CS6300 Speech Technology: Assignment 3 Report

-Team 7: Akash Reddy A, EE17B001 and Nikhil Mattapally, EE17B138

1 Linear Prediction Spectrum

For the first 4 questions, the sentence (b) ”Don’t ask me to carry an oily rag like that.” from Assignment 1 (`s_b1.wav`) has been used.

The waveform looks as follows:

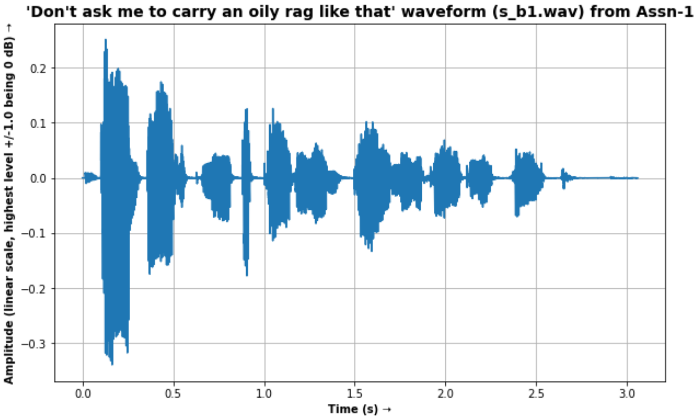
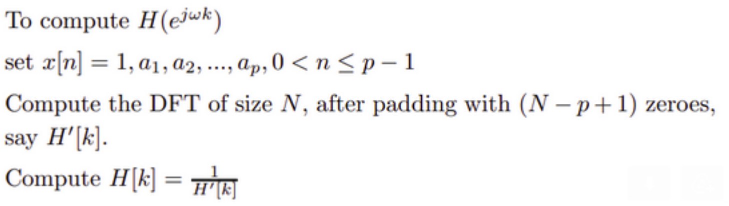


Figure 1: Waveform of Selected Sentence

1.1 Implementation

In order to compute the LP spectrum, the following steps are taken:

- The sampled waveform is windowed by a small window starting at the origin (here, 25 ms long). Since the sample rate is 16 kHz, the width of the window in terms of number of samples = 16000*0.025 = 400 samples.
- This windowed piece is zero-padded to obtain length = closest greater power of 2. This is for ease of DFT calculation using the FFT algorithm. Here, the windowed pieces are padded to 512 points.
- This windowed piece is multiplied by a **Hamming Window** to attenuate stop band interference because of truncation, and FFT is calculated for the piece to obtain a DFT spectrum.
- Next, the LPC coefficients are calculated using `librosa.lpc`, by giving the truncated piece to the function, and the required order p. We have chosen $p = 18$, so we obtain a 19-point LPC coefficient filter, where the first point is 1 and the next 18 are the 18 coefficients.
- Using this LPC coefficients filter, the LP spectrum is calculated as follows, where N is the power of 2 obtained in the second step (size of DFT), and p is the order of the LPC:



To compute $H(e^{j\omega k})$

set $x[n] = 1, a_1, a_2, ..., a_p, 0 < n \leq p - 1$

Compute the DFT of size N, after padding with $(N - p + 1)$ zeroes, say $H'[k]$.

Compute $H[k] = \frac{1}{H'[k]}$

- The reciprocal of the LPC filter's DFT gives the LP spectrum as shown above. The frequency axis is scaled to accurately reflect the unit of Hertz, and the **log-magnitude** LP spectrum is plotted against the **log-magnitude** original DFT spectrum. The LP spectrum neatly follows the trend in the original spectrum. We observe the smoothing effect of the LP spectrum in the frequency domain. It is as if the frequency domain DFT is passed through a low-pass filter.

- The window is shifted by a fixed shift duration (here, 10 ms = 16000*0.01 = 160 samples) and the process is repeated to obtain the LP spectrum for the next window.

- (The same window size and shift size have been used for windowing in further questions, such as short-term energy and ZCR in question 3, autocorrelation function in question 4, and the spectrogram and VAD in question 5.)

The formants of the windowed piece of audio can be easily obtained using the LP spectrum, whose clearly-defined peaks are the macro-level peaks (centres of dominant frequency) in the original DFT. The function `scipy.signal.find_peaks` has been used to find the peaks in the LP spectrum.

1.2 Observations

Below is the plot of the original DFT, LP spectrum, and identified formants for a random piece of the `s_b1.wav` audio.

