Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Ciudad de México

Project 1
Voice Characterization in time and frequency domain

Random Processes

PhD. César Vargas

Alfredo Zhu Chen	|A01651980
Luis Arturo Dan Fong	|A01650672
Juan Pablo Valverde López	|A01656127
Andrés Islas Bravo	|A01339391

27/03/2021

INTRODUCTION

The voice is a characteristic from most of the people who surround us and since the start of the times people have learned to identify each other by correlating the sound to an individual. Nowadays voice characterization is highly used as a voice recognition tool, as well as for other voice related applications such as the identification of emotions and biometrics [1], however the technology we know today has been in constant evolution from a couple of years ago until today. Around 1980 the first voice recognition technology was implemented mainly for childrens toys such as “Speak and spell” in 1978 and “Julie” an interactive doll created in 1987, both used speech chip technology[2].

In 1990 the company “Dragon” released the first prototype of what evolved to be the modern technology we know today. The “Dragon Dictate” debuted as the first speech recognition system and by 1997 the technology evolved into “Dragon Naturally Speaking” which is a software design for voice recognition. In recent years other companies such as Google, Apple and Microsoft have debuted their own AI assistant with voice recognition technology [2].

As humans it is important to keep innovating and developing new uses for current technology and turning them into new products. A modern use for voice recognition technology are the well known AI assistants from companies such as Microsoft (Cortana), Amazon (Alexa), Google (Google Assistant), Apple (Siri), Samsung (Bixby). All these AI assistants are much more advanced than a voice assistant, this new technology is capable of performing complex tasks and keeping a record of patterns and users. AI assistants are constantly evolving and in a few years they will perform tasks even more complex than the ones they are able to do today.

In the near future applications such as real-time audio translation will be available; companies such as “Waverly Labs” are constantly working on them. It shouldn’t be long until the technology debuts. Likewise the future implementation of voice recognition in the marketing area promises to make processes much more efficient, by analyzing voice and characterizing the results in look for different responses to a marketing campaign. Voice recognition and analysis technologies will keep on evolving and it looks that it will bring us an infinite amount of new opportunities [3].

New applications for voice recognition such as biometric security using voice fingerprinting are fairly new in the market and its development keeps scaliting every day, using more complex algorithms and technology. In newer voice fingerprinting a technology known as neural networks is implemented. Nowadays voice fingerprinting is used as a biometric security lock for several services and devices. It is divided into two main phases known as training and testing. In the training phase several samples are taken to teach the program to identify the characteristics of a particular voice, this is made by performing a Fast Fourier Transform to turn each sample from the time domain into frequency domain and using the Mel-Frequency Cepstrum Coefficients (MFCC) to obtain the energy feature of the given signal and keep it as the desirable result or match . Once the neural network is trained the testing phase can start. In the testing phase voice input is obtained by a microphone and both operations (Fast Fourier Transform and MFCC) are performed to obtain the energy feature and compare it with the desirable result [4] .

Other new applications for voice recognition lie in the modern techniques for analyzing sound signals and identifying them as voice signals. They are greatly used in voice recognition software in which the input signals are broken into individual sounds using filters and algorithms to find the most probable word fit from the language. A technology known as Natural Language Processing (NLP) and deep neural networks is a more complex way of understanding how machines identify voice. Once an input is given the software breaks it into bits it can understand, converts it to digital format and analyzes certain pieces of the content. Once those pieces are analyzed the software uses the training given by deep neural networks to create a hypothesis about what the user is trying to say and turn it into text format [5].

Another important aspect of voice recognition is the privacy and adequate treatment of speech data that is carried on the Internet. Speech data is very sensitive and richfull for many applications, such as biometric recognition for identification purposes. Biometric information is often used later in these years as a replacement of traditional passwords, this use due to the possibility that the user forgets the password or someone even could steal that information. But there are still some considerations to be taken into account in order to protect and preserve the privacy of the speaker. With this being said, there are a number of potential privacy-preserving techniques which concerns the protection of the reference model and the probe in a encrypted domain, some examples are mentioned by Nautsch et al [6] in their paper *Preserving privacy in speaker and speech characterisation*, such as Paillier cryptosystem-based methods, GMMs using secure two-party computation, distance-preserving hashing techniques etc. The same paper also discusses some evaluation measures on data privacy to tell how good is privacy preserved in biometrics algorithms such as voice recognition. It is important to establish this in order to accomplish the requirements for effective privacy preservation.

OBJECTIVE

Determine the time and frequency representation of the audio signal, differentiate and compare the same text said coming from different people, with the objective to characterize voice signals that would help for future speech recognition.

MATERIAL AND METHODS

- 1 PC or laptop with at least 4GB RAM and 3.1GB of available space.
- MATLAB R2019a or later versions.
- Script $'myvoice2.m'$provided by the instructor

The first step to do in the project is to record the voice signal, this can be done by defining an MATLAB *audiorecorder* object and calling the function *recordblocking(audiorecorder,t)* with *t* as the amount of time to record. After this, the data is stored in a double-precision array using the *function getaudiodata(audioRecording)*. It is important to consider that most of the signal could be at zero because of the waiting time before and after the recording, since this could affect the statistical analysis in time domain, the *nonzeros(A)* function is used to remove all the values of zeros from the recording only for time domain analysis.

By defining the time axis and the range of frequency with the number of samples and the time of the recorded audio, it is possible to show the voice signal in time and frequency domain with MATLAB *plot(x,y)* function. For the frequency domain, function *fft(y)* is used to get the Fourier transform of the signal, the zero-frequency component of it is rearranged with *fftshift(y)* and the absolute value is taken with *abs(y)*.

To conduct a statistical analysis of the amplitude of the voice signal in time domain, several functions of MATLAB are useful to accomplish this task. The mean, variance, standard deviation, skewness, kurtosis, dispersion are calculated and will be shown later for analysis purposes(see appendix for details of the code). The histogram, cdf and pdf of the signal are also plotted with *histogram(y),cdfplot(y)* and *histogram(y,"Normalization","pdf")* respectively.

The autocorrelation is also calculated using autocorr(y,"NumLags",value), it is important to specify the number of lags with the second and third argument of the function, this is because the function takes as default the minimum value between 20 and *T-1*, where *T* is the effective sample size of signal *y.* The Fourier transform of it is also obtained and both of them are plotted. Additionally, the magnitude square of the frequency domain representation, also known as the spectral density, is also calculated and plotted.

In order to store the voice data into the computer and reproduce it for analysis, a *InfinityRecording.m* function is created. The main idea of this function is to record the voice of the person and according to an input given, it can repeat the process several times. This would be useful to store lots of samples and analyze them, such as it is in experiment 2. The recording process is similar to the one as mentioned above, but this will be inside a *while* loop, which asks the user after each recording if another repetition is required and the commands entered in the command window are used to indicate this. Each voice sample is concatenated vertically in the variable *myRecording* (see appendix for details), so as it is required to average the samples in experiment 2, the averaged signal is created from all the samples with *mean(y,2)*, the 2 in the second argument is used to indicate that the average is taken by row.

RESULTS AND DISCUSSION

Experiment 1

Define a sentence or phrase to be read by different people.

- Phrase defined to be read: “Hola Hola”.

