EE386 Digital Signal Processing Lab **Jul-Dec 2021**

Experiment 5

Author: Prasad Fidelis D'sa *Email:* prasadfidelisdsa.191mt032@nitk.edu.in

1 Introduction

The exercise is based on the time frequency analysis of audio signals using spectrograms. We begin by investigating the frequency components of a chirp signal using FFT and spectrogram and compare the results. We try different windowing techniques and window length to investigate spectral leakage and resolution. Pitch extraction from an instrument signal is then performed using spectrogram and the frequency variations in an opera signal are analyzed. Finally the spectro-temporal analysis of speech is done to map the phonemes to the spectrogram. In our example we have $\alpha = 1 + mod(260, 3) = 3$.
All the code for this exercise and the relevant files are included in the `.zip` file submitted along with this report.

2 Problems

1. Spectrogram of Chirp Signal
(Part 1: Generating Chirp Signal)
(Solution)

The chirp signal $x(t) = \sin(2\pi F(t)t)$ is generated by increasing $F(t)$ linearly from 8 Hz to 20 Hz at a sampling rate $F_s = 100$ samples/second for a duration $t = 10$ seconds. Figure 1 shows the plot of the generated signal with respect to time.

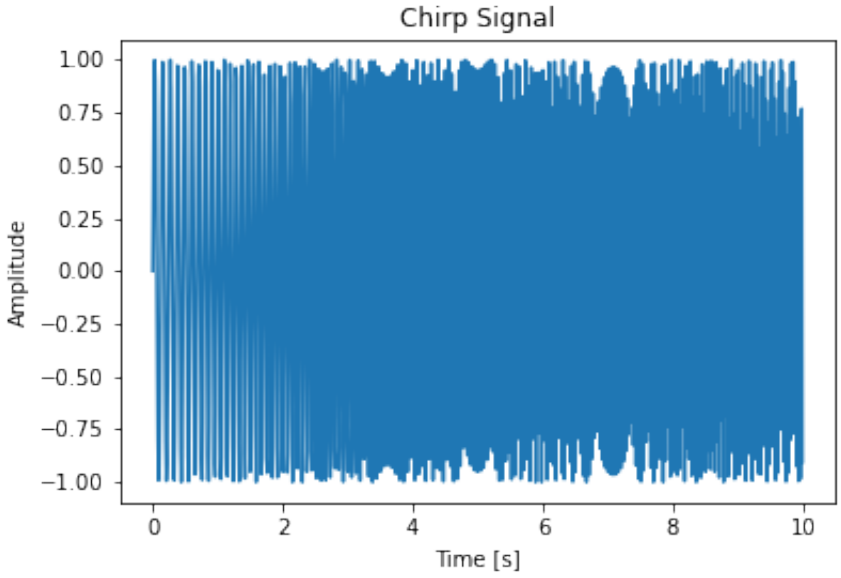

Figure 1: The time domain plot of Chirp Signal

The time-domain graph displays the changes in the signal over a span of time. It provides the transitory response of the system to be analyzed and it permits a better understanding of the flow of energies in wave propagation. When considered as an audio signal, $x(t)$ indicates the changes in air pressure from some ambient value as a function of time. However it is very difficult to analyze the frequency content from the waveform especially when the frequency is varying. For this we require a frequency domain representation and so we take to Fourier Analysis using FFT.

(Part 2: FFT of Chirp Signal)

(Solution)

The method of digitally computing Fourier spectra is widely referred to as the FFT (short for Fast Fourier transform). An FFT provides an extremely efficient method for computing approximations to Fourier series coefficients; these approximations are called DFTs (short for Discrete Fourier Transforms). The Discrete Fourier Transform (DFT) is used for the practical computation of the frequency content of real-world signals. The DFT transforms N samples of a discrete-time signal to the same number of discrete frequency samples, and is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}} \tag{2.1}$$

The DFT is invertible by the inverse discrete Fourier transform (IDFT):

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi nk}{N}} \tag{2.2}$$

The DFT and IDFT are a self-contained, one-to-one transform pair for a length-N discrete-time signal. Figure 2 shows the Magnitude Spectrum of the Chirp Signal.

