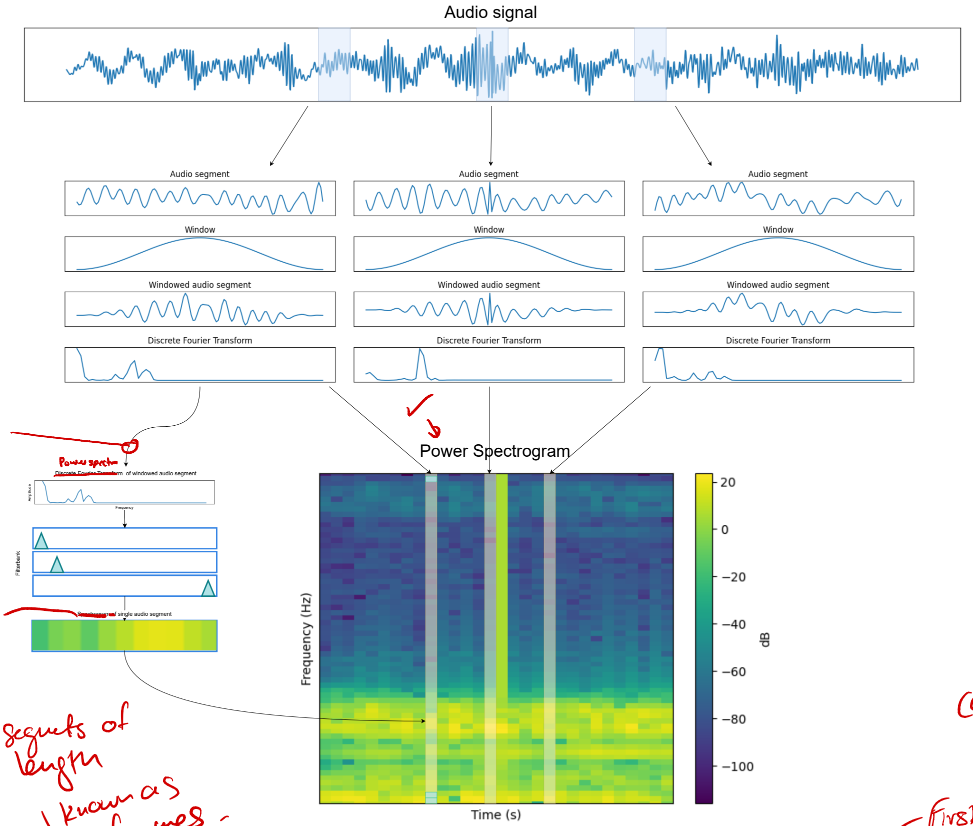
2.5. Acoustic features

The data for this project consists of audio cough recordings. These audio recordings are converted into ~~spectrogram images so that~~ a two-dimensional time-frequency representation to which image processing techniques can be applied. Figure 2.15 shows how an audio signal is converted into a spectrogram.

One such representation, known as



Over here the power spectrum is first calculated ie $\|X\|^2$

Filterbank energies for a

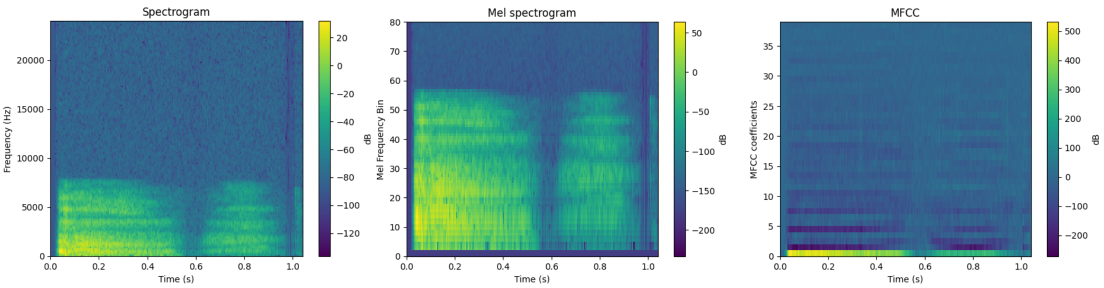
Short segments of equal length known as frames.

Figure 2.15: The process to convert a raw audio signal to a spectrogram. First, The audio signal is split into segments. Each ~~segment~~ such frame is multiplied by a window function, after which a Discrete Fourier Transform (DFT) is ~~applied to each of the windowed segments~~ computed. The DFT ~~is~~ coefficients are then normalised squared magnitude of the complex ~~multiplied with~~ passed through a filterbanks ~~and the result is~~. This is repeated for each frame and the results are stacked to form a spectrogram image.

The audio signal is divided into short, consecutive and sometimes overlapping audio segments often referred to as frames. A windowing function such as a Hamming window is then applied to each of the frames to reduce the spectral leakage caused by the truncation at the frame boundaries. A fast Fourier transform (FFT) is then applied to the windowed audio signal and the squared magnitude is taken to compute the energy with which each frequency is present in that audio signal. The resulting ~~magnitude~~ power spectrum is then multiplied by a set of bandpass filters constituting a filter bank to reduce the frequency resolution in a controlled manner. Finally, the resulting values are stacked as a series of vectors and used ~~as a~~ to display a contour plot with time on the x-axis, frequency on the y-axis and magnitude (in decibels) on the z-axis.

The contour plot is shown as a 2D image with the colours indicating the magnitude at the various frequencies for each frame as seen in Figure 2.16a.

Note this is not a contour plot, which would show lines of equal power. It is actually a "heat map"



(a) Spectrogram. **(b)** Mel-spectrogram. **(c)** MFCC

Figure 2.16: Spectrogram, mel-spectrogram and a MFCC of a cough audio sample.

A mel spectrogram is a spectrogram ~~where~~ in which the linear frequency scale ~~is~~ has been replaced by a mel frequency scale using the mel-scale ~~as~~ shown in Figure 2.16b. The mel-scale is a logarithmic frequency scale used to mimic how humans perceive sounds. Mel-scaled ~~spectrum~~ spectral energies can be computed from the magnitude spectrum by applying a mel-scaled instead of a linear-scaled filterbank.

and is D.

Mel-frequency cepstral coefficients (MFCCs) are calculated by applying a Discrete Cosine Transform (DCT) to the mel-scaled spectral energies as shown in Figure 2.16c. MFCCs capture uncorrelated information from audio signals that are useful to speech specific machine learning models.

2.6. Conclusion

This chapter discussed various theoretical concepts needed to understand contrastive learning and the terminology used throughout this project. The next chapter is a literature review of the existing work done in the field of TB classification and contrastive learning.