1. Plot of the recorded voice in time and frequency.

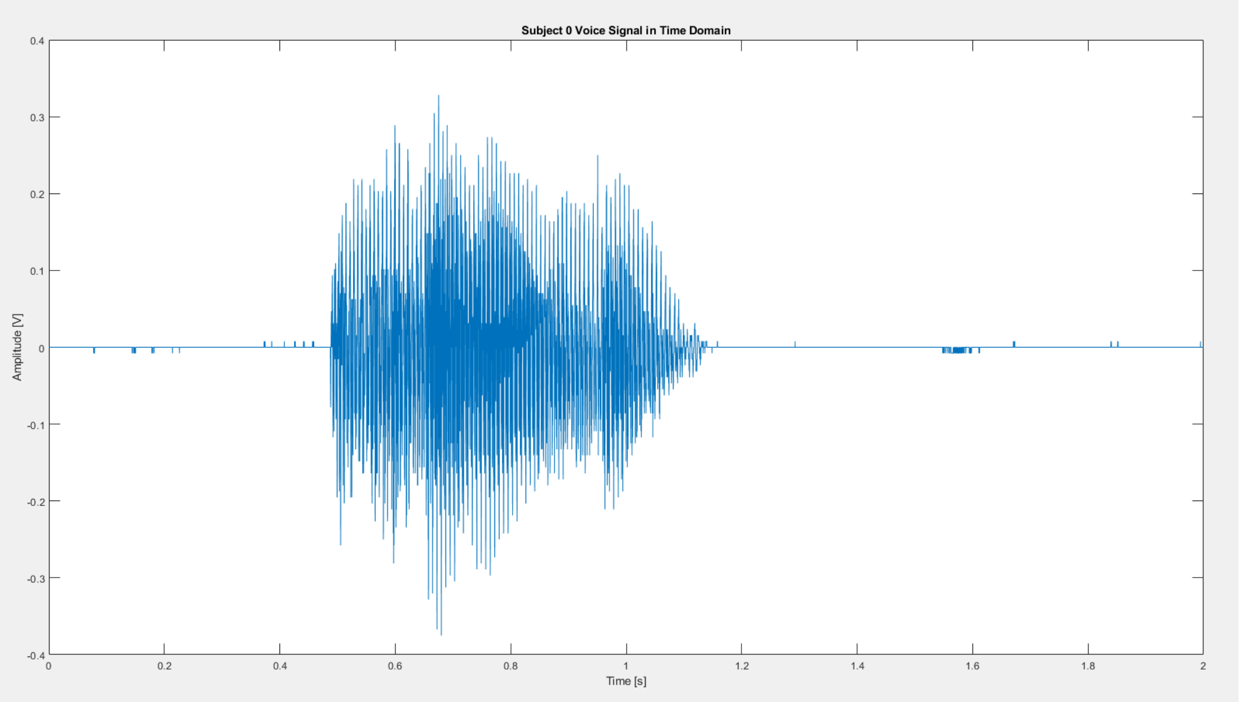

Figure 1. Voice in time domain

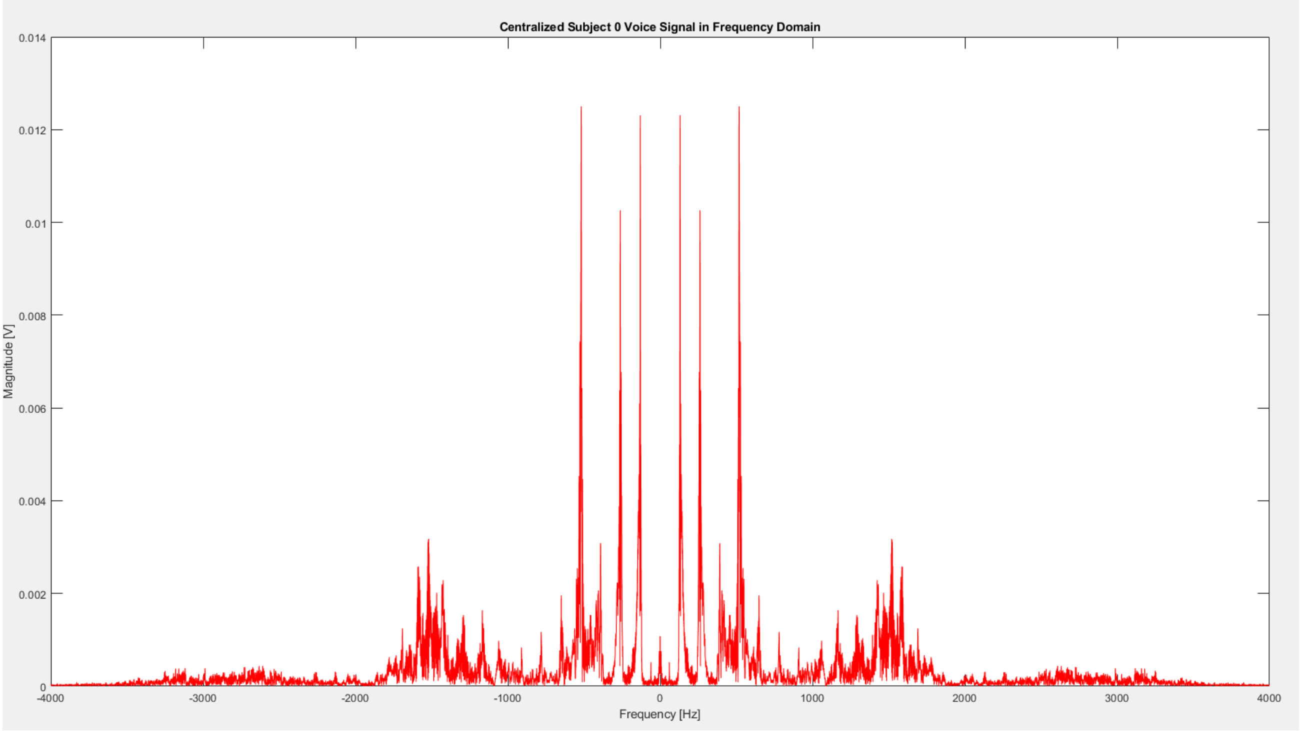

Figure 2. Voice in frequency domain

1. Conduct a statistical analysis of the amplitude of the voice in the time domain
   a. Histogram

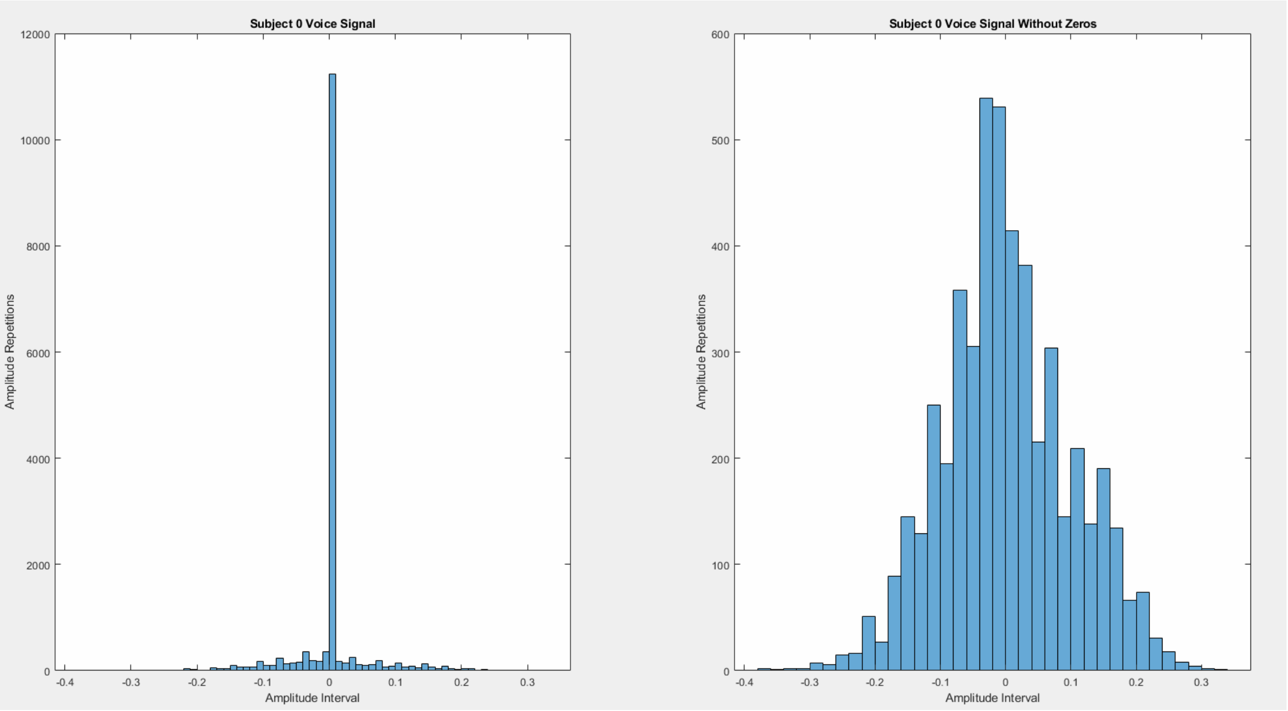

Figure 3. Histogram of voice signal

Table 1. Mean, variance, standard deviation, skewness, kurtosis, dispersion ($\frac{var}{std}$) of subject 0 voice recording

Statistical Parameter	Subject 0 Voice Signal With Zeros	Subject 0 Voice Signal Without Zeros
Mean	0.00046191	0.0014766
Variance	0.003043	0.0097277
Standard Deviation	0.055163	0.098629
Skewness	0.26363	0.11661
Kurtosis	9.6869	3.0257
Dispersion	0.055163	0.098629

b. CDF

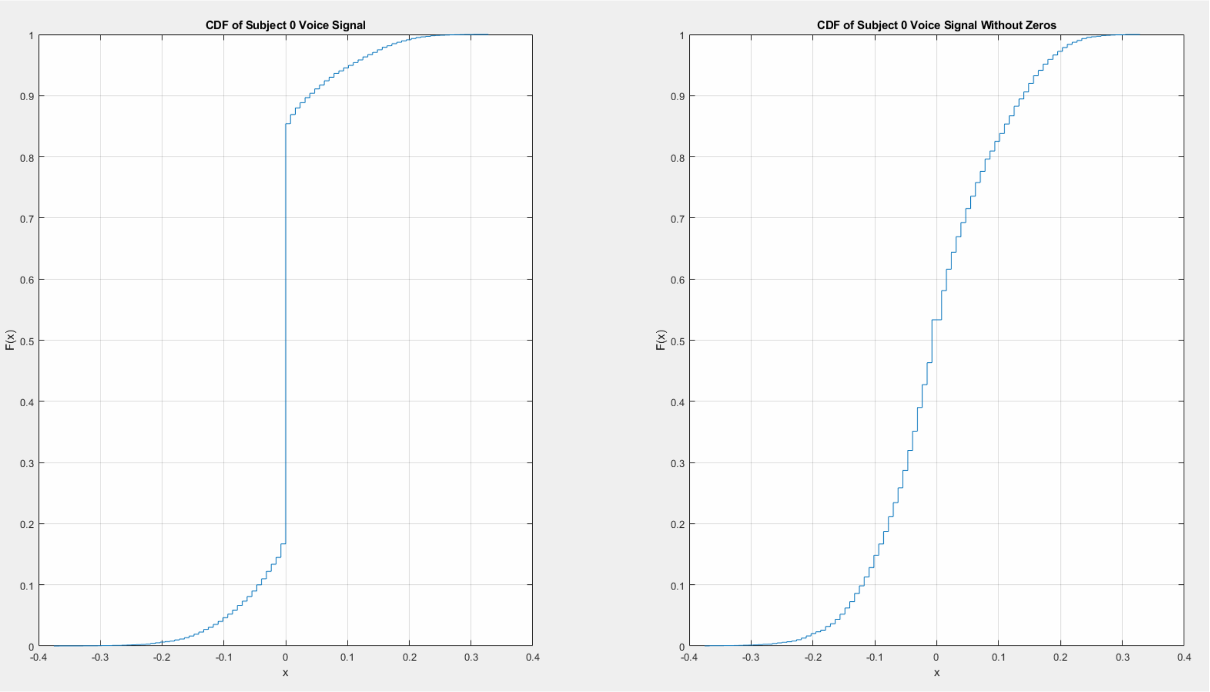

Figure. 4 CDF of voice signal

c. PDF

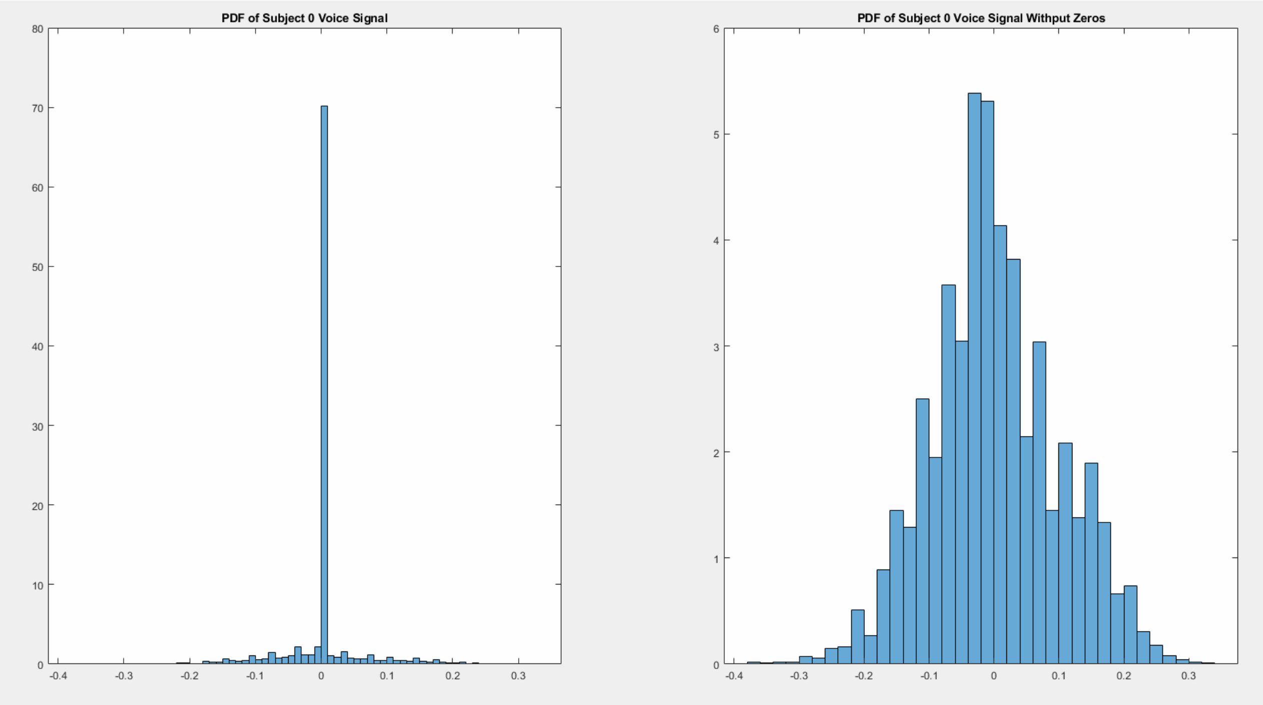

Figure. 5 PDF of voice signal

2. Obtain the autocorrelation of the voice signal

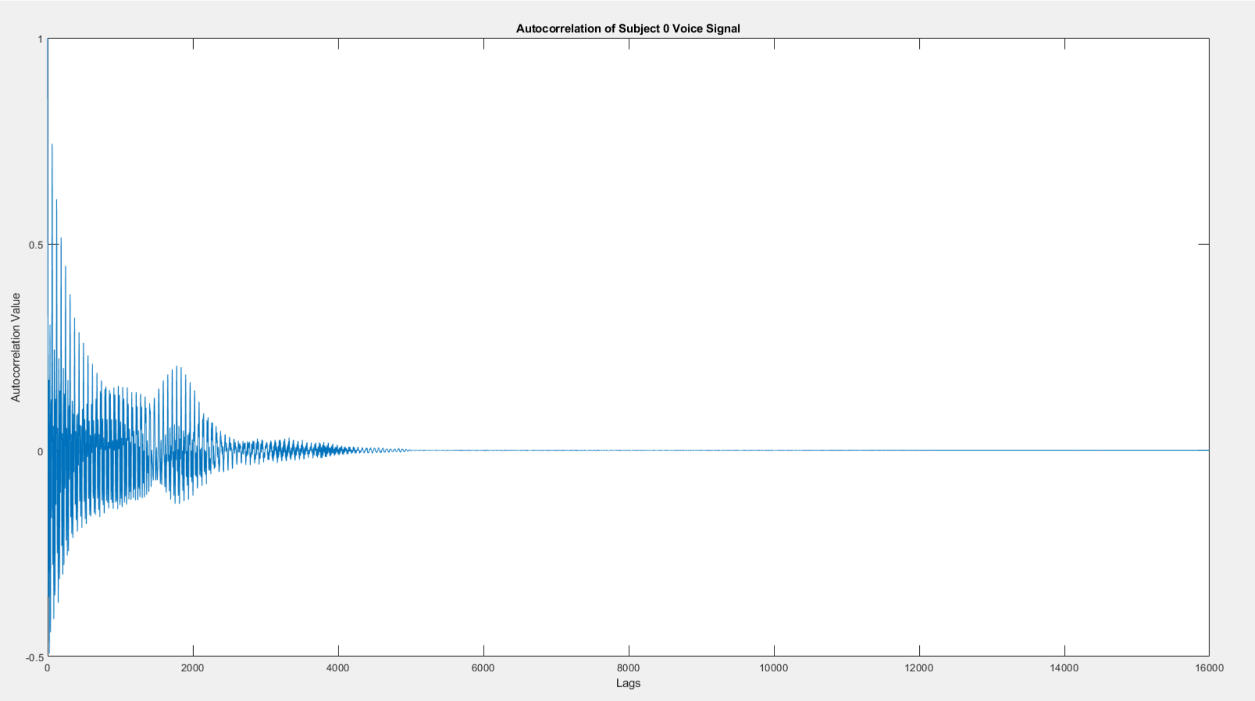

Figure. 6 Autocorrelation of voice signal

3. Since the frequency domain representation of your voice signal (Fourier transform) is a complex signal, get the magnitude square (spectral density) of it and plot it.