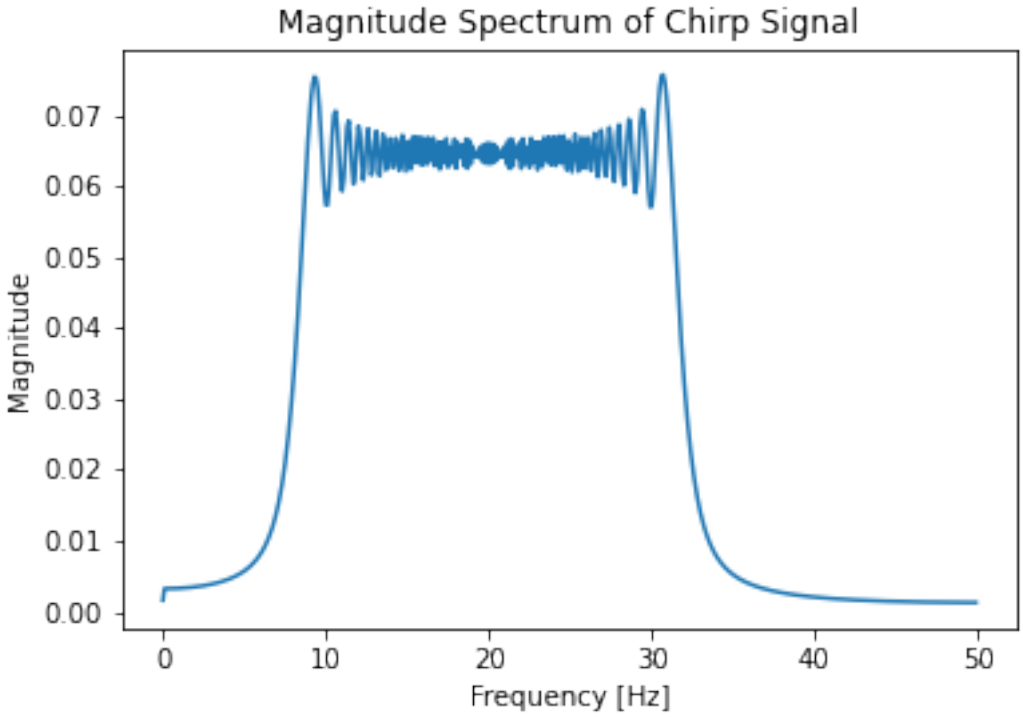

Figure 2: The Magnitude Spectrum of Chirp Signal

We can identify all the frequencies present in the entire signal as a whole. Thus FFT gives us all the frequencies in the chirp signal. However if we recall, the frequency was varying linearly as a function of time, meaning at a particular moment in time only one frequency was present in our signal. This sort of linear change of frequency with time was not captured in the magnitude spectrum. We observe that we have lost the track of time information. The magnitude spectrum is not able to tell us what frequency was present at $t = 0$ seconds or $t = 5$ seconds etc. While Fourier spectra do an excellent job of identifying the frequency content of individual tones, they are not as useful for analyzing several tones in a passage as it fails to capture changes in frequency. Thus it is desirable to associate each frequency with a certain partition of time that defines the intervals during which the corresponding note is emitted. We need to find a different way to calculate features for our system such that it has frequency values along with the time at which they were observed. Here Spectrograms come into the picture.

(Part 3: Spectrogram of Chirp Signal)

(Solution)

Visual representation of frequencies of a given signal with time is called Spectrogram. In a

spectrogram representation plot — one axis represents the time, the second axis represents frequencies and the colors represent magnitude (amplitude) of the observed frequency at a particular time with bright colors representing strong frequencies. Figure 3 shows the spectrogram for our chirp signal.

