

## Speech Signal Processing Assignment 1

### 1. Briefly explain about the following:

- **Coarticulation:**

Coarticulation is the way the brain organizes sequences of vowels and consonants, interweaving the individual movements necessary for each into one smooth whole. It takes about a fifth of a second to produce a syllable, or about a fifteenth or twentieth of a second for each consonant or vowel.

Eg: The word: Amma, we see that it's not Aa + Ma + Ma separately, instead each lip movement & constriction overlaps with each other.

- **Phonation**

The physical process behind sound production, called **phonation**, works the same way regardless of how much sound is being produced, or the reason for the sound. It starts with air being pushed into lungs, the quantity of air is proportional to the strength of the sound being produced, the air is pushed into glottis (opening between vocal chords). These membranes are stretched across larynx, which is an organ that acts like a passage to lungs. The larynx holds the vocal folds and glottis. When air is pushed through the glottis, it causes pressure to drop in the larynx. This in turn makes the vocal folds vibrate, and this vibration is what produces 'voicing' (another name for physical sound that is produced in this way).

- **Fundamental Frequency**

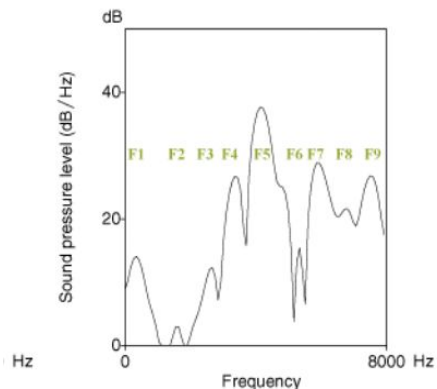
The fundamental frequency of a speech signal, often denoted by  $F_0$  or  $F_o$ , refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. The fundamental frequency is defined as the average number of oscillations per second and expressed in Hertz. Since the oscillation originates from an organic structure, it is not exactly periodic but contains significant fluctuations. Eg: The voiced speech of a typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz.

- Epochs

Epoch is the instant of significant excitation of the vocal-tract system during production of speech. For most voiced speech, the most significant excitation takes place around the instant of glottal closure.

- Formants

A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency, roughly one in each 1000Hz band. Or, to put it differently, formants occur at roughly 1000Hz intervals. Each formant corresponds to a resonance in the vocal tract. In the following diagram, we can see  $F_i$ ,  $i$  is the  $i$ th format in the speech signal.



- Pitch

Pitch, in speech, the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation. We generally see that children's pitch is highest, followed by female and male pitch. The pitch range for men's voices was 60–180 Hz, and the pitch range of women's voices was 160–300 Hz.

## **Q2. "Female pitch is more when compared to Male pitch." True or False. Justify the Statement with proper explanation.**

The larynx (voice box which holds the vocal folds) is more descended in males, which is why they tend to have an Adam's apple. This means that when the vocal folds generate a sound wave (from the vibration), the wavelength is longer (because it has further to travel along the vocal tract). Longer wavelength makes lower pitch. The sound wave carries all of the difference, beginning with the larynx and then being shaped by the vocal tract (longer for males) and the articulators (lips, tongue, etc). Men's vocal cords lengthen and thicken much more so than women's, resulting in the adult male voice pitch being on average half the frequency of adult female voice pitch. There's a measurable feature called "shimmer" which refers to the greater variability in pitch in female speech, but this along with other features are just a result of the physical differences in the vocal tract. Gender differences in jitter and shimmer at medium loudness may be mainly linked to different habitual voice loudness levels too.

## **3. What is speech ? How do speech signals differ from other signals?**

Speech in linguistics is defined as articulation of sounds with help of tongue, lips, jaws, vocal chords, and other speech organs. Speech sounds are mainly characterized by manner of articulation and place of articulation. Place of articulation refers to where the airstream in the mouth is constricted and manner of articulation in which organs interact, and how closely the air is restricted, what form of airstream is used - pulmonary, implosive, ejectives and clicks, whether or not vocal chords are vibrating, and whether nasal cavity is opened to the airstream.

The main difference between normal signals and speech signals is that most speech signals are non-stationary processes with multiple components that vary with time and frequency. We represent these signals with features in voiced-speech signals with fundamental frequency and formant. Each formant refers to concentration of acoustic energy around a particular frequency of the speech wave.

Q4. Record your mother's name which should be as "I am son of < mother's name >" or "I am daughter of < mother's name >" whichever category you belong to.

Ans : Heres, the link to the audio file of me voicing the following line :

"I am daughter of Padmaja "