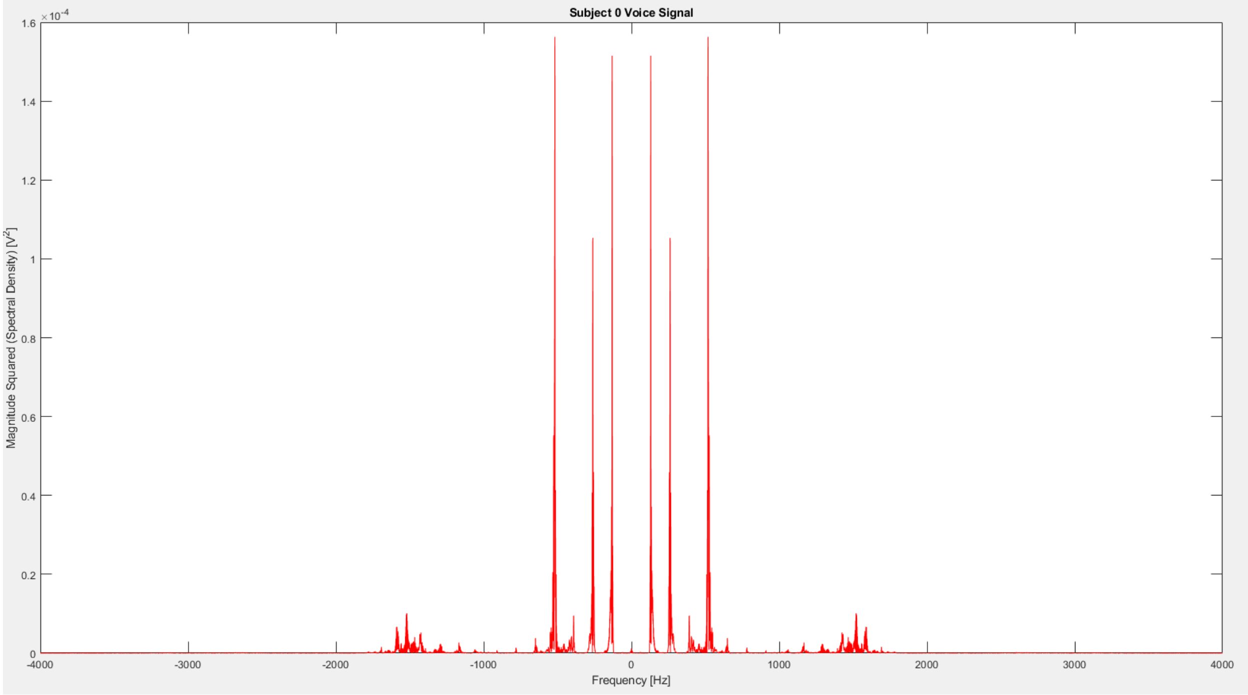

Figure. 7 Magnitude square (spectral density) of the voice signal.

4. Apply Fourier transform on your autocorrelation and plot it.

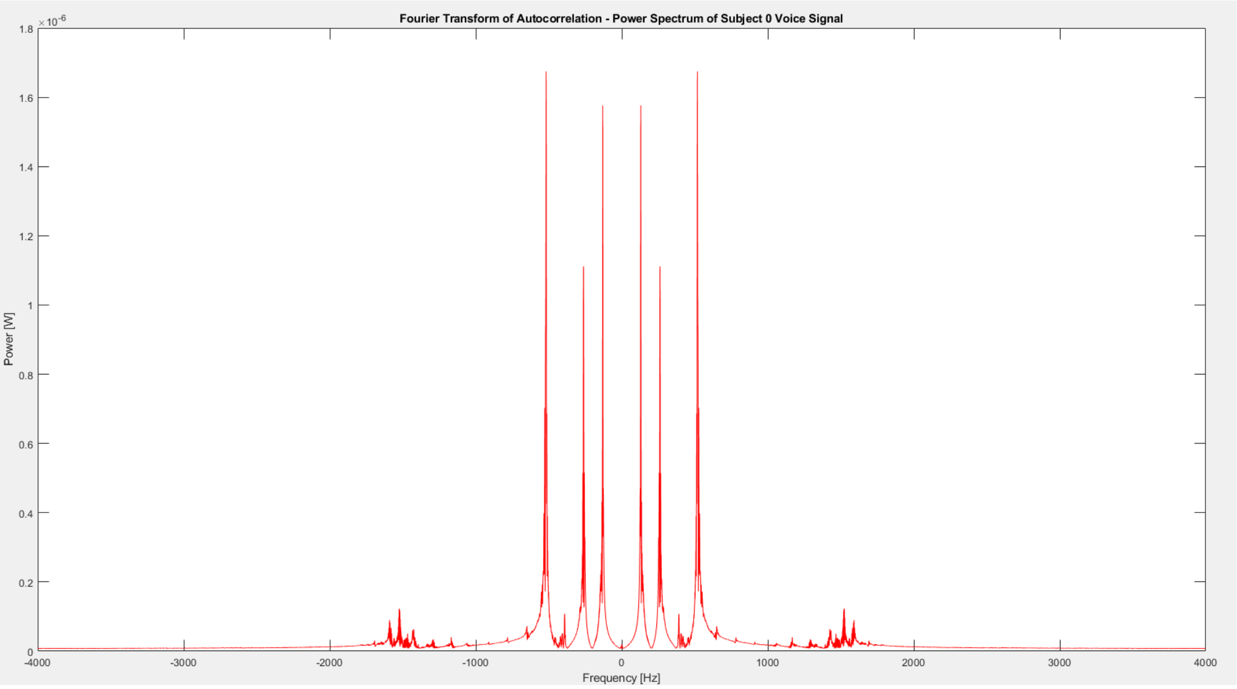

Figure. 8 Fourier transform of the autocorrelation of the voice

Experiment 1 analysis:

As the user does not start and finish speaking in synchrony with the recording statements, the signal presents a significant amount of zeros in the peripheria. Those zeros were significant for the time analysis, where the zero was huge at the histogram. But, for the frequency analysis, those extra samples were not taken into account; given a first useful characterization of the signal, which was improved with the magnitude square, making the important information bigger and the noise (values with amplitude smaller than 1) smaller.

The statistical parameters give valuable information. For example, as the data to be processed was a voice signal, there was expected an equivalent behaviour with respect to a sinusoidal. Which was confirmed with the mean pretty close to zero. On the other hand, parameters such as variance, standard deviation, and dispersion gives favourable signals as their values were considerably small. The smallness of those values were a good indicator that the data is concentrated in some values, these could be used in favour for speech recognition. Also, with special relevance, one of the kurtosis values was great enough to be considered a leptokurtic, and the other one, the analysis "without zeros", despite of not being leptokurtic, was mesokurtic, which is also for the previous characterization inferences.

Last but not least, the Fourier transform of the autocorrelation corresponds to the power spectrum. This information is useful as its behaviour is equal to the "Magnitude square (spectral density) of the voice signal", with an important different quality of not taking into account the phase. This is useful when the variable is aleatory, because, as there is no knowledge of the process, each sample will have an aleatory phase. In other words, the Fourier of the autocorrelation allows the frequency analysis of unpredictable signals, as in this case the voice could be.

