do.so i coundn't combine those two,they were independant packages and here i learned about FFmpeg package.it was a perfect one for my needs,it is a complete platform solution to record,convert and stream audio and video with synchronization.

With this project report i shall share 3 script files,one about spectrum animation mentioned earlier type Jupyter notebook,the second is the first attempted with Pyaudio library.even though it wasn't a successful shot but i managed to do some interesting things.i was able to write a program that can hear the speech on a real time from the microphone and convert it to a text and even translate it in any language of choice.
The Third one is "recordoffline.py" is the file where i used this library.the program activates both camera and microphone ,here you should speaks in french then after 8 second it will ends and you will get the output as output.mp4 made with subtitles. notice that on the second file,i mentioned at some point "translation",actually if it doesn't work by either showing no translation or resulting in an error like this:

$AttributeError$: NoneType object has no attribute group

Then you have to uninstall the current googletrans version and install the new one using the following commands:

(pip uninstall googletrans) and (pip install googletrans==3.1.0a0).

As i said before the original idea was to make a live streaming where the program write the interlocutors speech and convert it into subtitles but it was hard to do it because it uses a lot RAM. we got 3 process that working on the same time :
1-audio camera process where we activate our devices.
2-speech recognition process,capturing audio files and converting to text.
3-taking the text and concatenate it so to have a readable sentence and put i back on the video. remember that all this stages should be applied at he same time with synchronization and that what makes things complicated to do.

Notice:

On the recordoffline.py file please change the name of the camera and microphone and enter yours at line 9 you will see it in this form:
video="....":audio="...." .

Thankyou