Determining the Presence of COVID-19 Through the Processing and Analysis of Cough Samples.

Michael C. Moore (2118213)

Partner: Christopher Rawlings (2179595)

School of Electrical and Information Engineering, University of The Witwatersrand

ELEN3014A Biomedical Signals, Systems and Control

Prof. Vered Aharonson and Mr Abdul-Khaaliq Mohamed

October 30th, 2021

Abstract: The purpose of this project is to analyze cough samples obtained from patients who have tested positive and negative for COVID-19 and thus quantitatively differentiate the two using six different signal characteristics. The pitch, duration, average power, maximum power, number of peaks, and the most dominant frequency are the properties chosen to characterize the signals and thus detect the presence of COVID-19.

1. Introduction

After being declared a global pandemic on the 11th of March 2020 by the World Health Organization, COVID-19 has negatively impacted individuals, families, countries, and the world. Due to the devastation caused by this outbreak, professionals in the health industry have been working tirelessly to minimize the rapid spread through the production of vaccines and testing methods, such as lateral flow and PCR tests (1). These tests allow any individual to determine if they have contracted the virus and can thus self-isolate. These methods are useful and efficient; however, interaction with others is required to obtain the results, either through direct testing by a professional or through the purchasing of the kits (1). This interaction could lead to the further spread of the virus and thus a completely autonomous and isolated method of testing would prove beneficial.

Patients infected with COVID-19 present with various characteristics indicative of the virus, such as fatigue and low oxygen saturation levels; however, a change in the sound of the individual's cough is a discrepancy that could diagnose a patient (2).

The use of a cough audio signal to detect COVID-19 has multiple benefits, including the ability to perform the test at home using a smartphone or laptop and thus minimizing the spread of the virus and increased testing due to the increased accessibility, which in turn would result in improved COVID-19 tracking and safety. A potential drawback to using an audio signal to identify the virus is noise interfering with the signal data, and thus, a method of denoising must be employed.

This project will investigate the differences between the coughs of individuals who test positive and negative for COVID-19, and a comparative analysis will be employed to determine if the features of the signals investigated are successful at indicating the presence of the virus within an individual.

Within this project, the cough audio samples of various patients will be displayed in both the time and frequency domains to obtain a visual understanding of the differences between the coughs, and thereafter, six features will be selected and extracted from the sampled signals. These six features will be investigated, and a conclusion will be drawn as to whether there is a correlation between them and the presence of COVID-19.

2. Time-Domain Representation of Selected Signals

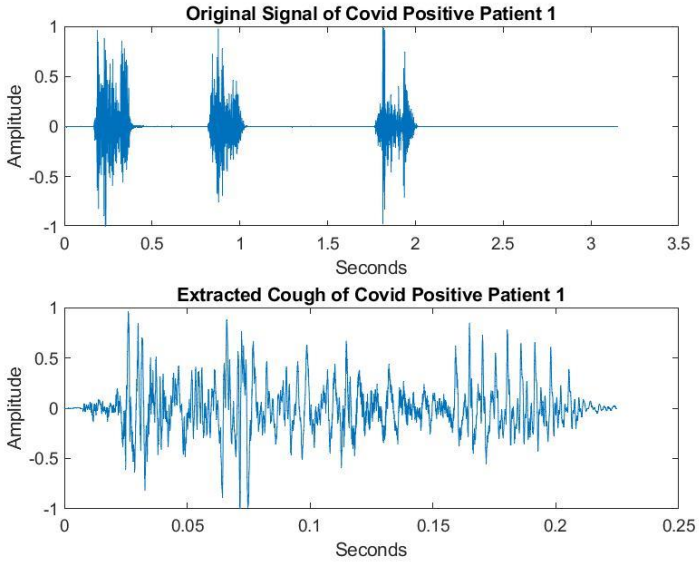

Figure 1: Original and cropped signal obtained from COVID-positive patient 1.

The first graph shown in *Figure 1* is that of a cough sample from a COVID-positive patient, with the extracted signal being that of the first cough within this audio file. The extracted cough has a duration of 0.22s and has many high peaks across the entire length of the signal. The highest peak within this sample has a magnitude of 1 and can be seen at the time of 0.07s, representing the loudest segment of this cough. The extracted graph has a relatively even distribution, and thus, it can be concluded that the loudness remained constant throughout its duration. 51.94% of the peaks within the extracted signal have magnitudes higher than 10% of the maximum peak, and this characteristic will be used as a possible indication of COVID-19. The characteristics presented can be further seen in *Figures 5 & 6* in Appendix A.

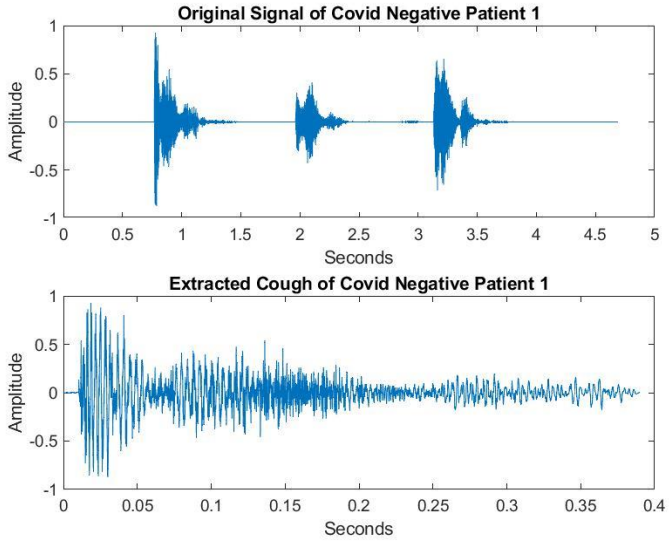

Figure 2: Original and cropped cough sample obtained from COVID-negative patient 1.