Experiment 2 - Repeated Recording

1. Plot of the recorded voice in time and frequency.

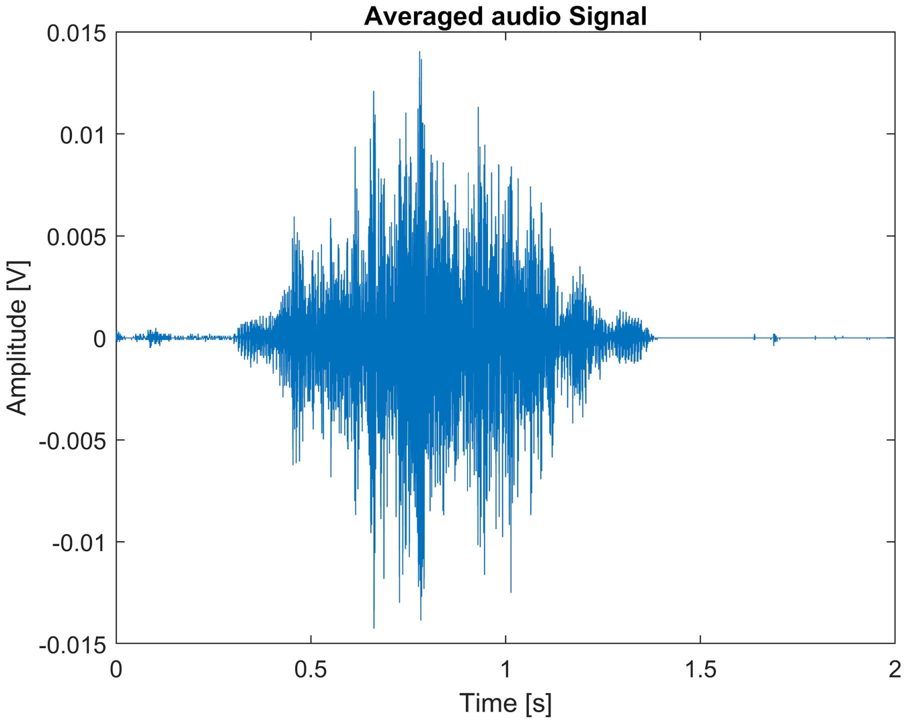

Figure 9. Averaged voice in time domain

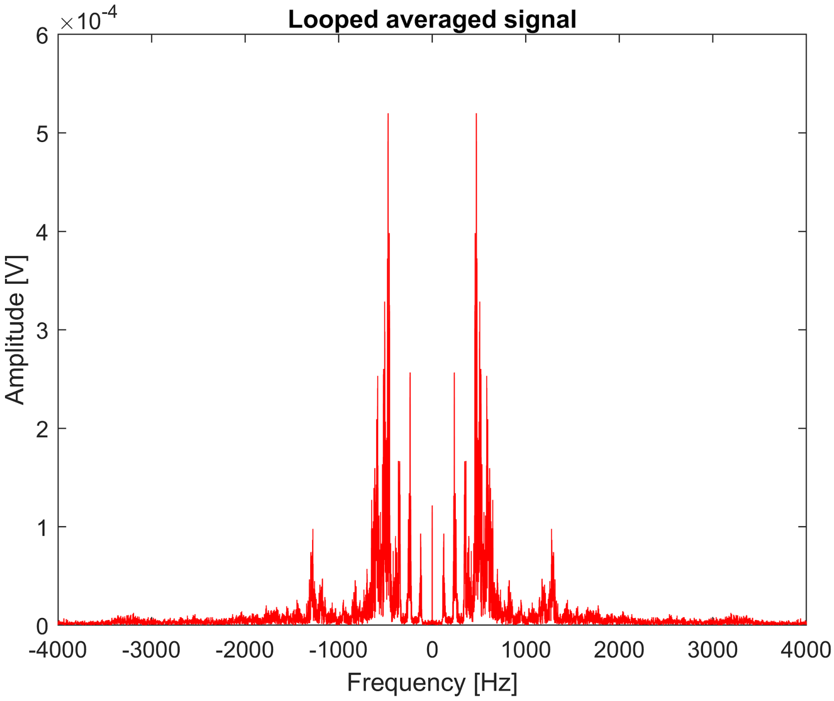

Figure 10. Averaged voice in frequency domain

2. Conduct a statistical analysis of the amplitude of the averaged voice in the time domain
    a. Histogram

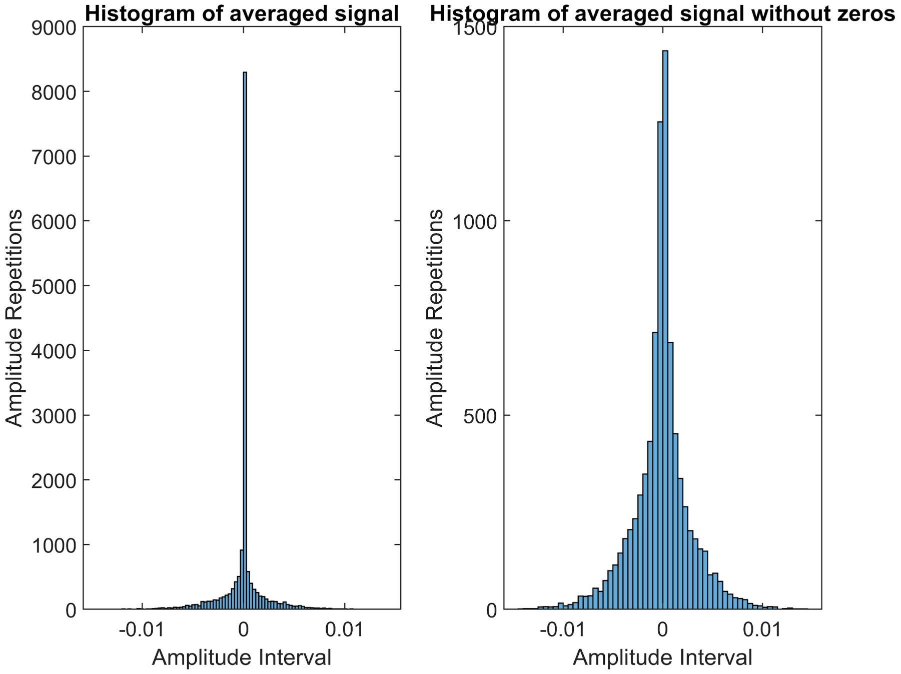

Figure 11. Histogram of averaged voice signal

Table 2. Mean, variance, standard deviation, skewness, kurtosis, dispersion ($\frac{var}{std}$) of averaged subject 0 voice recording

Statistical Parameter	**Subject 0 Averaged Voice Signal With Zeros**	**Subject 0 Averaged Voice Signal Without Zeros**
Mean	-0.000060754	-0.00011134
Variance	0.0000046266	0.0000084732
Standard Deviation	0.0021509	0.0029109
Skewness	-0.27912	-0.15426
Kurtosis	10.247	5.5864
Dispersion	0.0021509	0.0029109

b. CDF

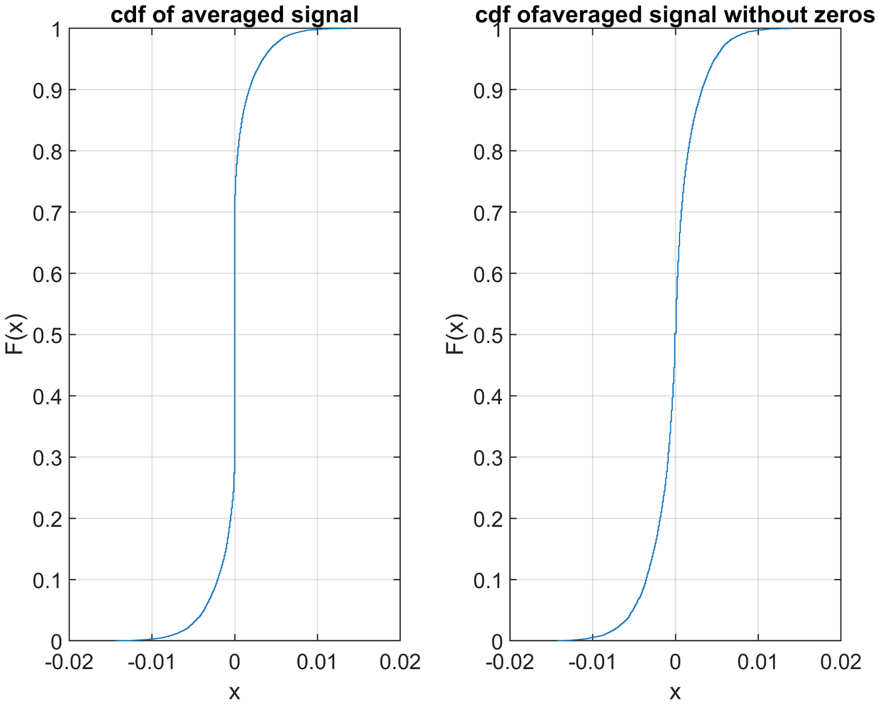

Figure. 12 cdf of averaged voice signal

c. PDF

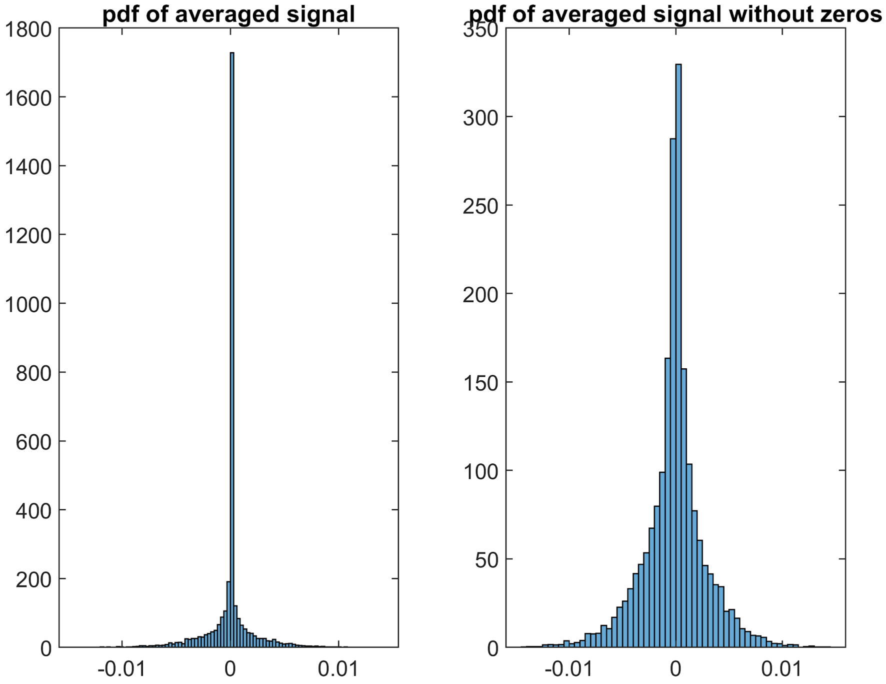

Figure. 13 pdf of averaged voice signal

3. Obtain the autocorrelation of the averaged signal.

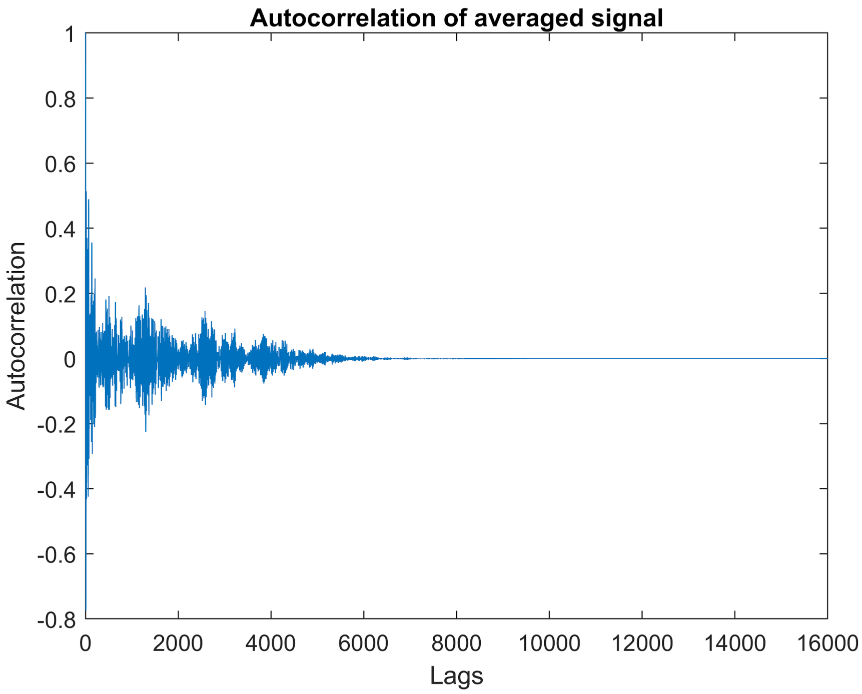

Figure. 14 Autocorrelation of averaged voice signal

4. Since the frequency domain representation of your voice signal (Fourier transform) is a complex signal, get the magnitude square (spectral density) of it and plot it.

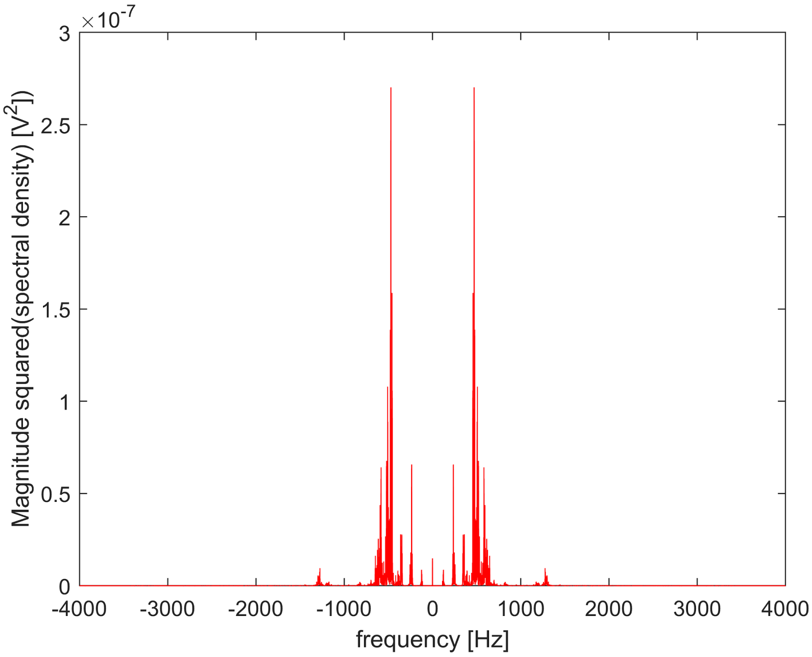

Figure. 15 Magnitude square (spectral density) of the averaged voice signal.