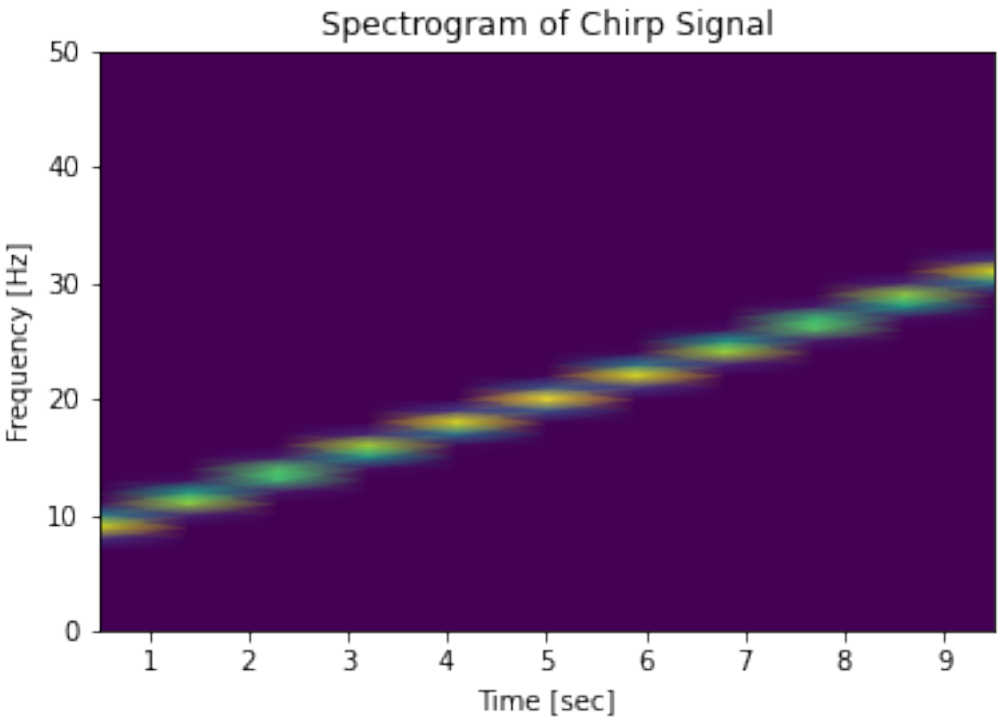

Figure 3: The Spectrogram of Chirp Signal

Spectrograms are a moving sequence of local spectra for the signal. In order to isolate the individual tones in the chirp signal, the signal $x(t)$ is multiplied by a succession of time-windows (or simply windows): $\{w(t-\tau_m)\}$, $m = 1, 2, ..., M$. Each window $w(t-\tau_m)$ is equal to 1 in a time interval $(\tau_m - \epsilon, \tau_m + \epsilon)$ centered at τ_m and decreases smoothly down to 0 for $t < \tau_{m-1} + \delta$ and $t > \tau_{m+1} - \delta$. We are assuming here that the values $\{\tau_m\}$ are separated by a uniform distance $\Delta\tau$, although this is not absolutely necessary. These windows also satisfy

$$\sum_{m=1}^{M} w(t-\tau_m) = 1 \tag{2.3}$$

Multiplying both sides of Eq 2.3 by $x(t)$ we see that

$$x(t) = \sum_{m=1}^{M} x(t)w(t-\tau_m) \tag{2.4}$$

so the sound signal $x(t)$ equals a sum of the subsignals $x(t)w(t-\tau_m)$, $m = 1, 2, ..., M$. Each subsignal $x(t)w(t-\tau_m)$ is nonzero only within the interval $[\tau_{m-1} + \delta, \tau_{m+1} - \delta]$ centered on τ_m.