The first signal presented in *Figure 2* is the cough sample of a patient who tested negative for COVID-19, with the extracted signal being the first cough within this patient's audio file. The extracted signal has a duration of 0.39s, and most of its higher peaks are clustered towards the beginning of the sample. The highest peak has a value of 0.93; however, the peaks subsequently decrease, causing the graph to taper and adopt a relatively uneven distribution; this contrasts with the sample obtained from the COVID-positive patient in which the peaks are evenly distributed. In the cough sample, only 32,6% of the peaks have magnitudes higher than 10% of 0.93, which is considerably lower than in the COVID-positive sample, indicating a possible method of differentiation. The trends noted within this sample extend further in *Figures 7 & 8* in Appendix A.

3. Frequency-Domain Representation of Selected Signals

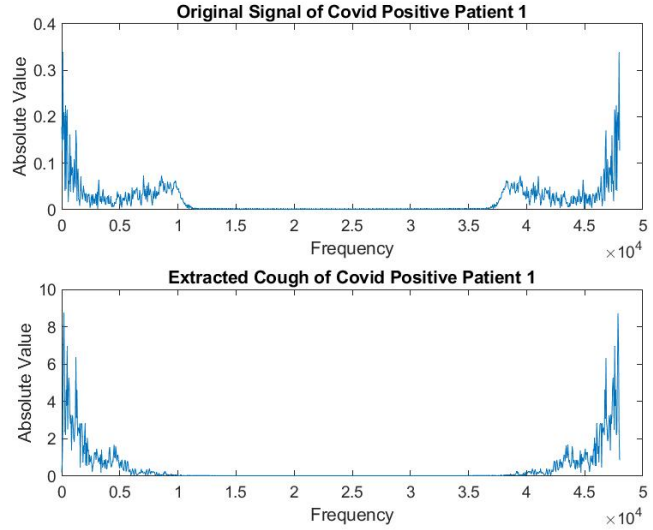

Figure 3: Frequency domain representation of original and cropped signal obtained from COVID-positive patient 1.

In *Figure 3*, the spectral representation of a COVID positive patient's audio file and extracted cough are presented. The dominant frequencies of the extracted cough are 188Hz and 47900Hz, and thus these frequencies have the peaks with the highest magnitude. There is an interval from 10000Hz – 38000Hz in which no peaks are present, and thus indicates that the extracted cough does not possess these frequencies. The pitch of the first cough is 158.75Hz, and thus this characteristic could serve as an indication of the presence of the virus. Further spectral representations of COVID positive patients can be found in *Figures 9 & 10* in Appendix B.

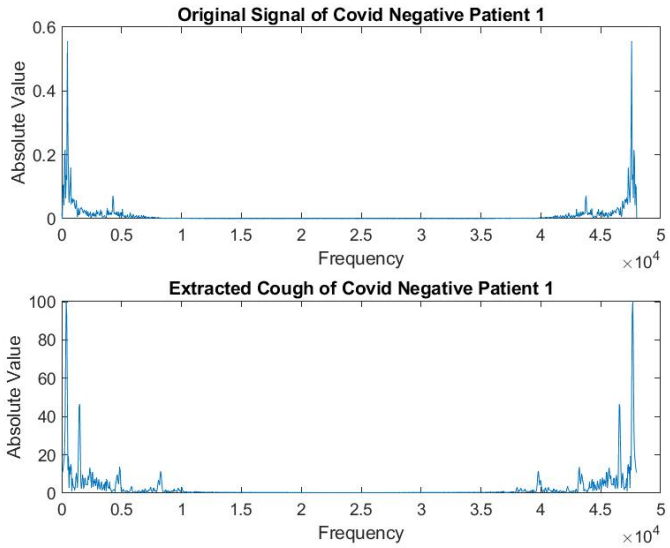

Figure 4: Frequency domain representation of original and cropped signal obtained from COVID negative patient 1.

The graphs presented in *Figure 4* consist of the spectral representations of the entire audio file and an extracted cough of a COVID-negative patient. The spectral presentation of the extracted cough consists of frequencies within the intervals 0 – 10000Hz and 38000Hz – 48000Hz, with the most dominant frequencies being 375Hz and 47700Hz, respectively. The lower dominant frequency differs between COVID-positive and negative individuals and thus serves as one of the characteristics used to distinguish the two groups. The highest peaks reach a magnitude of 100, which is greater than the samples obtained from COVID-positive individuals, and thus indicate that the frequencies correlating to these high peaks are more prevalent in COVID-negative individuals. The pitch of the extracted cough is 229.09Hz which contrasts with the lower value obtained from the positive sample. Further spectral representations of COVID negative patients can be found in *Figures 11 & 12* in Appendix B.

4. Signal Properties

- **Cough Duration**: The time over which a single cough occurs differs between individuals who test positive and negative for COVID and thus can be used to differentiate the two. When cropping the signals, it is assumed that prolonged regions without any peaks represent the breaks between coughs.
- **Maximum power of the signal**: The frequency that contains the highest power within the signal differs between the two groups, and thus this property can be exploited to characterize the samples. The maximum power represents the loudest point within the cough; however, the relationship is not linear (3).
- **The average power of the signal**: Each frequency within a sample possesses power, and thus taking the average of these values can be used to distinguish the two groups. There is a correlation between the average power and the average loudness of the cough sample (3).
- **The average pitch of the cough**: The pitch of a signal is its fundamental frequency, this value changes over the course of the sample, and thus an average will be taken, and the resulting value will be used for comparison (4).