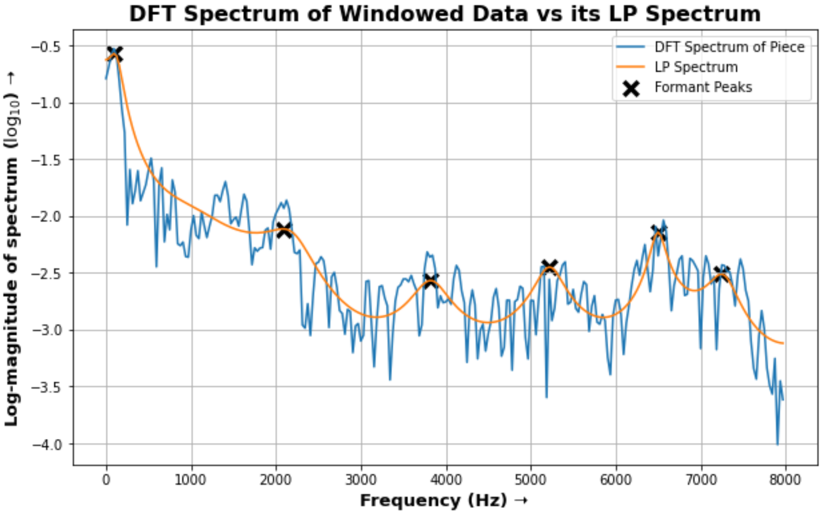


Figure 2: LP Spectrum and Formant Frequencies for a Window

2 Formant Contours

Using the LP spectrum method described above for each window, the formant contours vs time can be obtained. They are displayed in the below figure.

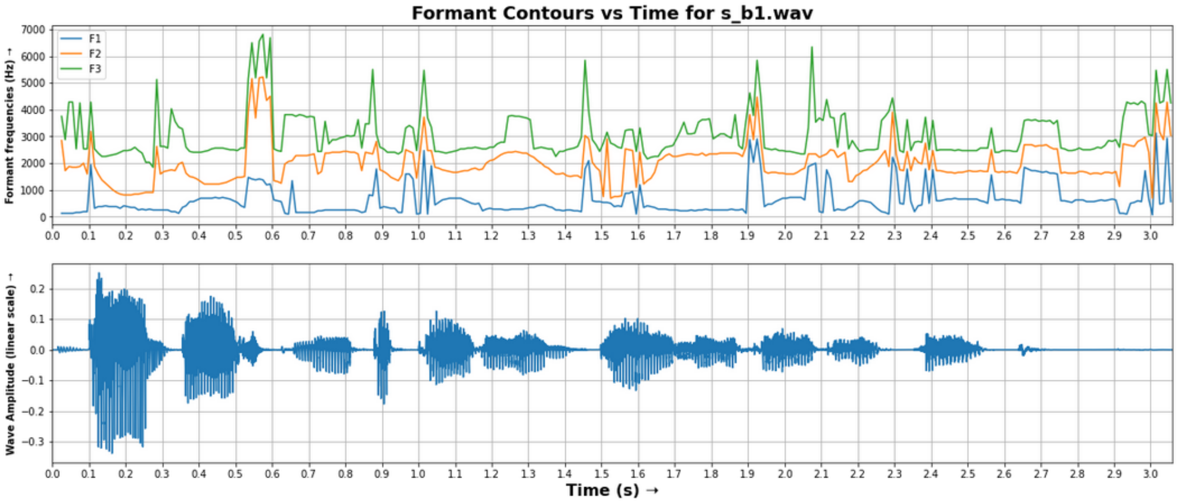


Figure 3: Formant Contours for s_b1.wav

And below is the Wavesurfer Formant Plot to compare:

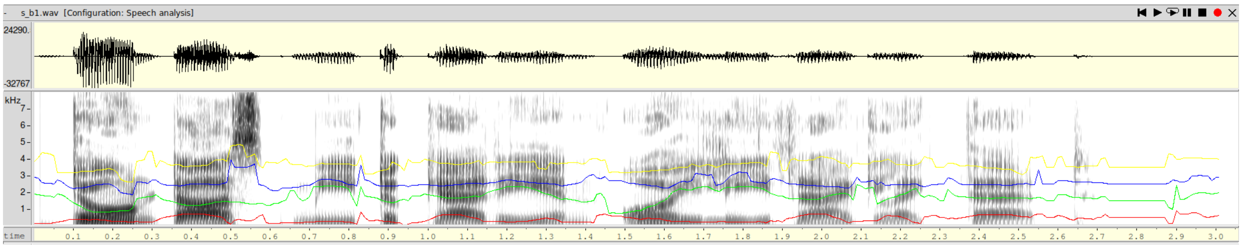


Figure 4: Wavesurfer Formant Contours for s_b1.wav

- We can see that, excepting for some errors (mostly in areas where the spectrogram is sparse or not prominent, which makes it hard to evaluate formants), the obtained formant contours roughly align with those obtained from Wavesurfer in the areas with a reasonably clear spectrogram atleast. Errors can also occur where the LP spectrum is noisy, with many peaks surrounding the one that we want to extract for formants.
- The formant contours for this particular sentence oscillate a lot, owing to combinations of voiced and unvoiced consonants, longer and shorter vowels, and diphthongs.
- We can see that F1 is roughly in the range of 100-1500 Hz. F2 remains roughly in the range of 1000-300 Hz, and F3 in the range of 2000-4000 Hz. This is consistent with our general idea of speech.

3 Voice-Activity Detection (VAD)

While formants can be obtained for both voiced and unvoiced parts of speech (since they are representative of the dominant frequencies of the entire vocal tract including articulators), quantities like pitch depend largely on the voice, or the vibration of vocal cords. It becomes extremely important to identify which regions of speech are voiced, so that we obtain accurate values of pitch.

Short-term Energy and **Short-term Zero Crossing Rate (ZCR)** can be used to perform Voice-Activity Detection.

Short-term Energy at a given point is calculated across the window preceding that point as:

$$e[n] = \sum_{k=n-N+1}^{n} s^2[k]$$

where N is the window size and $s[k]$ is the amplitude of the waveform at the kth index.