5. Apply Fourier transform on your autocorrelation and plot it.

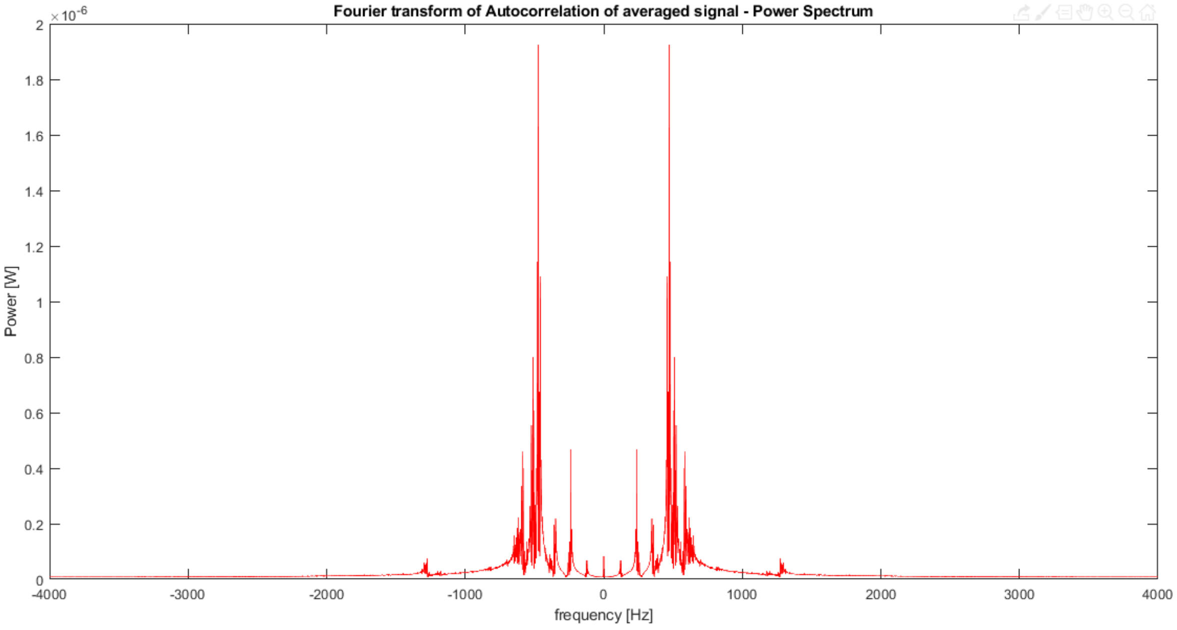

Figure. 16 Fourier transform of the autocorrelation of the averaged voice

Experiment 2 analysis:

The subject 0 is required to repeat the designated phrase a total of 50 or more times; particularly for our project the subject repeated the phrase a total of 80 times. Once all the samples were taken, the frequency and time domain of the average signal from all the samples were plotted. Finally a statistical analysis was performed for the amplitude of the averaged voice without the zero values in time domain, resulting in a histogram, a cdf and a pdf and a correlation plot, following the same procedure as in the so-called “Experiment 1” and conducting a Fourier transformation for the correlation plot.

The averaging of the whole samples had the disadvantages of losing valuable formation of samples as the signals were not always aligned, but by averaging all the samples, the resulting signal is more representative for characterization since it does not depend on a single sample and it is better to generalize the characterization of the voice, as long as the recorded samples are produced in the same time window and under the same circumstances.

By comparing the histograms from experiment 1 and 2, it is possible to see that the averaged voice has a smoother histogram than the one coming from experiment 1, although both of them have concentration of the data close to zero. The statistical parameters for this experiment have very similar results as in experiment 1. The variance, standard deviation and dispersion are much smaller than the values obtained in experiment 1 and this indicates that the average voice signal data is highly concentrated in some values and it is more representable for characterization of the voice than the signal used in experiment 1. The kurtosis parameter, with or without the zeros, shows a high value of 10.247 and 5.5864, which can be considered as being leptokurtic and it indicates that most values are close to the mean value.

The frequency content of the averaged voice seems to be smoother than the one from experiment 1, this case does not have several standing out frequencies in the center, instead of this, there is only one frequency on both sides center from cero which stands out than others. The information from the Fourier transform of the autocorrelation(power spectrum), and the magnitude square(spectral density) of the averaged signal has the same useful information in characterizing the voice of a person. As this experiment analyzes the averaged voice of 80 samples of the same phrase spoken, all these frequencies content has become more smoother than the ones obtained in experiment 1. By observing the resulting figures of frequency analysis in figure 10, 15 and 16, it is important to notice the high values of the frequencies components in the center, which could be very relevant to use as an indicator to differentiate the voice from one person to another.

Experiment 3 - Local Recording

1. Obtain samples from different test subjects reading the same text and make a comparison in the time and frequency domain with the first recorded signal and then with the averaged signal.

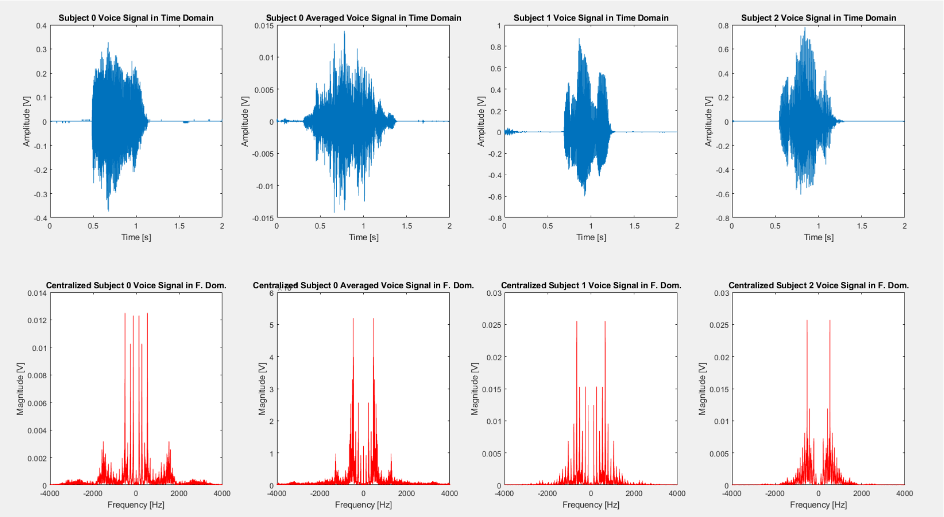

Figure 17. Time and frequency domain from subject 0, subject 1 and subject 2 voice local recordings.

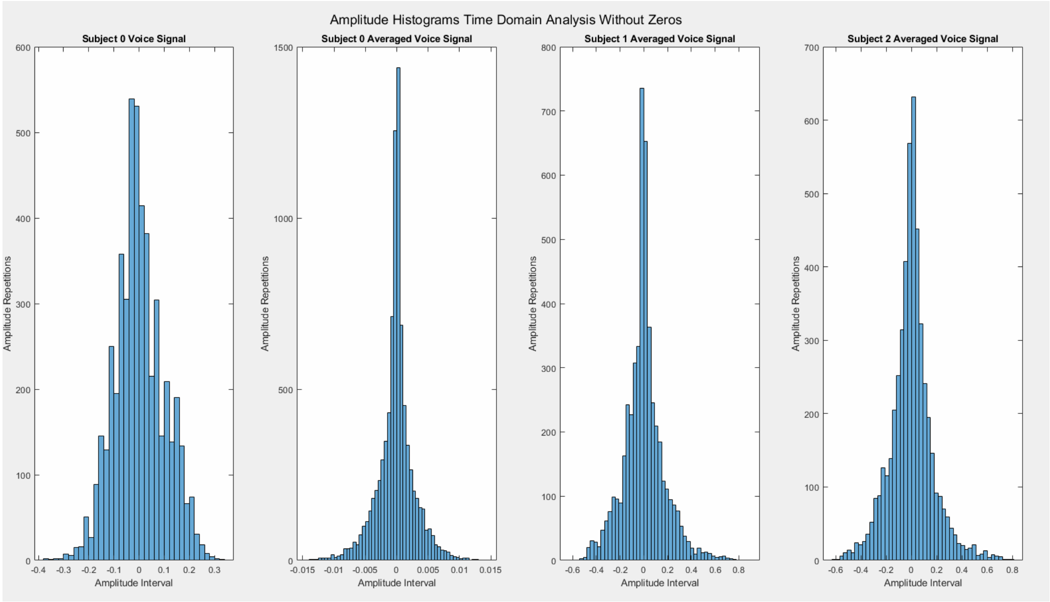

Figure 18. Histograms from subject 0, subject 1 and subject 2 local voice recordings.

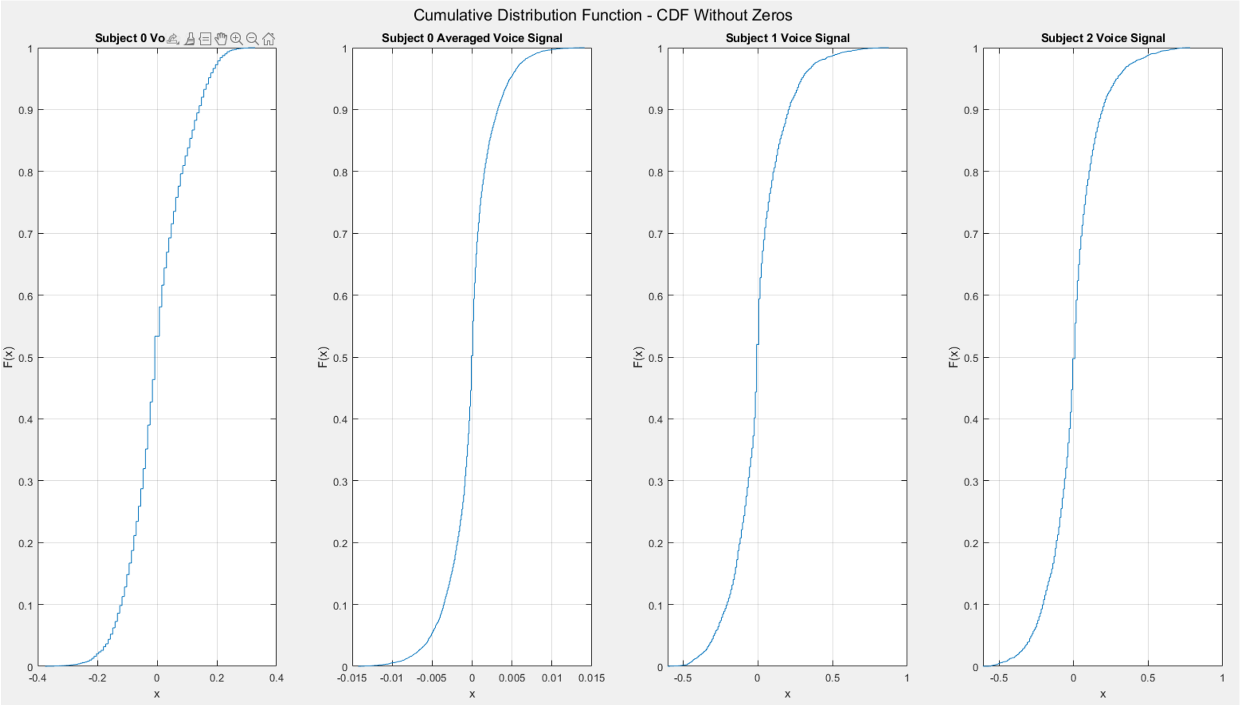

Figure 19. CDF from subject 0, subject 1 and subject 2 local voice recordings.

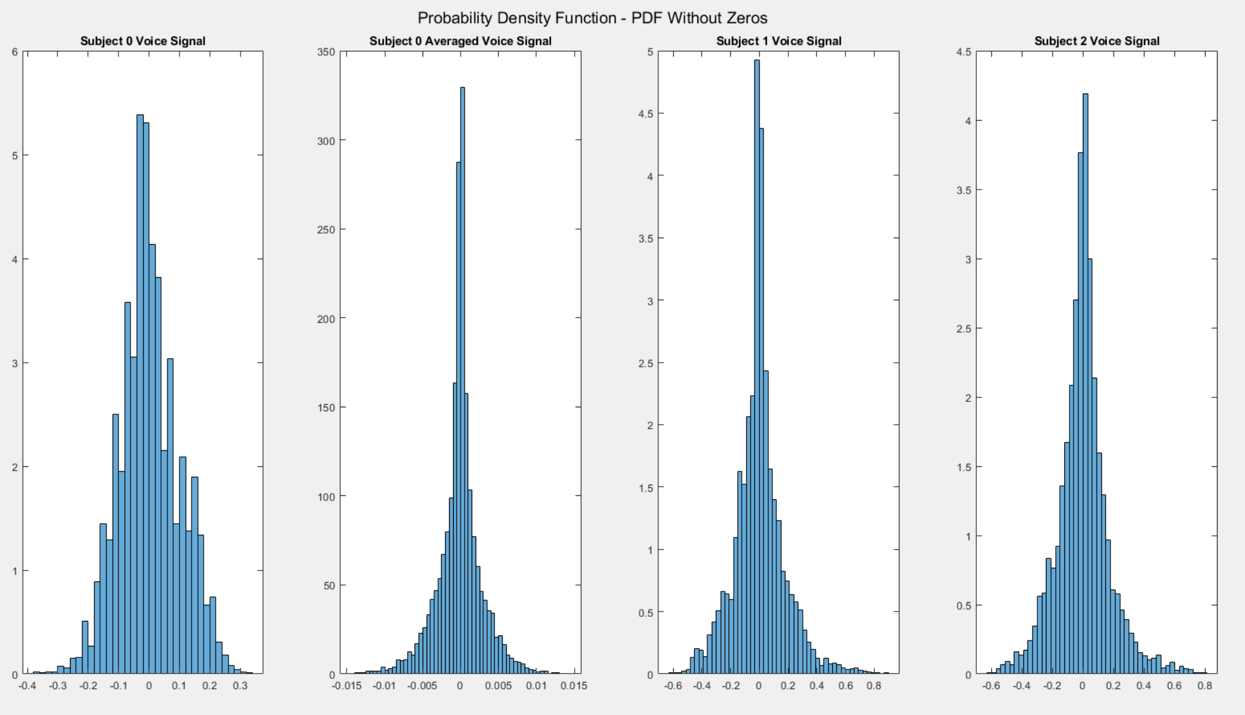

Figure 20. PDF from subject 0, subject 1 and subject 2 local voice recordings.