- **The lower dominant frequency of the signal**: The cough samples are made up of various frequencies over a large range, and thus determining the frequency that occurs most often will allow for differentiation (5).
- **The number of peaks above 10% of the maximum peak**: The distribution and magnitude of the peaks differ between COVID positive and negative patients, and thus this property will allow for the differentiation between the two groups. This characteristic will provide information about the loudness of a signal over the course of the cough's duration.

5. Feature Tables

Table 1: Data collected from cough samples of COVID-negative patients.

Coughs COVID19 Negative	Average Power (W)	Lower Dominant frequency (Hz)	Duration (s)	Maximum Power (W)	Number of peaks above 10% (%)	Pitch (Hz)
Cough 1	0.0236	375	0.39	21.94	32.6	229.09
Cough 2	0.0079	1410	0.26	2.43	49.94	267.6
Cough 3	0.0274	1500	0.22	18.06	54.23	205.73
Cough 4	0.025	328	0.38	5.61	50.5	338.12
Cough 5	0.02	375	0.35	2.11	50.35	278.8
Cough 6	0.016	422	0.18	0.78	49.8	243.67
Cough 7	0.018	610	0.21	5.12	50.28	284.61
Cough 8	0.022	610	0.21	7.99	52.62	337.38
Cough 9	0.019	610	0.22	4.00	53.2	311.39
Cough 10	0.0069	235	0.28	3.62	54.34	194.16
Cough 11	0.048	422	0.4	15.21	62.8	147.55
Cough 12	0.0072	610	0.3	2.58	35.98	148.96
Cough 13	0.013	431	0.32	3.29	38.29	258.66
Cough 14	0.01	259	0.32	2.25	59.24	197.42
Cough 15	0.019	259	0.34	5.54	33.51	220.18
Cough 16	0.055	422	0.52	12.99	60.89	250.47
Cough 17	0.056	516	0.55	17.27	59.65	278.11
Cough 18	0.039	1880	0.31	11.85	59.44	245.70
Cough 19	0.041	563	0.39	12.23	42.84	288.2
Cough 20	0.046	610	0.23	9.67	23.94	217.15
Cough 21	0.057	657	0.19	13.6	26.64	274.04
Cough 22	0.08	563	0.23	12.37	58.57	229.99
Cough 23	0.027	469	0.3	6.54	55.59	293.63
Cough 24	0.0019	845	0.22	0.31	52.22	302.49
Cough 25	0.018	1640	0.23	1.38	27.08	265.29
Cough 26	0.017	1720	0.23	1.43	26.7	225.72
Cough 27	0.02	1770	0.21	1.69	23.59	252.85
Cough 28	0.04	235	0.36	18.66	60.1	183.12
Cough 29	0.034	282	0.3	14.23	51.01	190.69
Cough 30	0.029	375	0.39	14.98	53.03	141.76
Average	0.02813	698.43	0.301	8.324	46.98	243.42

Table 2: Data collected from cough samples of COVID-positive patients.

Coughs COVID19	Average Power (W)	Lower Dominant frequency (Hz)	Duration (S)	Maximum Power (W)	Number of peaks above 10% (%)	Pitch (Hz)
Cough 1	0.052559	187.68	0.22469	25.352	51.943	158.75
Cough 2	0.027852	1736.1	0.22094	6.3566	39.397	172.61
Cough 3	0.023864	1736.1	0.23896	3.5422	28.605	156.29
Cough 4	0.024609	517.3	0.2442	3.156	36.921	278.98
Cough 5	0.0075461	517.3	0.19698	1.1447	18.038	293.82
Cough 6	0.016016	431.09	0.30093	2.538	40.984	272.05
Cough 7	0.0040185	517.3	0.20488	1.8657	42.18	292.36
Cough 8	0.0028858	140.76	0.2151	1.2014	63.012	311.35
Cough 9	0.058572	656.89	0.34837	23.482	38.488	241.67
Cough 10	0.077262	1454.5	0.31281	29.643	39.56	229.28
Cough 11	0.040711	187.68	0.5469	24.071	32.566	221.3
Cough 12	0.057341	328.45	0.33687	23.653	37.458	164.86
Cough 13	0.071195	1548.4	0.16681	23.704	42.55	192.17
Cough 14	0.094042	1360.7	0.36896	16.128	47.493	225.82
Cough 15	0.00056737	1032.3	0.23721	0.037398	45.617	261.27
Cough 16	0.076003	656.89	0.33523	14.938	72.351	333.44
Cough 17	0.027159	609.97	0.37281	4.3185	50.4	258.24
Cough 18	0.010482	656.89	0.21215	0.91556	18.659	185.53
Cough 19	0.0015789	375.37	0.2064	0.20441	41.96	149.05
Cough 20	0.0416	328.45	0.87769	34.818	43.571	233.8
Cough 21	0.0046325	281.52	0.30981	5.9421	67.433	190.17
Average	0.0343	726.74	0.3085	11.7625	42.8184	229.6578

Graphs summarizing the information above are presented in Appendix C.