Similarly, short-term ZCR at a given point is calculated as:

$$z[n] = \frac{1}{N} \sum_{k=n-N+1}^{n} |sgn[s[k]] - sgn[s[k+1]]|$$

and it is a normalized measure of how many times the signal crosses zero in the window.

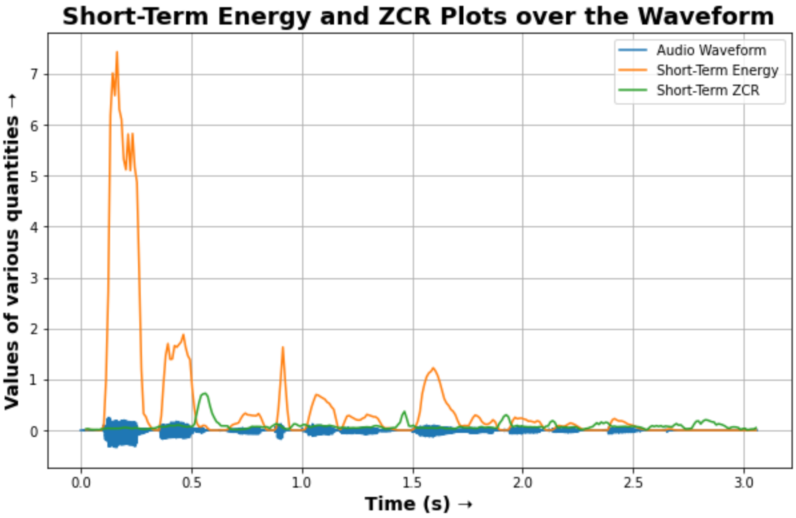


Figure 5: Short-Term Energy and Zero-Crossing Rate Curves on top of Waveform

Above is the plot of the short-term energy and ZCR curves lying on top of the waveform. We can already observe that the voiced regions seem to have **high energy** and **low ZCR**, and vice versa for the unvoiced regions.

We now formally identify the voiced and unvoiced regions. The thresholds for short-term energy and ZCR have been obtained from the paper titled "An Algorithm for Determining the Endpoints of Isolated Utterances" by Rabiner and Sambur, 1976. They are evaluated as follows:

- **Threshold for ZCR:**

  $$ZCRT = min(25/10ms, ZCR' + 2\sigma)$$

  where ZCR' is the mean and σ is the standard deviation of the ZCR in silence region. We replace $10ms$ with the number of samples in 10 ms (here, 160) to obtain the first term in the $min(.)$ operator.

- **Threshold for Energy:**

  $$\begin{aligned} IMX &= max(e[n]) \\ IMN &= mean(silence_energies) \\ I1 &= 0.03 \times (IMX - IMN) + IMN \\ I2 &= 4 * IMN \\ ET &= min(I1, I2) \end{aligned}$$

  In the paper, what we have calculated as the final energy threshold (ET) is actually the LOWER energy threshold (ETL) above which the region is considered voiced ONLY if it also crosses another UPPER energy threshold (ETU) before dropping back below the ETL. However, we have considered only one energy threshold (ETL) as the final ET for simplicity, and it gives good enough results.

We now take an ensemble of both thresholds to obtain the voiced regions (logical 'AND' of waveform samples that satisfy BOTH thresholds - higher energy than ET and lower ZCR than $ZCRT$). The voiced regions are marked over the original waveform in the plot below.

We observe that the voiced regions agree very well with our observation of the waveform.

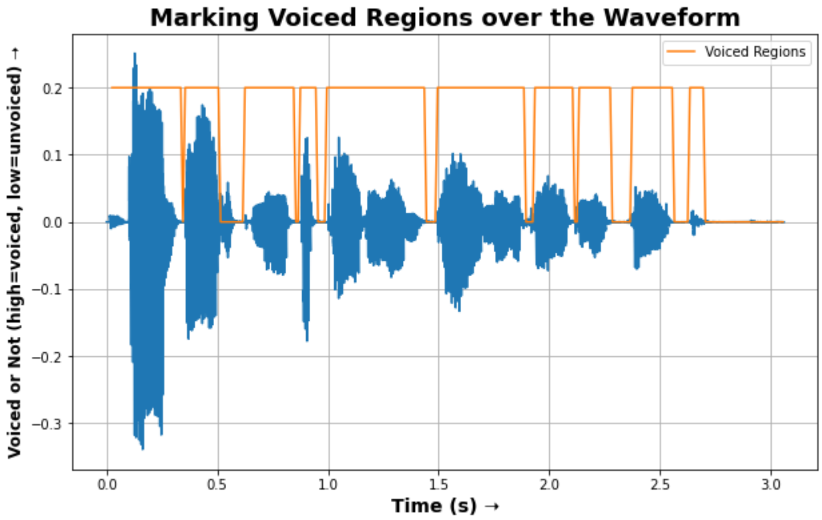


Figure 6: Voiced Regions on top of Waveform

4 Pitch Extraction using Autocorrelation

- Pitch can be identified well in voiced regions as it is a characteristic of the voice. Therefore, the unvoiced regions in the waveform are first masked out.