Table 3. Mean, variance, standard deviation, skewness, kurtosis, dispersion ($\frac{var}{std}$) of subject 0, subject 1 and subject 2 local voice recording.

Statistical Parameter	**Subject 0 Voice Signal with zeros**	**Subject 0 Voice Signal without zeros**	**Subject 0 Averaged Voice Signal with zeros**	**Subject 0 Averaged Voice Signal without zeros**	**Subject 1 Voice Signal with zeros**	**Subject 1 Voice Signal without zeros**	**Subject 2 Voice Signal with zeros**	**Subject 2 Voice Signal without zeros**
Mean	0.00046191	0.0014766	-0.0000607	-0.000111	0.00005468	0.00017591	-0 00002832	-0.000090
Variance	0.003043	0.0097277	0.000004626	0.0000084	0.0093039	0.029932	0.0091318	0.029074
Standard Deviation	0.055163	0.098629	0.0021509	0.0029109	0.096457	0.17301	0.09556	0.17051
Skewness	0.26363	0.11661	-0.27912	-0.15426	0.8377	0.46497	0.63401	0.35643
Kurtosis	9.6869	3.0257	10.247	5.5864	16.213	5.0388	16.085	5.0532
Dispersion	0.055163	0.098629	0.0021509	0.0029109	0.096457	0.17301	0.09556	0.17051

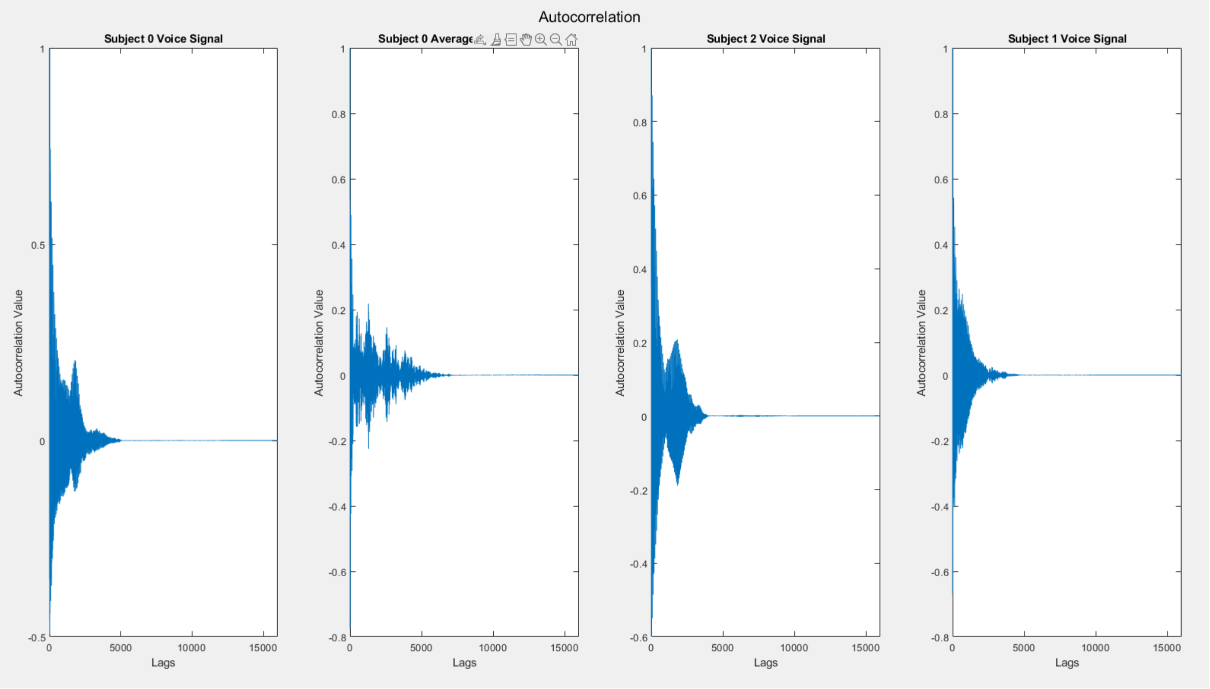

Figure 21. Autocorrelation from subject 0, subject 1 and subject 2 local voice recordings.

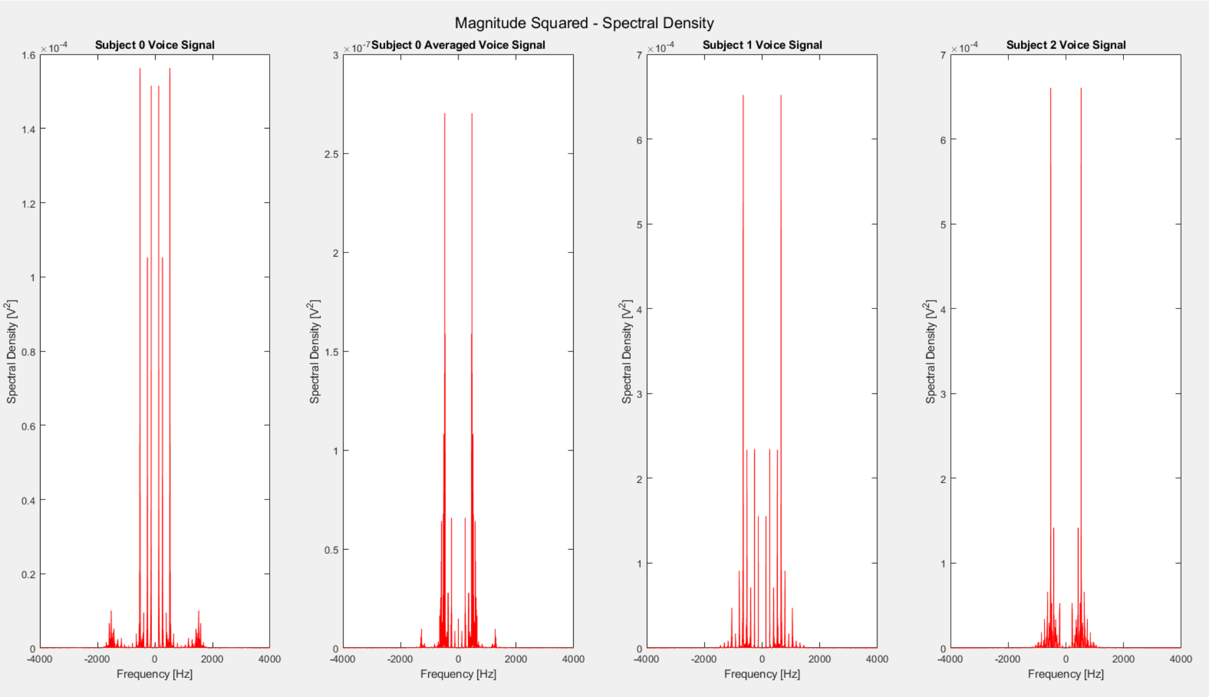

Figure 22. Spectral Density from subject 0, subject 1 and subject 2 local voice recordings.

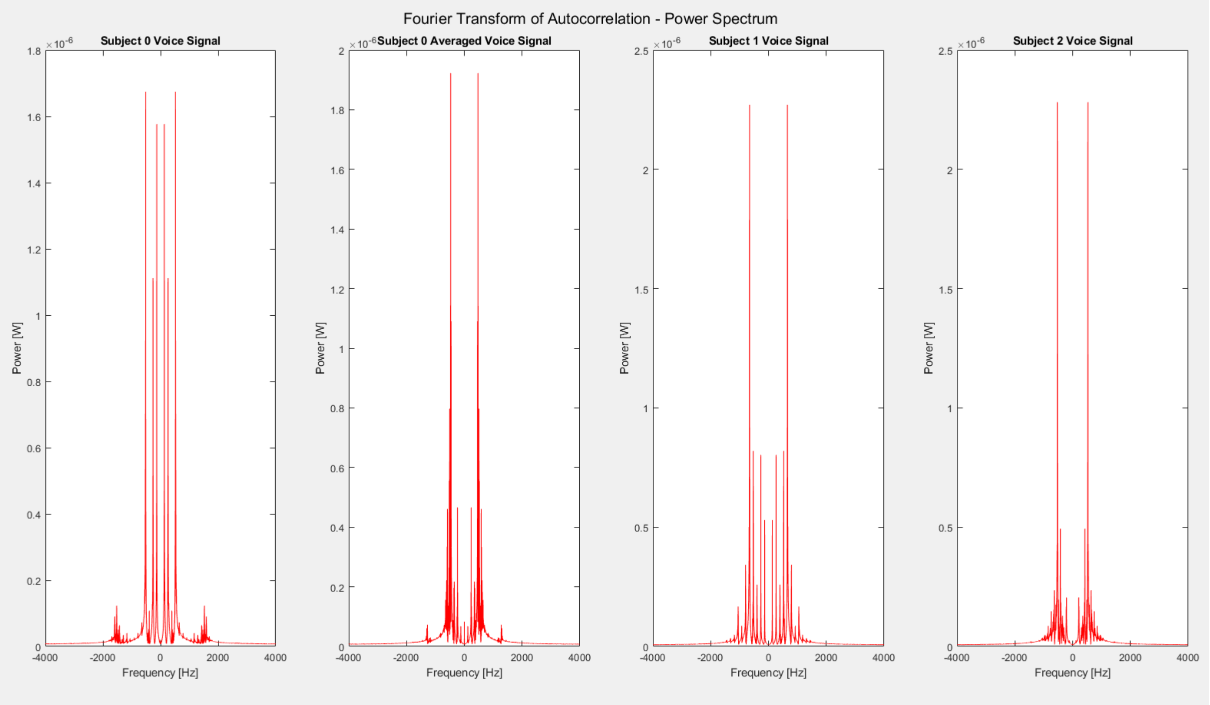

Figure 23. Power spectrum from subject 0, subject 1 and subject 2 local voice recordings.

Experiment 3 - Local recording analysis

Amplitude with respect to time of Figure 17 is a photograph of the recording, in terms of characterization it does not provide sufficient information to discard a voice signal from another. In that matter frequency provides substantial information about the voice subject, the spikes are the most important data points because those characterize a voice and can compare with future subjects to determine whether it is the subject of interest. It is no casualty that Figure 17 along with Figure 21 and Figure 22 resemble the same information about the subjects voice, but the last two mentioned figures are a kind of average computed to smooth the non-substantial spikes of larger frequencies (keeping in mind the fact that voice range goes approximately from 300 to 3600 kHz. Having more number of spikes facilitate comparison between test subjects' voices, as an example subject 0 (one time recording) has plenty of differentiation from subject 1 (one time recording). Subject 1 averaged signal can be improve in how it was recorded, some amplitudes are attenuated when average function is computed, losing information, this can be corrected by selecting a time window for all repetitions and perform correlation process to determine when the signals have the higher correlation factor and thus, it can provide a better overall characterization of a subject voice signal. The pre-processing of the signal is a crucial section for all experiments of this kind, one recording it is not reliable to characterize a signal.

The statistical parameters such as the variance, standard deviation, kurtosis and dispersion yield to very similar results, all of them are close to zero and this indicates, as it has been analyzed in experiment 1 and 2, that the data is very close to the mean value. Although these parameters of the average voice signal of subject 0 tend to be smaller than the others because of the smoothing effect of averaging, there is not much information to differentiate one voice from another using these statistical values.