6. Comparative Analysis

Table 3: Accuracy of the program at determining the presence of COVID-19 based off individual characteristics as well as all properties combined.

Characteristic	Accuracy of COVID Positive Predictions (%)	Accuracy of COVID Negative Predictions (%)
Average Power	42.86	66.67
Lower Dominant Frequency	28.57	76.67
Duration	42.86	66.67
Max Power	42.86	60
% Peaks above 10%	66.67	63.33
Average Pitch	57.14	53.33
Combined	66.67	80

In this analysis, the six signal characteristics were tested separately to determine their effectiveness at differentiating between COVID positive and negative individuals, and subsequently, the properties were used together to calculate the overall accuracy of the program's

prediction. The accuracy of the program was determined using both the COVID positive and negative samples as these differed due to various factors.

The tool used to compare the signals was a program written in MATLAB that calculated the six characteristics of a specific cough sample and then compared those results to the averages of both the COVID positive and negative patients. The absolute difference between the value and the averages was determined, with a lower difference indicating a higher similarity. By comparing the values to the averages of both the positive and negative data, the program determined the diagnosis.

Once the program had made a prediction, the result was compared to the true diagnosis of the patient, and the number of correct predictions out of the total sample size was recorded. The recoded percentages represented the accuracy of the program and are presented in *Table 3*.

From the data presented in *Table 3*, the overall accuracy of the program differs between the COVID positive and negative samples, with the accuracy being higher for the latter. The program had an accuracy of 66.67% for the positive coughs due to the data obtained from these patients being like that of the negative samples, and thus the program produced an incorrect prediction. The program had an 80% accuracy when analysing the negative coughs; due to this data having more outliers, and thus the program could easily distinguish the sample.

The different properties varied in their ability to accurately predict the presence of COVID-19 within a patient, with a lower accuracy indicating a greater similarity between the positive and negative data. The lower dominant frequency only had an accuracy of 28.57% and thus was not a useful indication of the virus; whereas the percentage of peaks above 10% of the highest peak had an accuracy of 66.67% and thus indicates that this characteristic differed between both the positive and negative coughs, which could be seen in *Figures 1 & 2*. The average pitch accurately diagnosed COVID-19, 57.17% of the time, and thus proved to be a more useful characteristic. The maximum power, average power, and duration all had an accuracy of 42.86%, and thus are unreliable within this comparison when used by themselves.

7. Discussion

Within this project, six different signal properties were analyzed in both patients who tested positive and negative for COVID-19, which allowed for accuracy of 66.67% when predicting the presence of the virus within a patient. Certain signal characteristics, such as the percentage of peaks above 10% of the highest peak and the average pitch proved to be useful with accuracies of 66.67% and 57.14% respectively. Other properties such as the lower dominant frequency were ineffective at differentiating positive and negative samples and negatively impacted the overall accuracy of the program's diagnosis.

The relatively low accuracy of the diagnosis is due to various factors negatively influencing the recorded data. Noise was present in many of the cough recordings, and this altered the cough's waveform. A method of removing this noise was not implemented within this project due to the difficulty of differentiating between the frequencies of the true signal and the background interference. The presence of noise within the sample influences the power, frequency, and pitch values, and thus could serve as a reason for the inaccuracy of these characteristics.

The inaccuracies are also in part due to the characteristics chosen to differentiate the samples. Properties such as the duration and lowest dominant frequency are very similar between the positive and negative coughs, and thus the program could not effectively perform a diagnosis with the information provided. These properties resulted in the number of incorrect predictions to increase and decrease the overall accuracy of COVID-19 identification.

To improve upon this project; a method of filtering should be employed that removes all unwanted frequencies from the cough data before conducting an analysis. This would allow for the comparison of the true data, which would likely vary more between the positive and negative samples and thus improve the accuracy of COVID-19 diagnosis. The duration and lower dominant frequency should be replaced by characteristics that are more effective at differentiating the two sample groups, and the number of properties used in the comparison should be increased to allow for greater accuracy. The total number of samples should also be increased as more data would allow for more accurate averages to be calculated and thus improve the accuracy of the diagnosis.

8. Conclusion

After analyzing six different signal characteristics, an accuracy of 66.67% for COVID-19 diagnosis was achieved. It can be concluded that both the peaks and pitch of the signal served as effective indicators of the virus, whereas a combination of noise and ineffective properties, negatively impacted the accuracy of the diagnosis.

9. References

1. **Discovery.** What is the test used to diagnose COVID-19 infection? What are PCR and antibody tests all about? *Discovery*. [Online] 09 01, 2020. [Cited: 09 11, 2021.] https://www.discovery.co.za/corporate/covid19-test-used-to-diagnose-infection?.

2. *An ensemble learning approach to digital corona virus preliminary screening from cough sounds.* **Mohammed, Emad A, et al.** 15404, s.l. : Scientific Reports, 2021, Vol. 11.

3. **MathWorks.** Measure the Power of a Signal. *MathWorks Help Center*. [Online] 2021. [Cited: 10 3, 2021.] https://www.mathworks.com/help/signal/ug/measure-the-power-of-a-signal.html.

4. —. pitch. *MathWorks Help Center*. [Online] 2021. [Cited: 09 10, 2021.] https://www.mathworks.com/help/audio/ref/pitch.html.

5. —. Practical Introduction to Frequency-Domain Analysis. *MathWorks Help Center*. [Online] 2021. [Cited: 09 10, 2021.] https://www.mathworks.com/help/signal/ug/practical-introduction-to-frequency-domain-analysis.html.