- The autocorrelation method works by taking short-term autocorrelations of windowed pieces of the waveform, and identifying the period of the highest peak in the time domain. This gives an approximate value of the pitch period. The short-term autocorrelation is given by:

$$R_n(k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)x(m+k)w(n-m-k)$$

  where w is the window function - therefore the piece that is correlated with itself is windowed and the other samples are made into zeroes.

- The short-term autocorrelation function obtained this way is usually noisy because it tends to retain too much information about the speech signal, more than we require to extract the pitch. There are many more peaks than we require due to the oscillations of the vocal tract response, which determine the shape of each period of the signal. Therefore it becomes difficult to extract the required peak that corresponds to the pitch period.

- Therefore, it helps to make the periodicity prominent while suppressing other features of the signal, and this is done by ”center clipping” of the waveform as follows:

$$\begin{aligned} s_c[n] = \; & s[n] && |s[n]| > C_L \\ & 0 && |s[n]| \leq C_L \end{aligned}$$

- According to *Digital Processing of Speech Signals* by Schafer and Rabiner, a ”3-level clipping” does even better in this aspect, and is also easier to compute the autocorrelation function for.

$$\begin{aligned} s_c[n] = \; & +1 && s[n] > C_L \\ & 0 && |s[n]| \leq C_L \\ & -1 && s[n] < -C_L \end{aligned}$$

So finally, shown below is the voiced-region-masked, 3-level clipped version of the waveform.

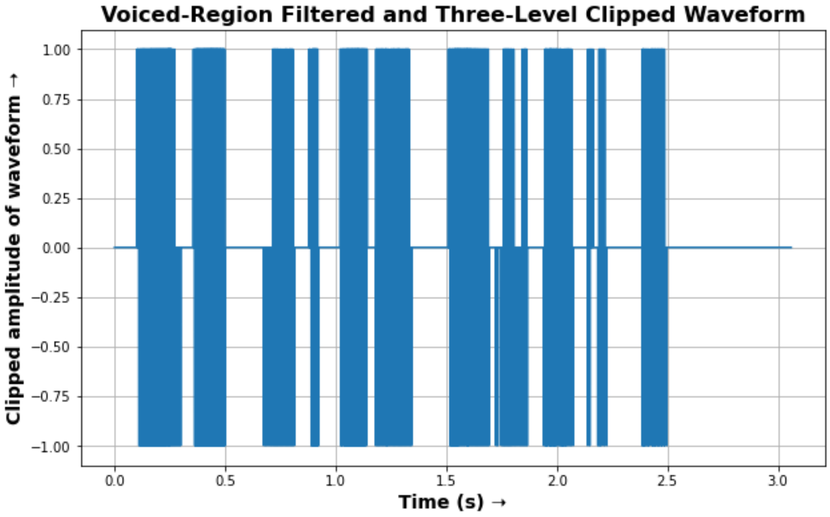


Figure 7: Voiced-Region-Masked, 3-Level Clipped Waveform

The pitch is then calculated by performing short-term autocorrelations on this waveform, and calculating the peaks to obtain the pitch periods in terms of samples. Then, the pitch frequency is calculated as:

$$\text{Pitch frequency} = \frac{\text{Sample Rate}}{\text{Pitch Period in Samples}}$$

for each window. Finally, the pitch contour is plotted. The points in the unvoiced regions are taken care to be ignored in this plot.

In the below figure, it is compared with the pitch contour with Wavesurfer, and we see that the extracted pitch contour agrees well with the Wavesurfer contour.

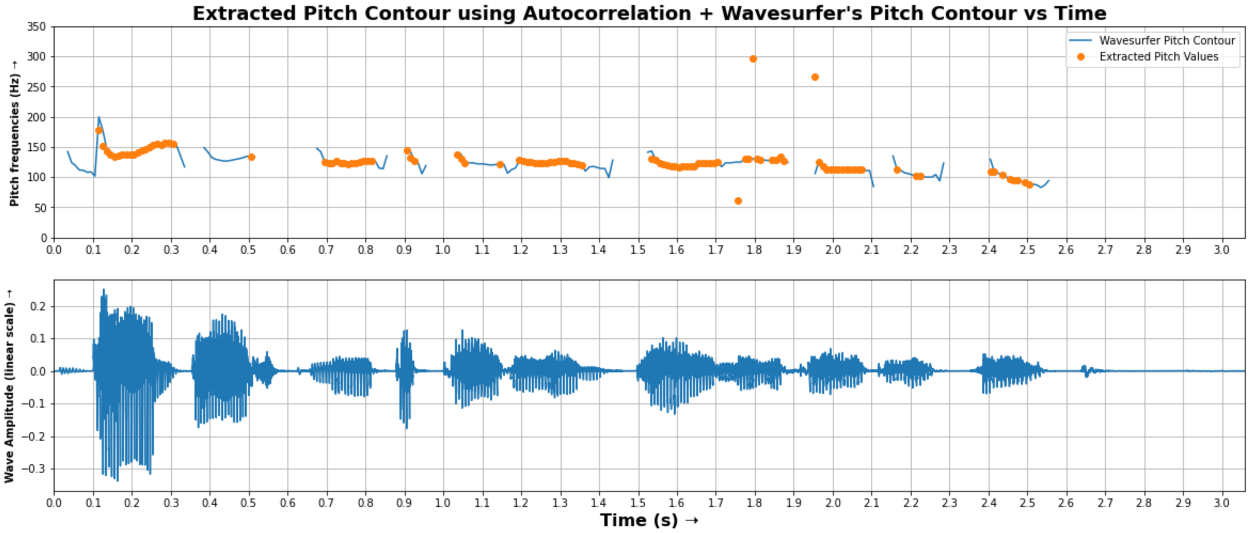


Figure 8: Extracted Pitch Contour and Wavesurfer Pitch Contour

5 Analysis of Recordings of We Were Away a Year Ago

The waveforms of both the male and female recordings of the sentence, ”We were away a year ago,” are shown below:

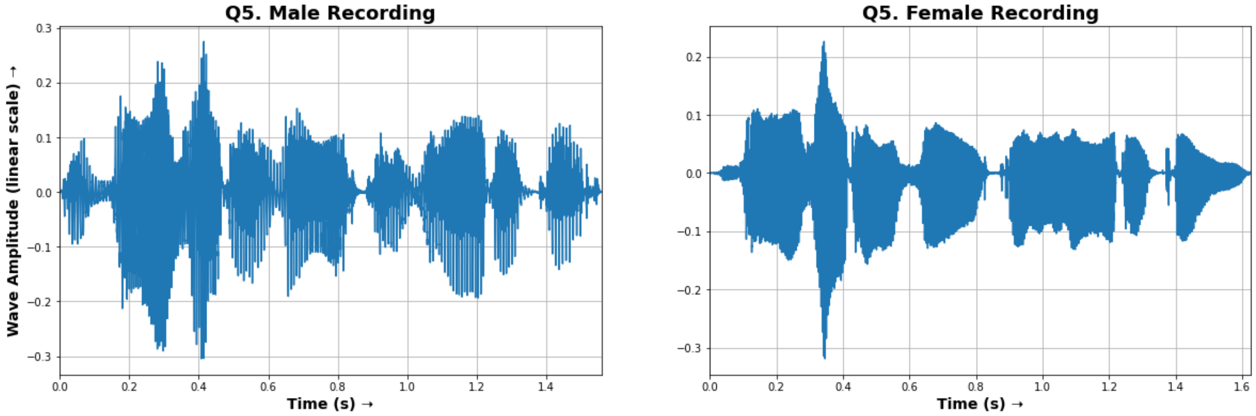


Figure 9: Waveforms of both recordings

- The spectrogram for the recordings is calculated by calculating the DFT of each window in the recording, and plotting it vertically in a column (for the corresponding time frame), with the x-axis representing time.
- The formant contours are calculated similar to question 2, using the LP spectrum method and finding its peaks.
- The pitch contours are calculated after performing VAD as in questions 3 and 4. Below is the plot of short-term energy and ZCR curves on top of the waveforms:

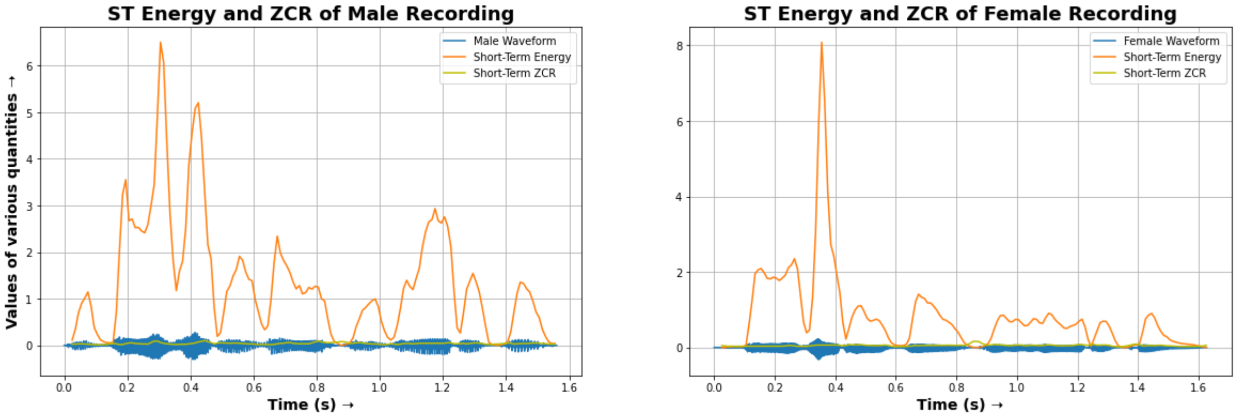


Figure 10: Short-Term Energy and ZCR plots for both recordings

- These are used to mark the voiced regions in the waveforms, and to mask out their unvoiced regions, as shown below:

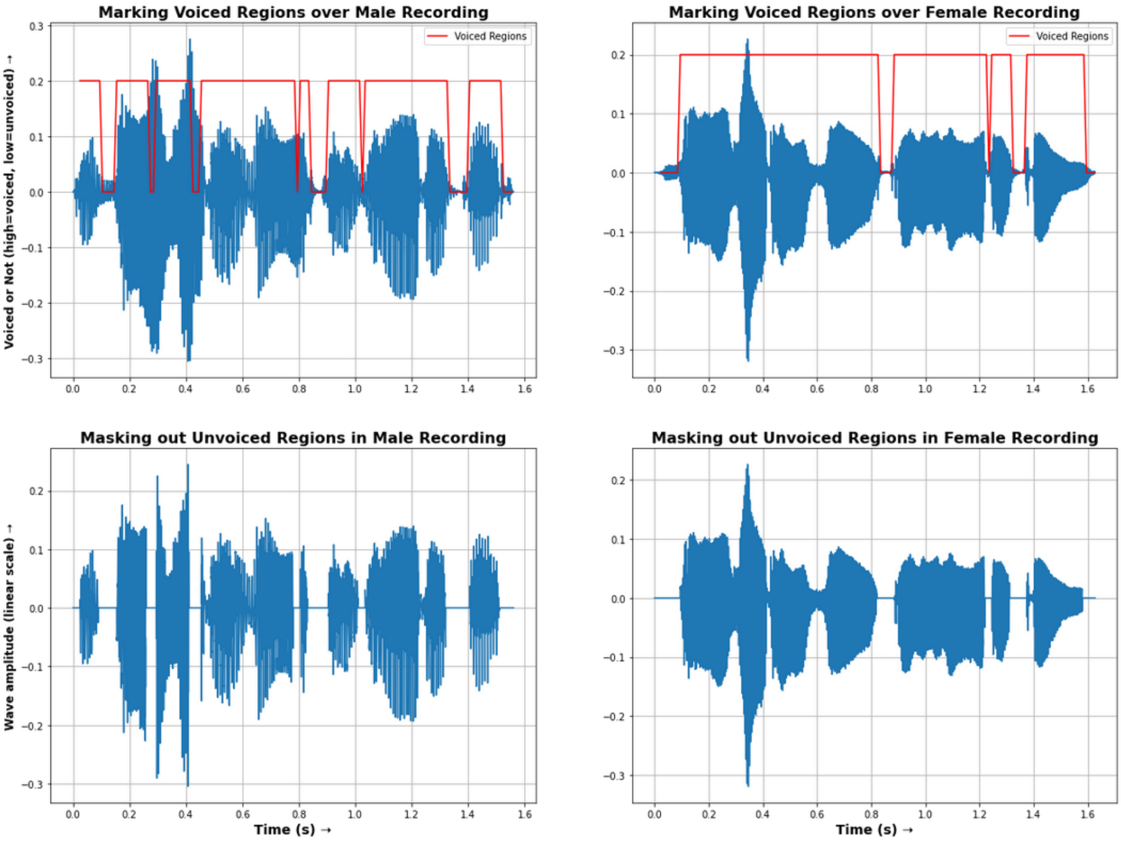


Figure 11: Voiced regions marked and unvoied regions masked out for both recordings

- Three-level clipping is then performed on the masked waveforms:

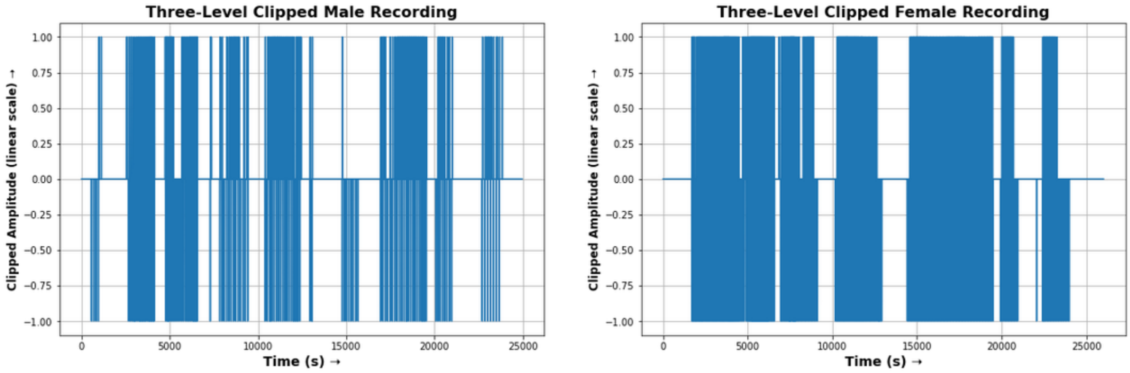


Figure 12: 3-level clipping for both recordings

- Finally, the pitch contours are extracted similar to question 4. The formant curves and pitch curves are overlaid on top of the spectrogram as expected. First, here is the male recording's obtained spectrogram. Below it is the Wavesurfer spectrogram for comparison. We see that apart from a few noisy points, the formant curves match up quite well. Even the spectrogram looks very similar in both plots.