The domain of each subsignal $x(t)w(t-\tau_m)$ is restricted so that when an FFT is applied to the sequence $\{x(t_k)w(t_k-\tau_m)\}$, with points $t_k \in [\tau_{m-1}, \tau_{m+1}]$, then this FFT produces Fourier coefficients that are localized to the time interval $[\tau_{m-1}+\delta, \tau_{m+1}-\delta]$ for each m. This localization in time of Fourier coefficients constitutes the spectrogram and the method of obtaining it is called the Short Term Fourier Transform. The choice of window presents important trade-offs. Intuitively, if we expect the signal to have rapidly fluctuating properties, in order to enable fine temporal resolution we should use a narrow window. In contrast, if we expect our signal to have slowly fluctuating properties, we should use a wide window. There is no optimal way to pick the STFT window. Indeed, this trade-off between temporal and frequency resolution is a fundamental feature of the STFT. An appropriate amount of overlap will depend on the choice of window and on the requirements. One may wish to use a smaller overlap (or perhaps none at all) when computing a spectrogram, to maintain some statistical independence between individual segments.

In Figure 3 the spectrogram is computed using Hamming Window length of 100 samples and an overlap of 10 samples. The larger values of these spectra are displayed brighter; the plain regions represent values that are near zero in magnitude. The vertical scale on this figure is a frequency scale (in Hz), and the horizontal scale is a time scale (in sec). Thus we can investigate the frequency components at specific points in time. As expected we can observe the linear variation of frequency with time in our chirp signal. As seen from the FFT plot in Figure 2 the magnitudes of the frequency components are somewhat similar and hence we observe a similar level of brightness with time in the spectrogram. Unlike FFT, the spectrogram gives us a visual representation of the spectrum of frequencies of the chirp signal as it varies with time. Thus the spectrogram provides us a complete description of the sound signal with the benefits of both time and frequency domain representations in the time-frequency plane and it is preferable over FFT for such a signal. This gives us two axes to work with in terms of resolution. Spectral resolution or frequency resolution of a spectrogram is its ability to resolve two signals with similar spectral content and time resolution of a spectrogram is its ability to resolve two signals in time, that is to differentiate between 2 pulses that are certain distance apart. For a given window function, the width of the main lobe defines the ability to differentiate sharply between 2 nearby frequencies, and the width of the window in the time domain defines the ability to differentiate between 2 pulses in the time domain. A narrower window means better time resolution, but a poorer frequency resolution due to widening of the main lobe, and a wider window leads to narrowing of the main lobe, i.e better frequency resolution,but a poorer time resolution. Figure 4 demonstrates this trade-off by taking spectrograms for Hamming Window of different window lengths.