6. **Abdullah, Saifuddin.** Signal Characteristics in Data Communication. *Techwalla.* [Online] 2021. [Cited: 09 05, 2021.] https://www.techwalla.com/articles/signal-characteristics-in-data-communication.

Appendix A

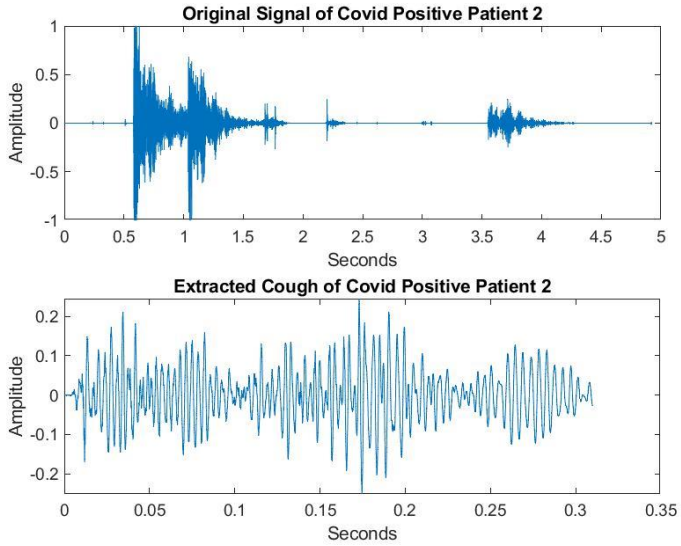

Figure 5: Original and cropped signal obtained from COVID-positive patient 2.

Patient 2's extracted cough presented in *Figure 5* is 0.31s in duration and has high peaks distributed evenly over the entire length of the signal. This sample differs from patient 1 in terms of the magnitude of the peaks, the highest peak observed within patient 2's cough is 0.25 in magnitude, while amplitudes as high as one are noted in patient 1's sample and thus suggest that patient 2 had a quieter cough. The extracted cough presented is the third waveform from the original audio file, which spans from 3.5s to 3.81s.

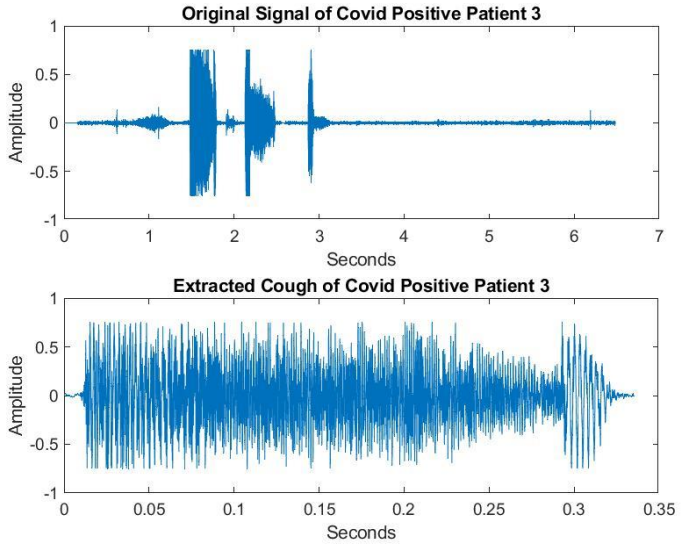

Figure 6: Original and cropped signal obtained from COVID-positive patient 3.

The cough sample presented in *Figure 6* has a duration of 0.34s and many peaks with large amplitudes distributed over the entire length of the signal. This signal differs from the previous COVID-positive patients in that it appears to be very noisy, and thus it is likely that the recording device registered both the cough and background noise. The largest peaks recorded in this sample have a magnitude of 0.75; however, due to numerous peaks presenting with this exact value, it is most likely due to a constant noise present in the recording.

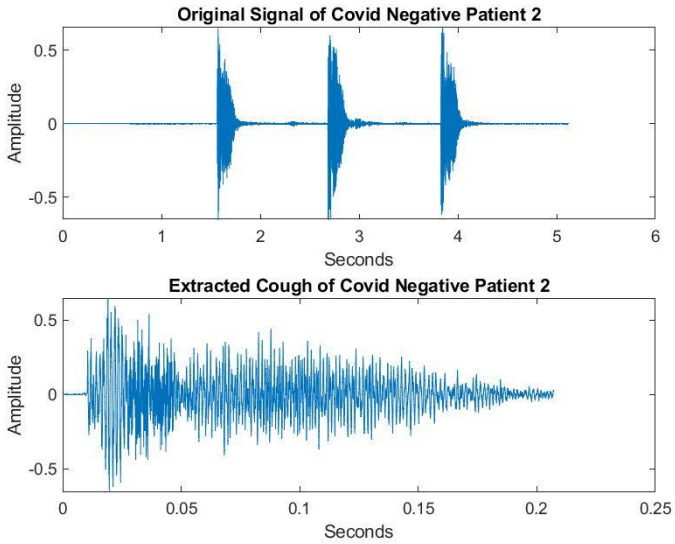

Figure 7: Original and cropped signal obtained from COVID-negative patient 2.

The sample presented in *Figure 7* has a similar distribution to that of *Figure 2* with the larger peaks being clustered at the start of the cough and then proceeding to taper towards the end. The highest peak present within this sample is 0.66 in magnitude, making it smaller than the previous COVID negative patient. Another aspect in which this signal deviates from the previous cough is in terms of its duration only being 0.21s, making it relatively short.