Experiment 3 - Remote Recording

- Consider a remote voice recording. i.e., using phone or zoom. Show a time and a frequency domain plots together with the statistical characterization.

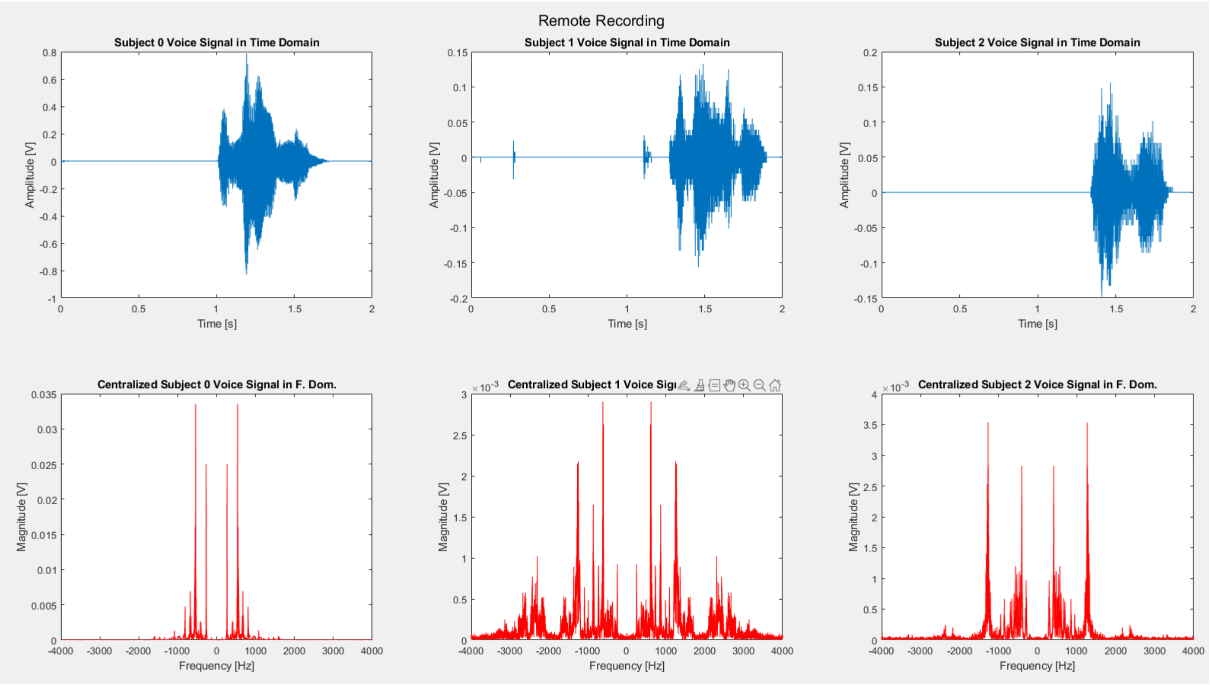

Figure 24. Time and frequency domain from subject 0, subject 1 and subject 2 voice remote recordings.

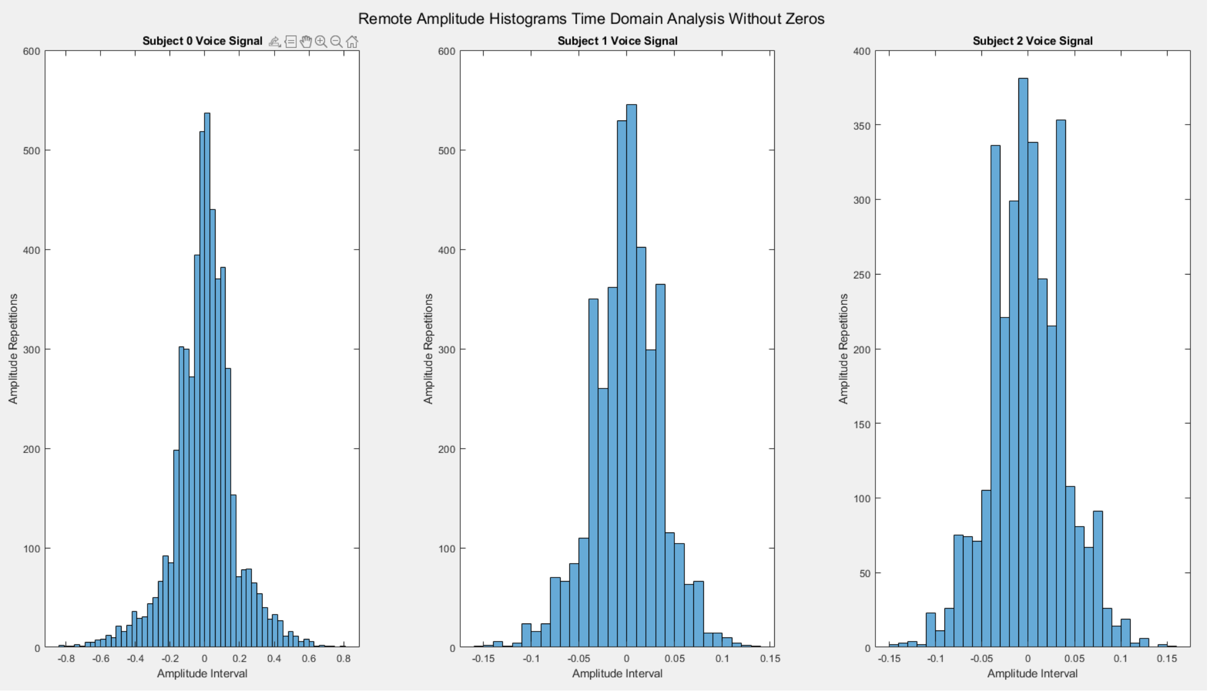

Figure 25. Histograms from subject 0, subject 1 and subject 2 remote voice recordings.

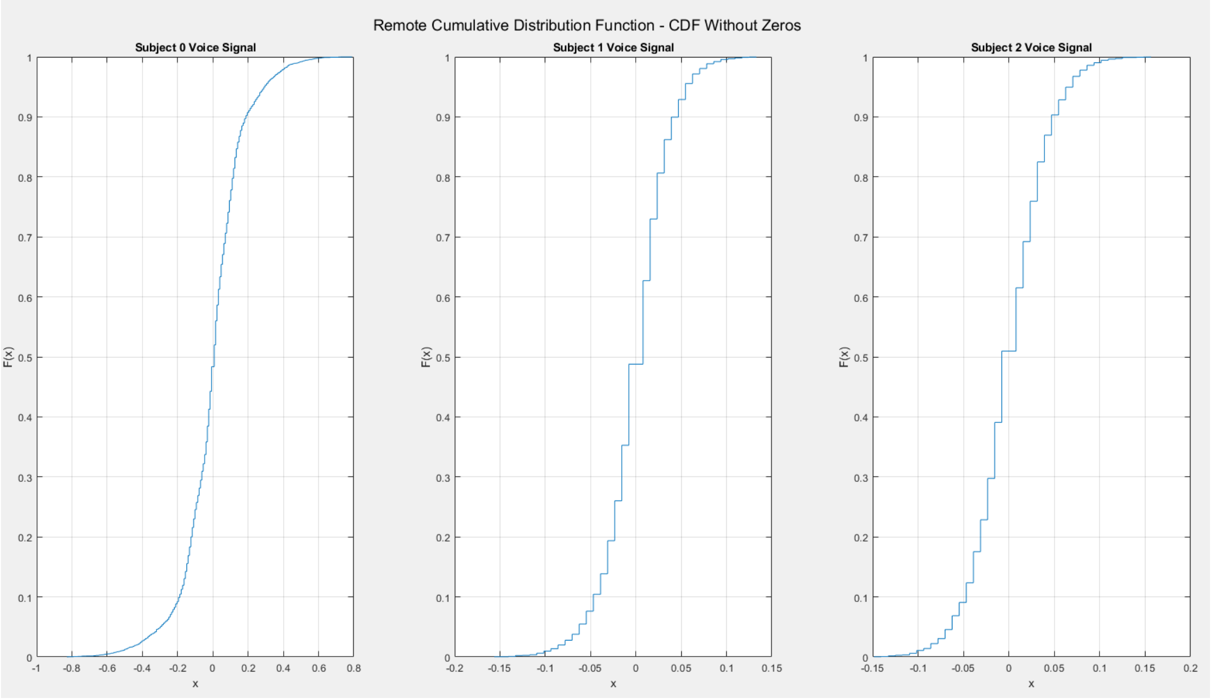

Figure 26. CDF from subject 0, subject 1 and subject 2 remote voice recordings.

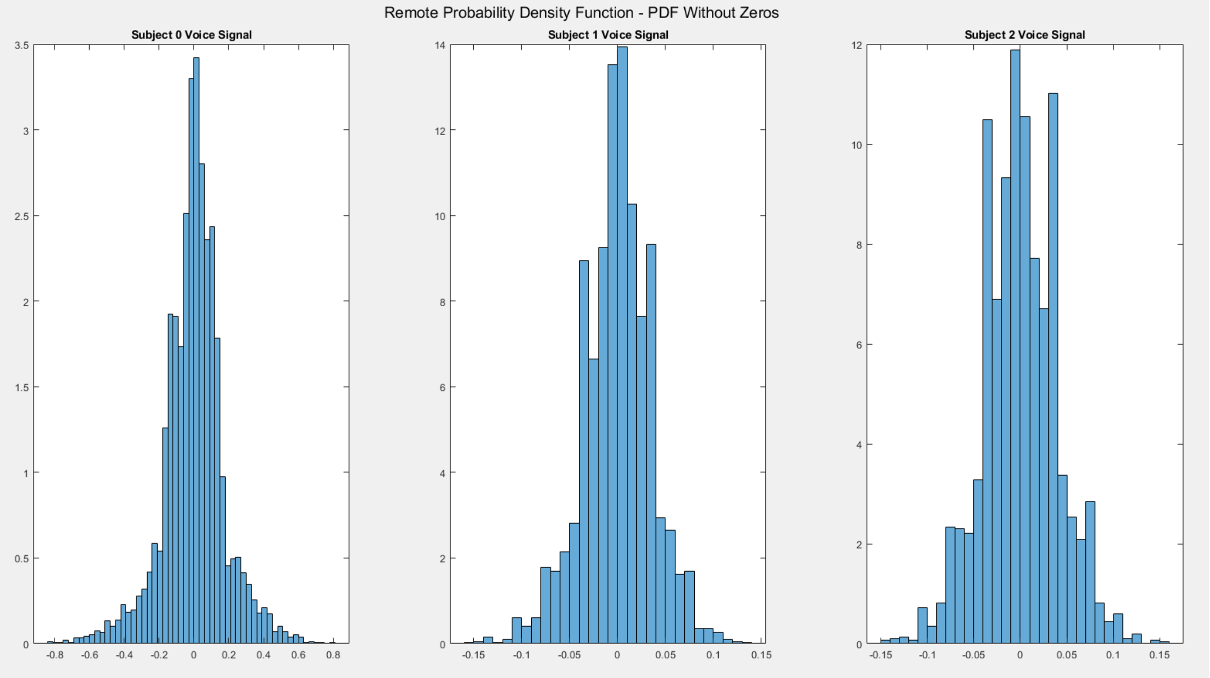

Figure 27. PDF from subject 0, subject 1 and subject 2 remote voice recordings.

Table 4. Mean, variance, standard deviation, skewness, kurtosis, dispersion ($\frac{var}{std}$) of subject 0, subject 1 and subject 2 remote voice recording.