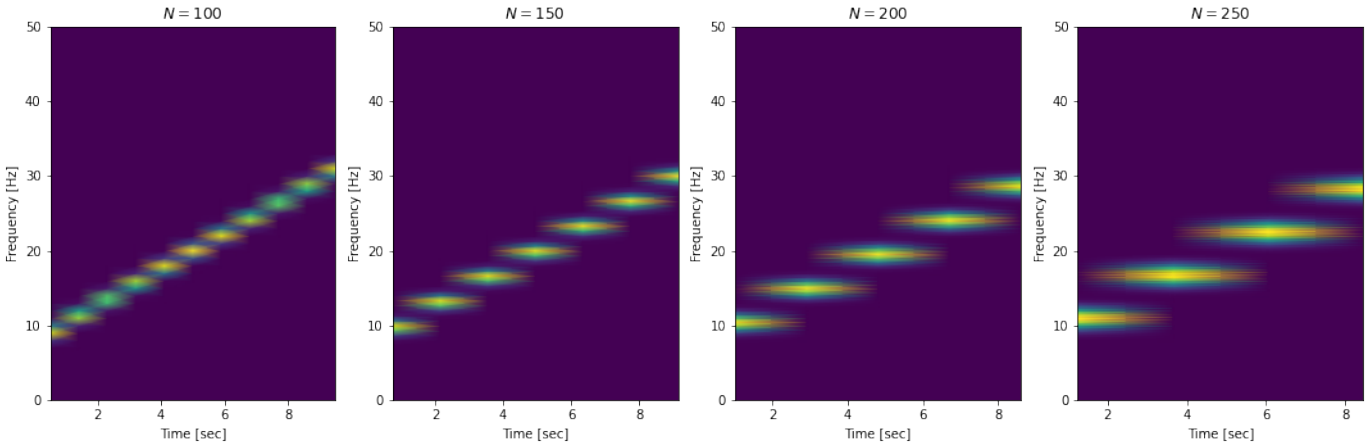

Figure 4: The Variation of Spectrogram of Chirp Signal with Window Length N

As the value of N increases from left to right, we observe that the lines get brighter indicating a better ability to distinguish similar frequency content and hence a higher frequency resolution. However we notice that we lose the ability to exactly pin point the time during which a particular frequency occurs as higher length indicates fewer number of windows for a given signal length and hence it assumes similar frequency content for the whole duration of the window reducing the time resolution.

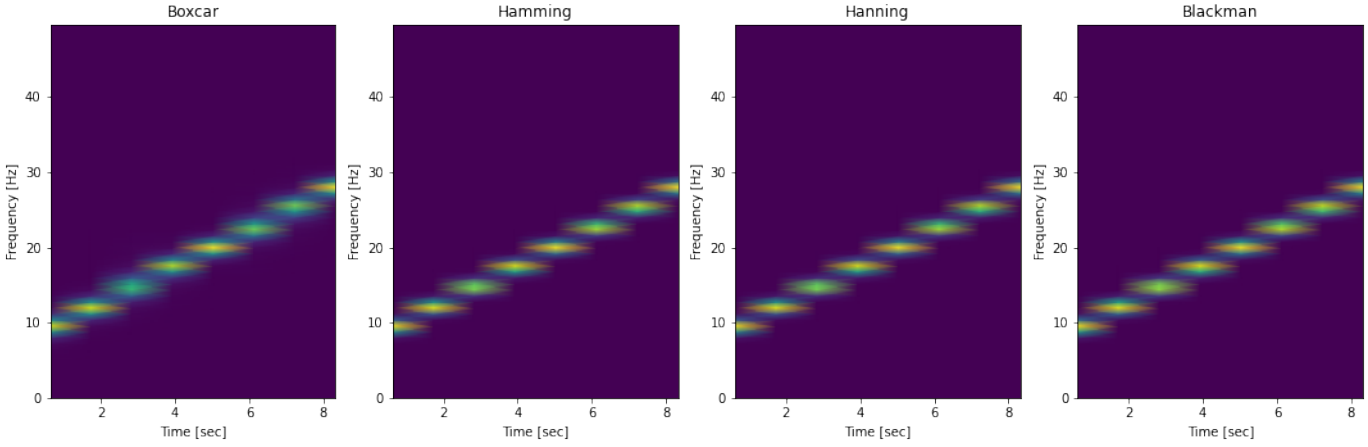

Figure 5: Spectrogram of Chrip Signal for Different Windows

We next investigate the spectrogram for different window functions of length 125. Figure 5 shows the spectrograms computed using different window functions. In the last experiment we discussed that when the number of periods in the acquisition is not an integer multiple of the period, the endpoints are discontinuous. These artificial discontinuities show up in the DFT as frequency components not present in the original signal. The spectrum we get by using a the DFT, therefore, is not the actual spectrum of the original signal, but a smeared version. It appears as if energy at one frequency leaks into other frequencies. This phenomenon is known

as spectral leakage, which causes the fine spectral lines to spread into wider signals. During STFT as we are acquiring the signal over several windows and calculating the FFT, we can expect spectral leakage to occur each time we compute FFT. This materializes as sort of blur corresponding to the side lobes in the spectrogram. Indeed in Figure 5 we can observe that the Boxcar or no window spectrogram has the highest blur between spectral lines as it has the highest spectral leakage and the Blackman window spectrogram has the least blur because it does a good job of suppressing the side lobes. Between the Hamming and the Hanning window, although difficult to see, the Hamming window does a better job of cancelling the nearest side lobe, but a poorer job of cancelling the rest.