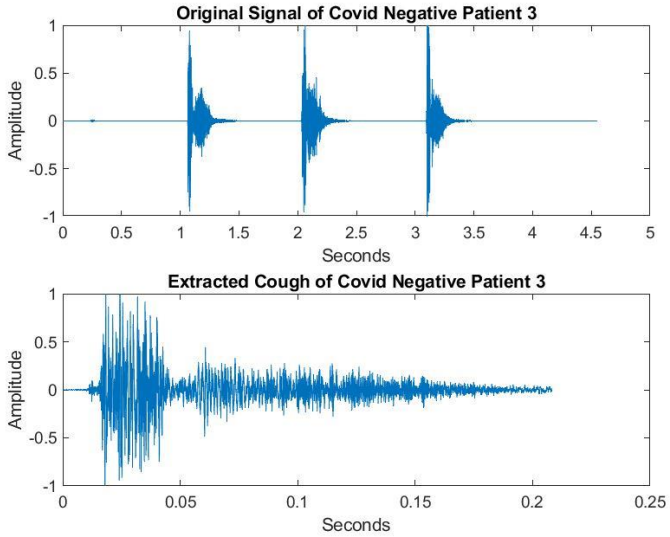

Figure 8: Original and cropped signal obtained from COVID-negative patient 3.

The sample displayed in *Figure 8* continues with the trend amongst the COVID negative patients of the larger peaks being found towards the start of the signal, with a maximum value of 1. The peak heights gradually decline towards the end, resulting in the signal having an uneven distribution and thus differentiates itself from the COVID positive patients. This cough sample has a duration of 0.21s and is thus shorter than the cough in *Figure 2*.

Appendix B

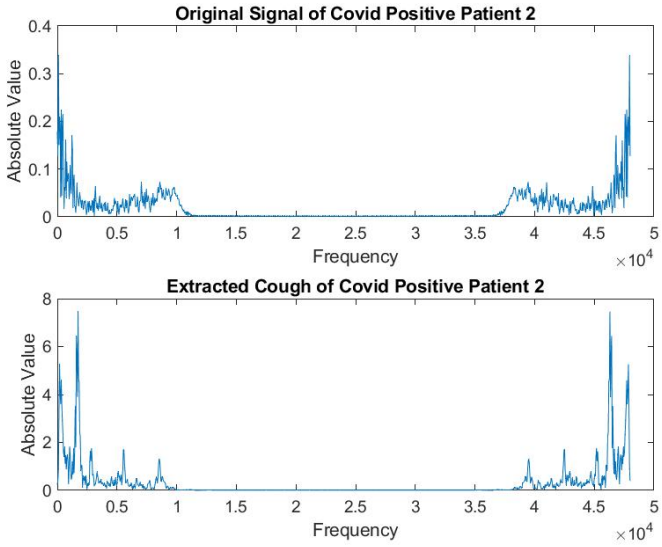

Figure 9: Frequency domain representation of original and cropped signal obtained from COVID-positive patient 2.

In *Figure 9*, the dominant frequencies are 1736Hz and 47800Hz; however, the peaks are higher in this sample and thus imply that these frequencies make up an even higher proportion of this signal as compared to the signal in *Figure 3*. Frequencies between 10000Hz and 38000Hz are not found within this signal, and thus the variety of frequencies present in this patient's cough is relatively low.

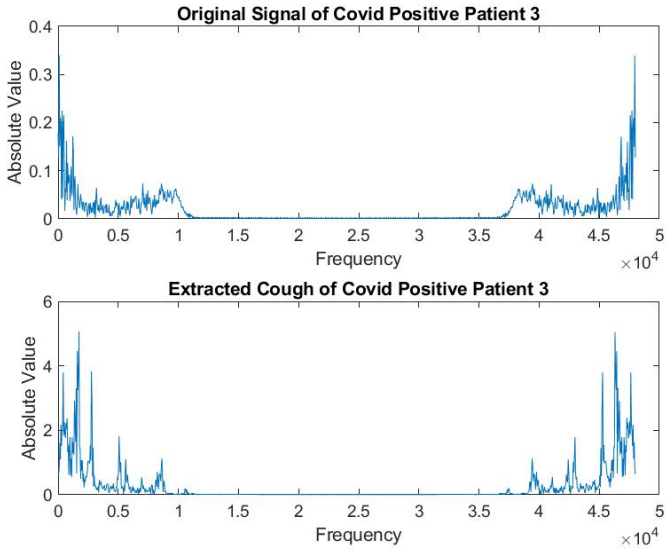

Figure 10: Frequency domain representation of original and cropped signal obtained from COVID-positive patient 3.

The sample seen in *Figure 10* consists of frequencies across the entire interval from 0Hz to 48000Hz; however, the frequencies, 1736Hz, and 47400Hz are more prevalent. The magnitude of the frequencies in this signal is larger than the previous COVID positive patients which indicates that there was more noise of the same frequency as the cough present in this sample.