Remote Recording						
Statistical Parameter	**Subject 0 Voice Signal with zeros**	**Subject 0 Voice Signal without zeros**	**Subject 1 Voice Signal with zeros**	**Subject 1 Voice Signal without zeros**	**Subject 2 Voice Signal with zeros**	**Subject 2 Voice Signal without zeros**
Mean	-0.000025879	0.00007914	-000024414	-0.000099827	0.000029297	0.0001463
Variance	0.010113	0.030929	0.0003113	0.0012731	0.00031682	0.0015825
Standard Deviation	0.10056	0.17587	0.017644	0.035681	0.017799	0.039781
Skewness	-0.353	-0.20095	-0.45039	-0.21639	0.063654	0.01966
Kurtosis	15.869	5.189	15.799	3.8619	16.833	3.3705
Dispersion	0.10056	0.17587	0.017644	0.035681	0.017799	0.039781

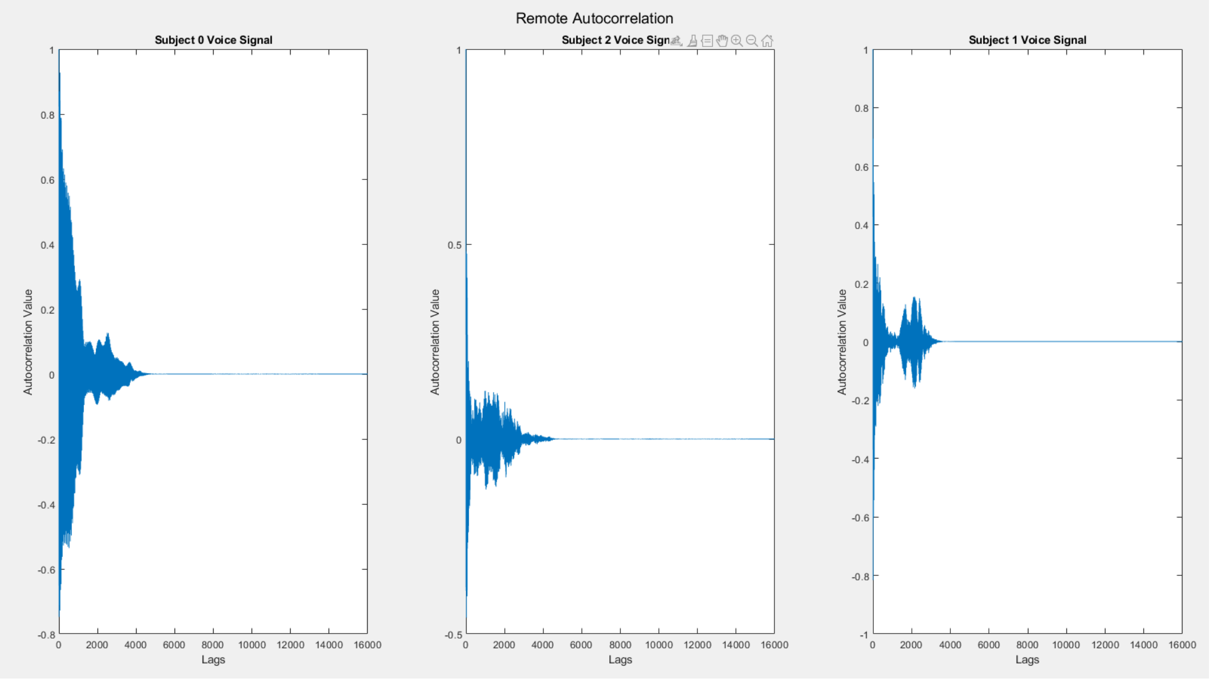

Figure 28. Autocorrelation from subject 0, subject 1 and subject 2 remote voice recordings.

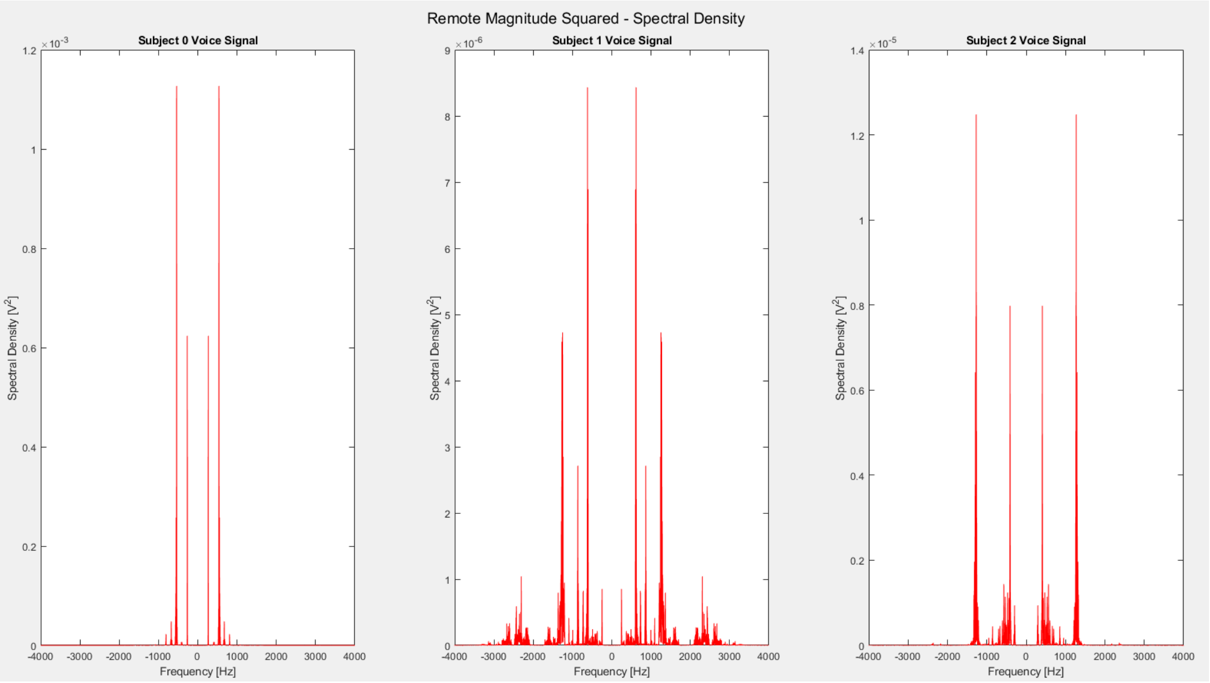

Figure 29. Spectral Density from subject 0, subject 1 and subject 2 remote voice recordings.

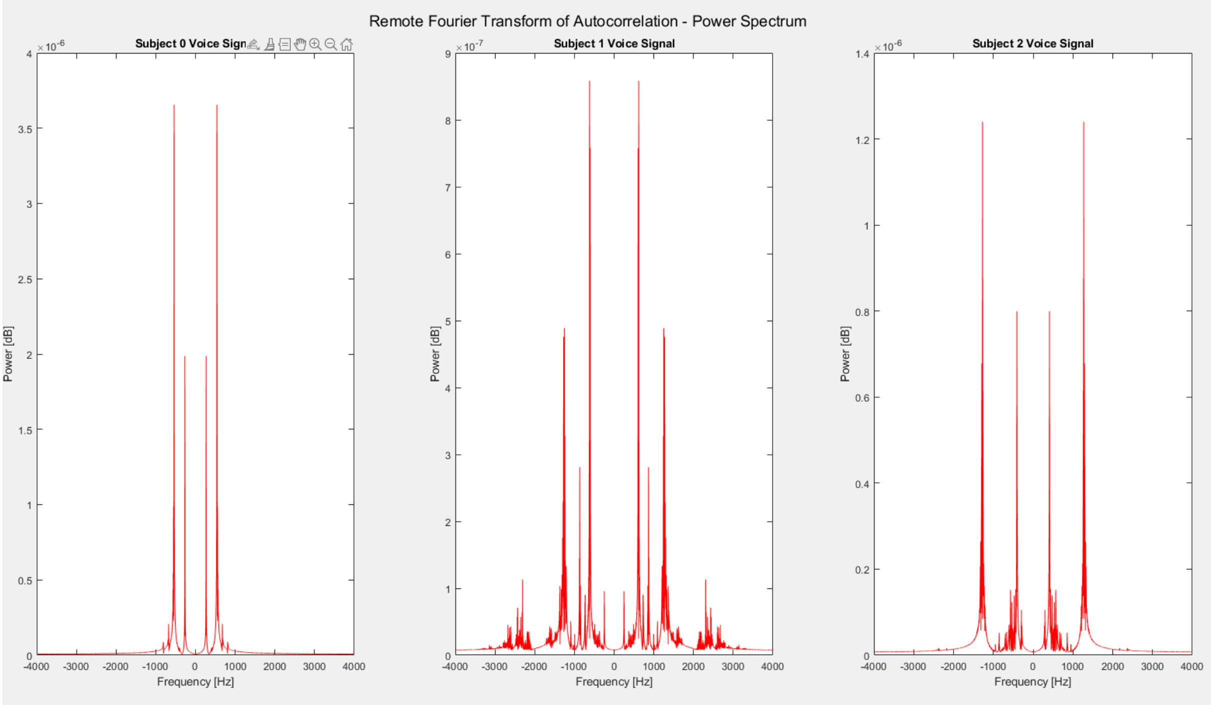

Figure 30. Power spectrum from subject 0, subject 1 and subject 2 remote voice recordings.

Experiment 3 - Remote recording analysis

Subjects 0, 1 and 2 presented particular characteristics between them at the remotely recording. First, the amplitude was different in each case as a consequence that each one speaks at a certain volume. Then, the most important, is that the peaks distribution was so different, giving a more centric concentration for the bass and baritone voice tones, and more peripheral distribution for the countertenor and tenor. Even though, voices when speaking are note pure in tonality. In example, the subject 1 has peaks all around the spectrum, with a predominance around the 1000Hz. This whole mixture of tonalities and amplitudes result in what is known as “timbre”, which opens the possibility of recognizing the voice subject origin.

When first recording and analyzing the power spectrum from subject 0 using autocorrelation (local recording), we can observe that there are significant peaks in frequencies from -1000 Hz to 1000 Hz, however when comparing it with the subjects 0 power graph (Remote recording); there are a significant differences between the amplitudes in the same range and peaks from frequencies, where the highers than $|1000|Hz$ and lowers than $|200|Hz$ were eliminated. Meaning that even though subject 0 used the same phrase for both analysis, the amplitude difference lies in the fact that the first test was made locally and the second one was made remotely using Zoom, which realizes a software filtering and an amplitude standardization before sending is done.

CONCLUSION

The realization of the experiment under different parameters gives enough arguments to define future improvements that will be necessary in order to have a more accurate characterization of the signals. An example of this was the differences between remote signals and local, where the Zoom filtering reduced the noise presence. A type of filtering was performed when the 80 samples were averaged, but the lack of alignment ended up attenuating information that was part of the desired signal; opening an improvement area were a cross-correlation would be helpful at the moment of aligning signals, at once aligned now its possible to cut out the areas without information.

In general, the objective planned was fulfilled as the voice inputs were correctly characterized by conducting a time and frequency domain analysis. The relevance of those analyses were clearify by comparing the data and confirming that different subjects have different qualities of power spectrum, dispersion, kurtosis and skewness; meaning that future subject identification based on their voice will be possible. Likewise, the investigation conducted for the introduction provided the team with enough information and data, regarding the origines, modern and future uses and algorithms used for voice recognition technology. All these allow us to have a comprehensible development of all three experiments and the results interpretation.

REFERENCES:

[1] Duque Sánchez, C., & Morales Pérez, M. (2007). Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones. Retrieved 25 March 2021, from https://core.ac.uk/download/pdf/71394315.pdf

[2] Van Der Velde, N. (2021). Speech Recognition Software: History, Present & Future. Retrieved 25 March 2021, from https://www.globalme.net/blog/speech-recognition-software-history-future/

[3] Van Der Velde, N. (2015). Innovative Uses of Speech Recognition Today: A Complete Guide. Retrieved 25 March 2021, from https://www.globalme.net/blog/new-technology-in-speech-recognition/

[4] Shah, H. N. M., Ab Rashid, M. Z., Abdollah, M. F., Kamarudin, M. N., Lin, C. K., & Kamis, Z. (2014). Biometric voice recognition in security systems. *Indian journal of Science and Technology*, *7*(2), 104.

[5] Lyden, C. What is Speech Recognition Software? How Does it Work?. Retrieved 25 March 2021, from https://www.callrail.com/blog/speech-recognition-software/

[6] Nautsch A. et al. (2019). *Preserving privacy in speaker and speech characterisation.* Retrieved 26 March 2021, from https://www.sciencedirect.com/science/article/pii/S0885230818303875#!

[7] Semmlow, J. (2012). *Signals and systems for bioengineers* (2nd ed., pp. 131-165). Waltham, MA: Elsevier/Academic Press.