2. Pitch Extraction

(Part 1: Spectrogram of `instru3.wav`)

(Solution)

Sounds from musical instruments are time-evolving superpositions of several pure tones, or sinusoidal waves. The frequencies of these sinusoidal waves are all integral multiples of a smallest base frequency. This base frequency is called the fundamental and the integral multiples of this fundamental are called overtones. In music theory, pitch is the perceived fundamental frequency of a sound, however the actual fundamental frequency may differ from the perceived because of the overtones. Our notion of pitch in experiment 3 was an absolute, precise and unambiguous measurement, but pitch is relative. Though pitch and frequency are not equivalent, they are correlated. This means that as one goes up, the other does as well. A higher frequency produces a higher pitch, and a lower frequency produces a lower pitch. A pure tone having a single pitch can be thus associated with a single frequency. Frequency describes a physical phenomenon, while pitch describes a perceptual phenomenon. Figure 6 shows the spectrogram of the instrument signal using Hanning window of length 512 samples and 25% overlap. We can see that each line corresponds to the different harmonics present in the signal. The fundamental pitch as described is the smallest base frequency and can located as the first line in the spectrogram at 783.278 Hz. Following the conventional approach Figure 7 shows the magnitude spectrum of the instrument track and we locate the fundamental pitch using peak picking and we find that it is equal to the 783.278 Hz, the same answer obtained using the spectrogram. As a matter of fact, the magnitude spectrum can be thought of as a side view of the spectrogram with the time axis facing towards us and each line in the spectrogram corresponding to a peak in the FFT.