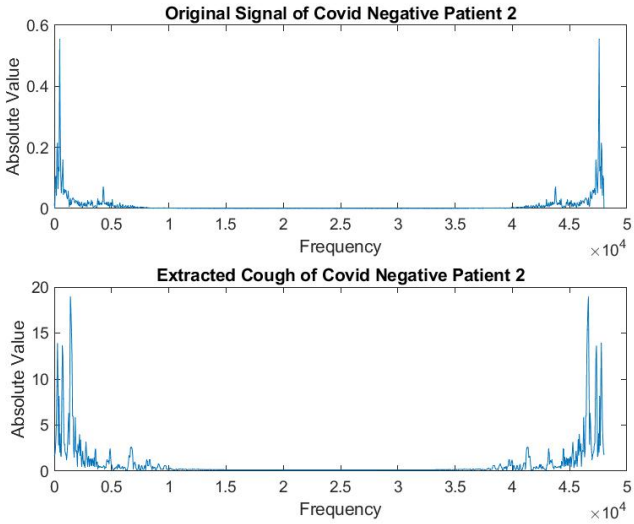

Figure 11: Frequency domain representation of original and cropped signal obtained from COVID-negative patient 2.

Within *Figure 11*, a similar distribution can be seen as compared to the other spectral densities. The frequencies between 10000Hz and 38000Hz are not present within this signal, whereas the frequencies 1500Hz and 47400Hz have a magnitude of 18 and thus constitute the dominant frequencies in the sample.

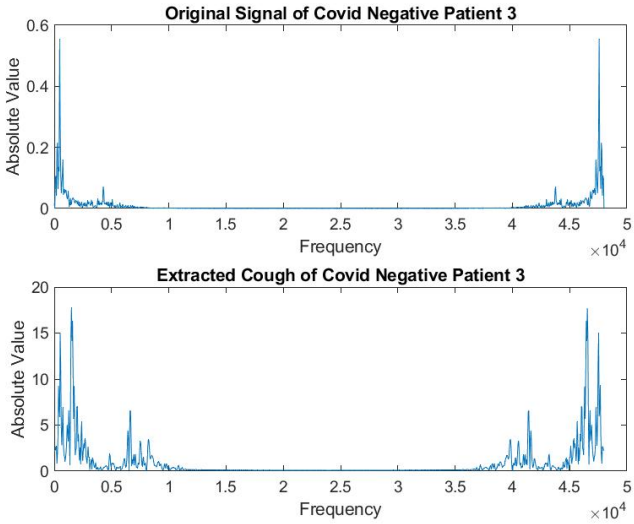

Figure 12: Frequency domain representation of original and cropped signal obtained from COVID-negative patient 3.

Patient 3's cough sample presented in *Figure 12* has its dominant frequencies of 1640Hz and 47000Hz, along with less dominant frequencies in the range of 5000Hz – 10000Hz, which could be a combination of noise and the patient's cough. The magnitude of the highest peaks in this sample appears to be lower as compared to the previous two COVID negative coughs, and thus indicates that less noise of that frequency was present during the recording.

Appendix C

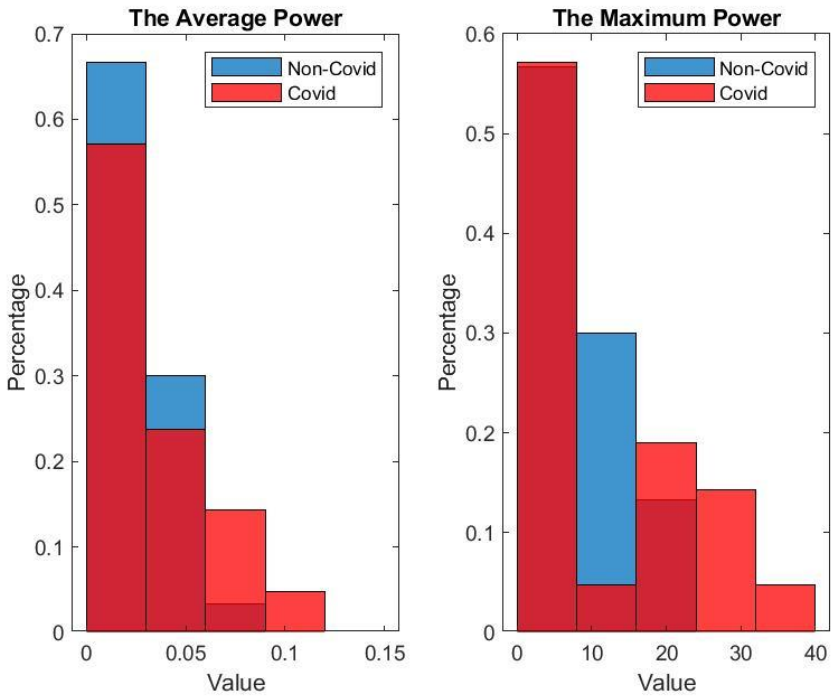

Figure 13: Graphs showing the percentage of COVID positive and negative samples that have different average and maximum powers.