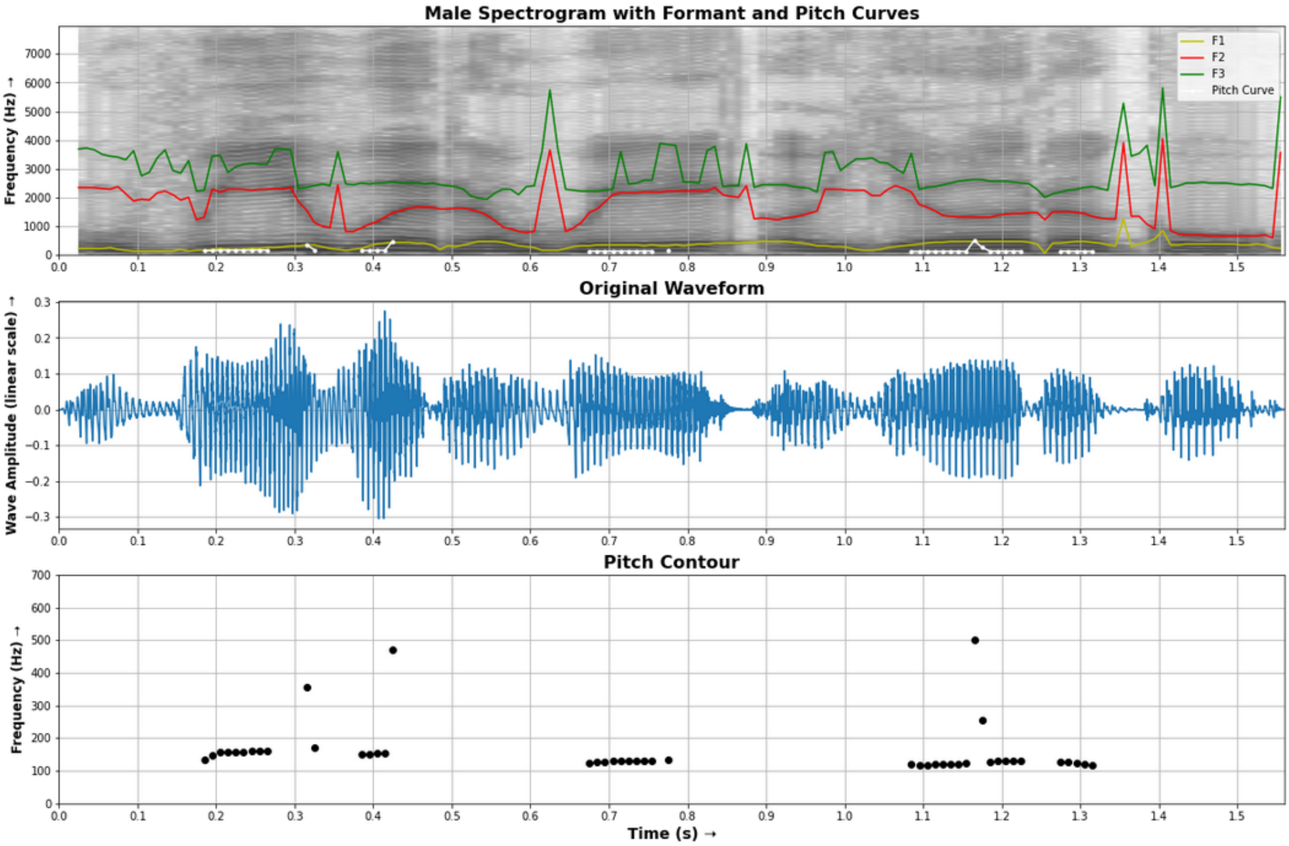


Figure 13: Spectrogram with formant and pitch curves for male recording

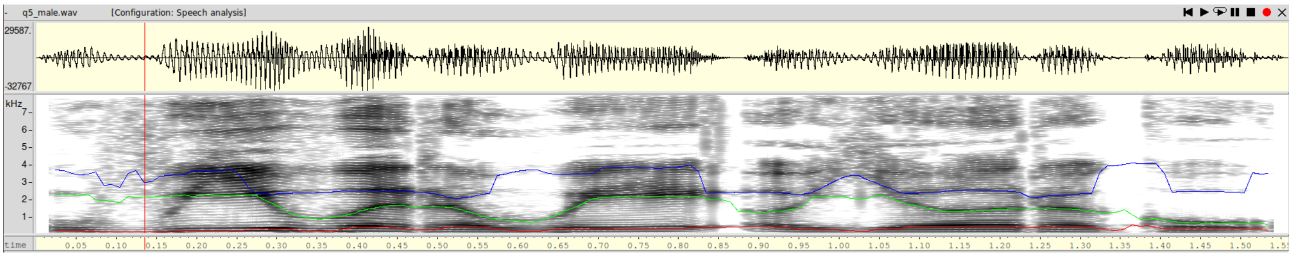


Figure 14: Wavesurfer spectrogram for male recording

- Similarly, the spectrogram and curve plots for the female recording are shown below, and the similarity is observed once again:

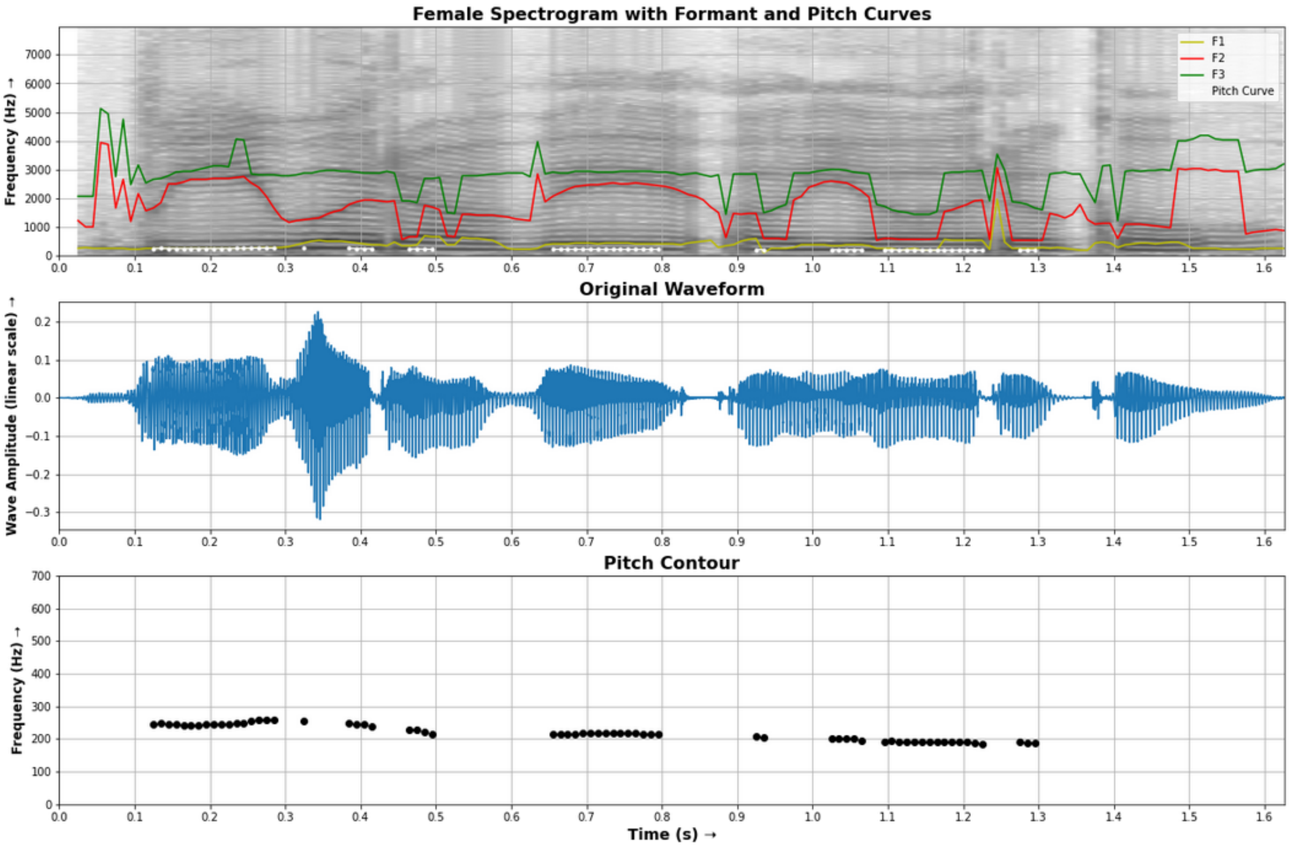


Figure 15: Spectrogram with formant and pitch curves for female recording

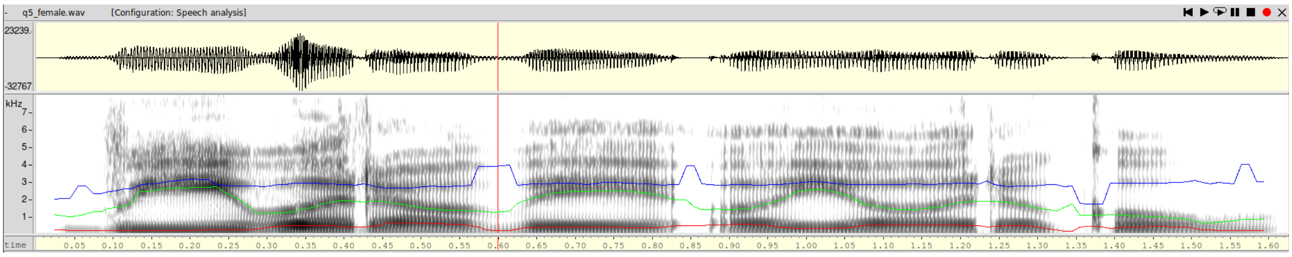


Figure 16: Wavesurfer spectrogram for female recording

- We have obtained good spectrograms and formant curves, as verified by the Wavesurfer plots.

- The pitch contours look much neater apart from very few noisy points, and clearly indicative of a pitch (as opposed to the question 4 plot). This is because the entire sentence ”We were away a year ago” is a voiced sentence. As a result, most of the sentence shows up as voiced in the VAD, and the pitch can be uniformly calculated.

- The pitch contour lies pretty close to the first formant as seen in the spectrogram. Both are representative of the fundamental frequency of the voiced sentence.

- The difference in pitch between male and female voices is clearly visible from the pitch contour. The male pitch contour wanders close to 150 Hz throughout, whereas the female pitch contour is closer to 250 Hz.

- In both sentences, we observe that the pitch drops very slowly towards the end of the sentence. This is because when we start uttering a sentence, we begin more energetically. As we reach the end of the sentence, our energy and subsequently, pitch drops as we have communicated most of our message and are in a hurry to wrap up the sentence. This fact is reflected in the pitch contour.

* *