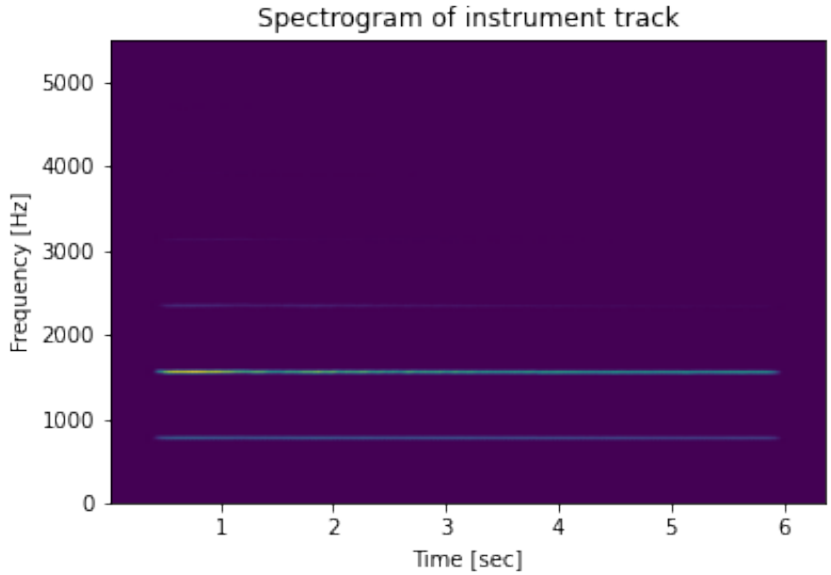

Figure 6: Spectrogram of instrument track

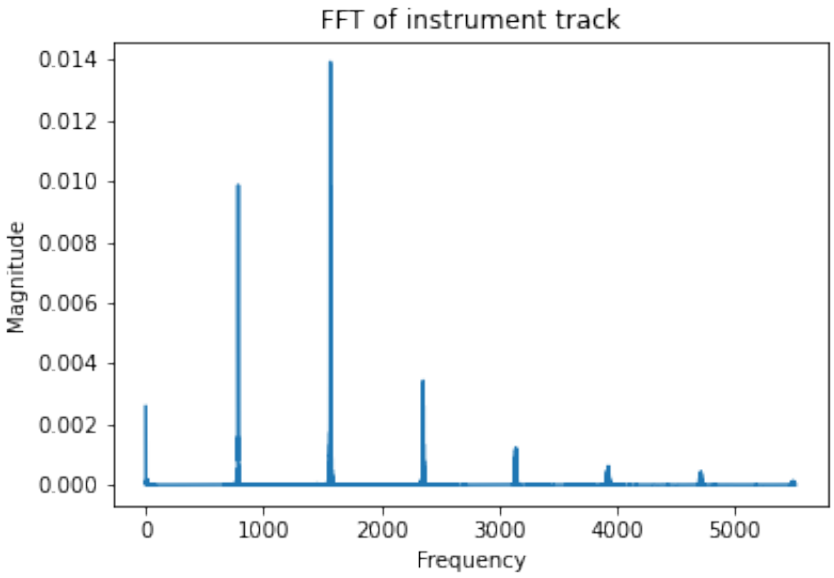

Figure 7: Magnitude Spectrum of instrument track

(Part 2: Spectrogram of `opera.wav`)

(Solution)

In experiment 3 we tried to capture the temporal variations of the spectrum of `opera.wav` by manually performing STFT by dividing the signal into 10 intervals and then computing FFT. In the light of the knowledge of spectral leakage and the importance of windowing, we realize that the approach was not correct as there is possibility to misinterpret the side lobes as part of

the signal. More importantly interpreting variations by comparing FFTs of multiple intervals was a tedious task and we had to only assume that the frequency remained the same in that interval, also there was no direct correlation of the frequency to time and we had to resort to interval numbers and interval length complicating the entire process. Figure 8 shows the spectrogram of the opera signal resampled to 8000 samples/second, using Hanning window of length 200 and overlap of 12.5%

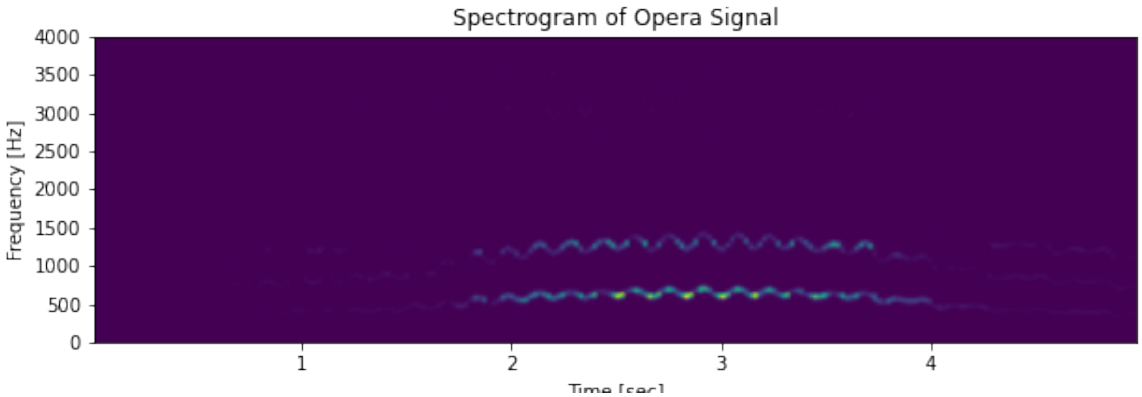

Figure 8: Spectrogram of Opera Signal

We can at once see the variation of frequency with time. The larger values of these spectra are displayed brighter; the plain regions represent values that are near zero in magnitude. The vertical scale on this figure is a frequency scale (in Hz), and the horizontal scale is a time scale (in sec). Thus we can investigate the fundamental pitch at specific points in time. We observe that the frequency increases with time, becomes maximum at around $t = 3$ seconds and decreases from there. If we listen to the track this conforms well with our perception of pitch as we hear the singer starts singing with a lower pitch which increases till about $t = 3$ seconds and lowers from there. Hence spectrogram gives us a visual representation of the spectrum of frequencies of the opera signal as it varies with time and the use of Hanning window reduces spectral leakage leading to better interpretation of results in comparison to experiment 3. We can also observe multiple lines corresponding to the overtones of the human voice. The wavy nature of the spectral lines illustrates vibrato, a vocal technique used by opera singers.