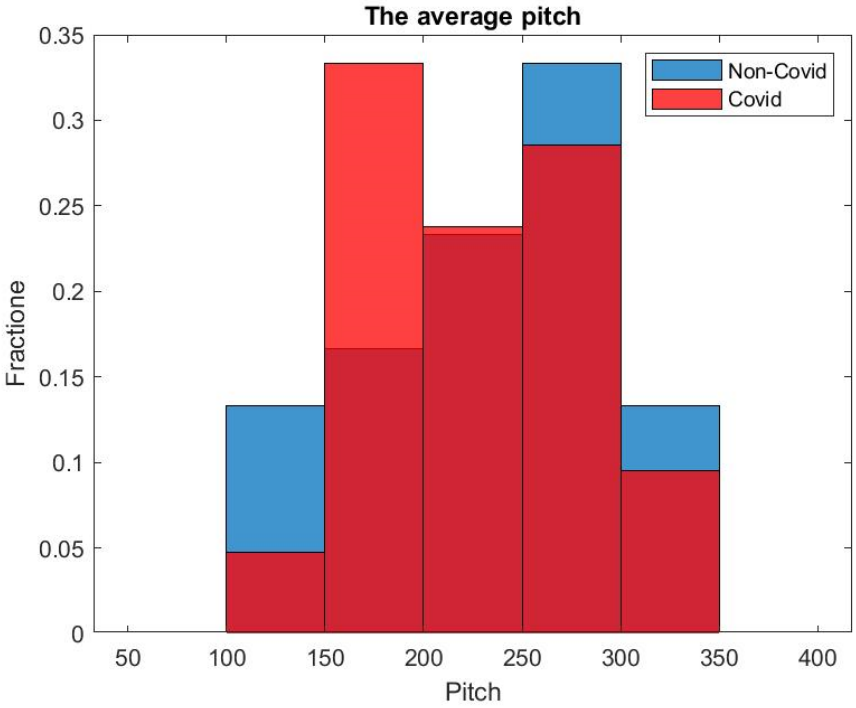

Figure 14: Graph showing the percentage of COVID positive and negative samples that have different average pitch values.

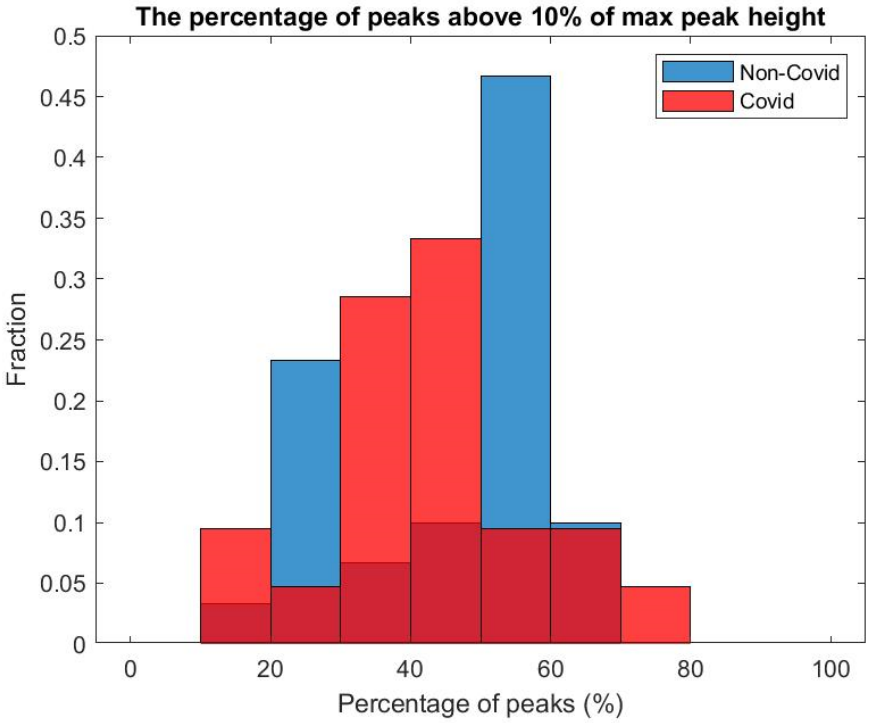

Figure 15: Graph showing the percentage of COVID positive and negative samples that have different peak values.

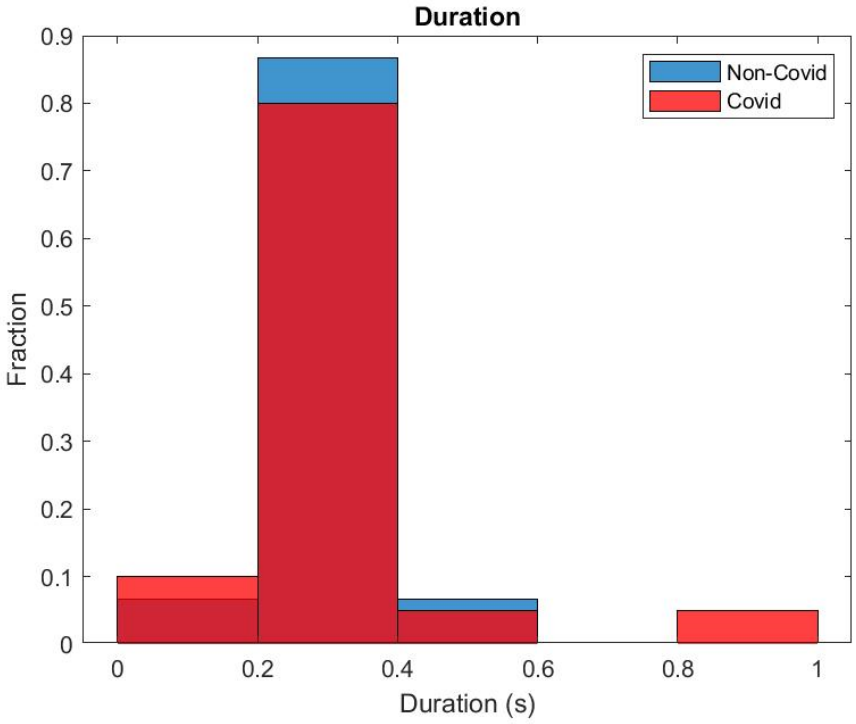

Figure 16: Graph showing the percentage of COVID positive and negative samples that have different duration values.