3. Spectro-Temporal Analysis of Speech

(Mapping Location of *phoneme* to the Spectrogram)

(Solution)

Phonemes are any of the perceptually distinct units of sound in a specified language that distinguish one word from another, for example p, b, d, and t in the English words pad, pat, bad, and bat. In English, there are 44 phonemes, or word sounds that make up the language. They're divided into 19 consonants, 7 digraphs, 5 'r-controlled' sounds, 5 long vowels, 5 short vowels, 2 'oo' sounds, 2 diphthongs. Our goal is to map the phoneme locations to the spectrogram. Figure 9 shows the spectrogram of the audio signal of my name at $F_s = 4000$ samples/second computed using Hanning window of length 350 and overlap of 12.5%.

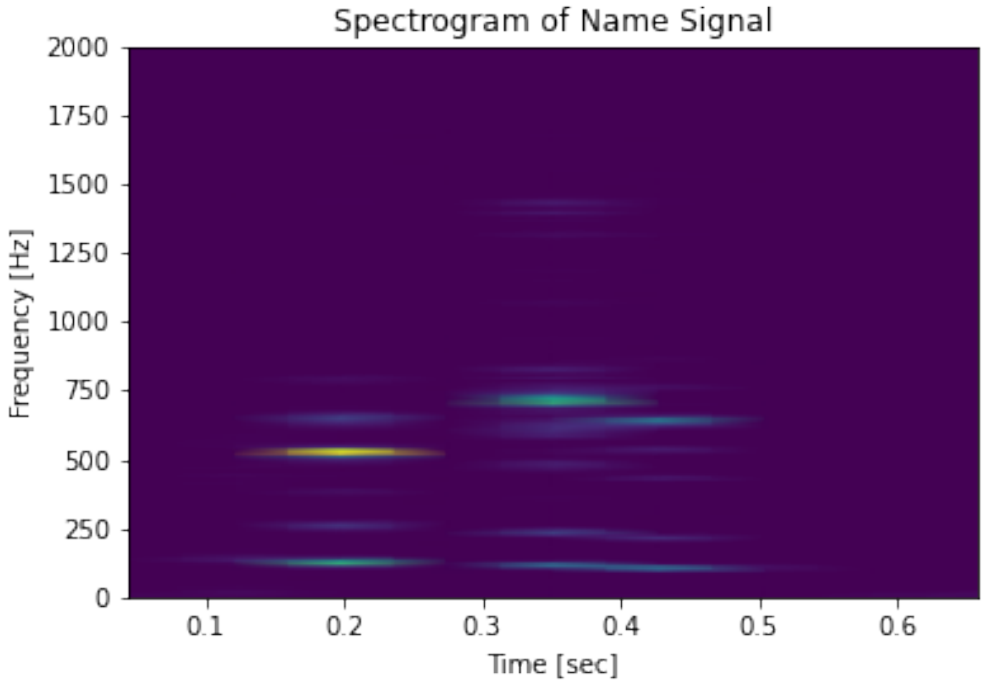

Figure 9: Spectrogram of Name Signal

We can break down the word *Prasad* phonetically as prAs-uhd. As phoneme refers to the smallest phonetic unit of sound in a language, we can further identify the phonemes as /p/ /r/ /a/ /s/ /a/ /d/. We can observe two proper groups of spectral lines at about $t = 0.2$ seconds, $t = 0.35$ seconds and light group of lines at $t = 0.45$ seconds. These lines correspond to the fundamentals and overtones of the human voice. The two major groups of lines correspond to the two syllables in *Prasad*. The first syllable contains the phonemes /p/ /r/ /a/ and hence we can say that those phonemes are mapped to the first group of lines. The second syllable in *Prasad* contains the phonemes /s/ /a/ /d/ and hence we can say that those phonemes are mapped to the second group of lines.