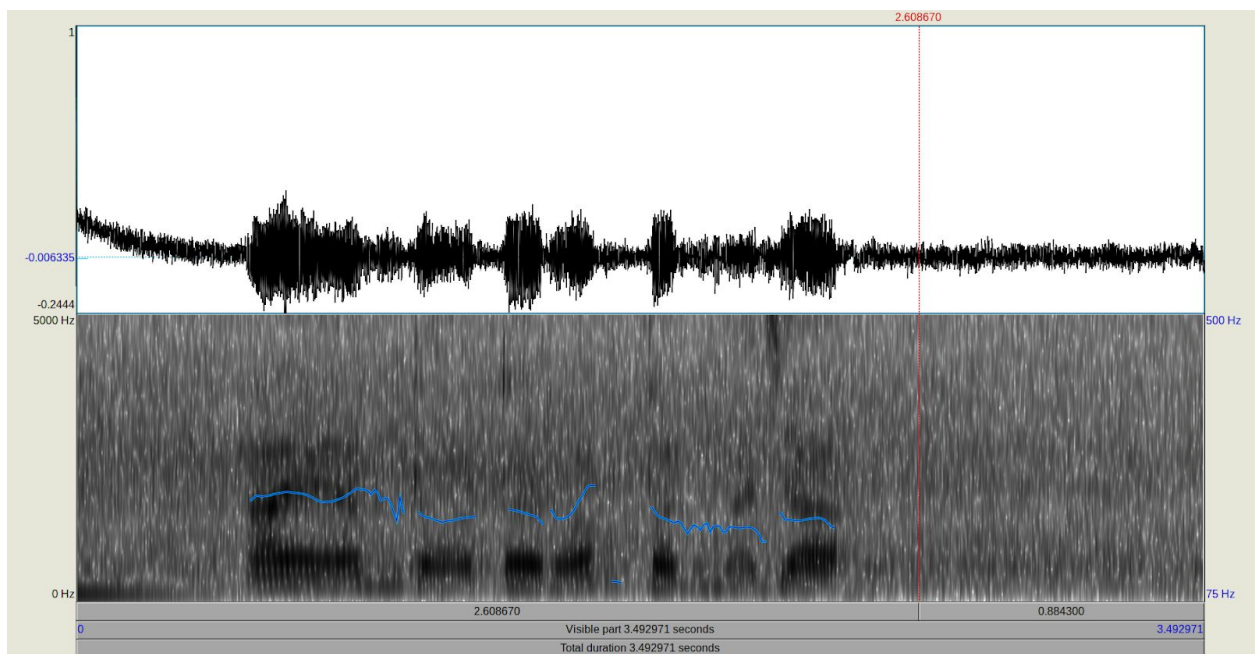
Voice Sample Link :

[https://drive.google.com/file/d/1OQhLO\\_ZjbEe0j1hwwCnEbFrPRAI6iq3g/view?usp=sharing](https://drive.google.com/file/d/1OQhLO_ZjbEe0j1hwwCnEbFrPRAI6iq3g/view?usp=sharing)

Text Grid Annotation Link :

<https://drive.google.com/file/d/1dRgu5l8aK-mfGqUPqgsKSCPdzQYnaEio/view?usp=sharing>

1. The following is the snapshot of the spectrogram of the entire speech signal , displaying the amplitude of various harmonics .

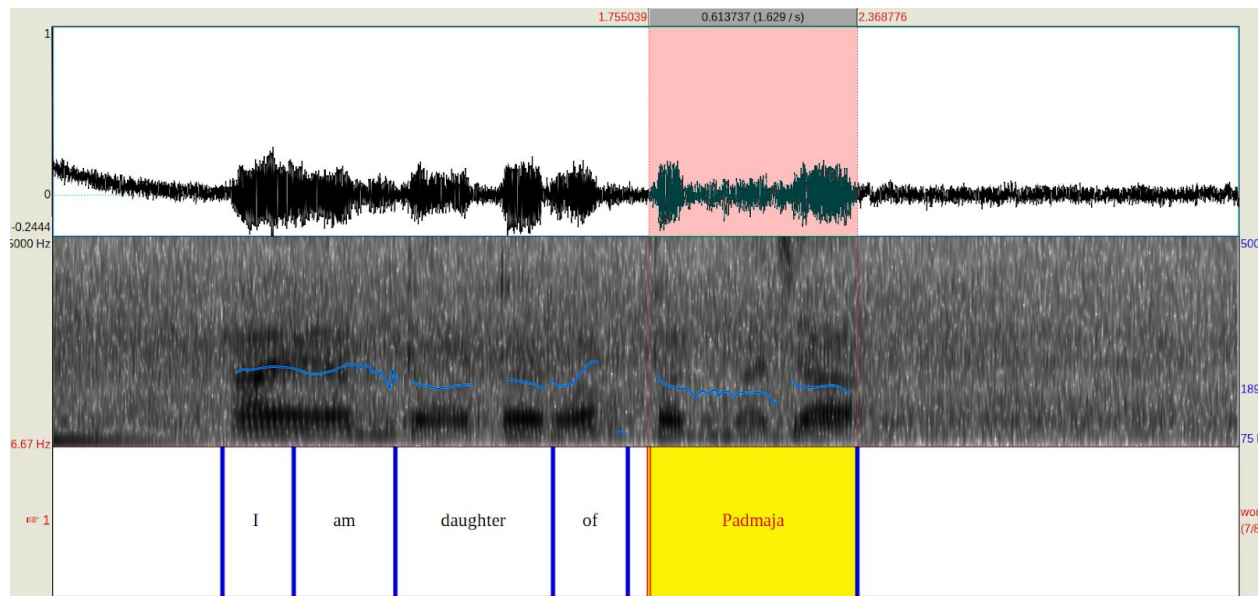


Spectrogram of the voice sample

2. So , we need to mark the following regions , first let's understand them :
- Voiced: In the case of vowels a regular formant structure (3 to 4 formant frequencies) and pitch harmonics (vertical striations in the case of wideband spectrogram) are used for identifying the voiced regions, where as nasals and voiced stops low frequency regions and pitch harmonics are used as clues. They are quasi - periodicity and high amplitude regions .
  - Unvoiced: Energy at high frequency regions and no regular formant structure. Here , the signals are non-periodic and random like noise .
  - Plosive: A silence bar followed by energy at high frequency regions. Noise burst like signal indicates the sudden release of constriction at different tracks in the vocal tract system .
  - Silence: No frequency components (white region), almost zero amplitude .

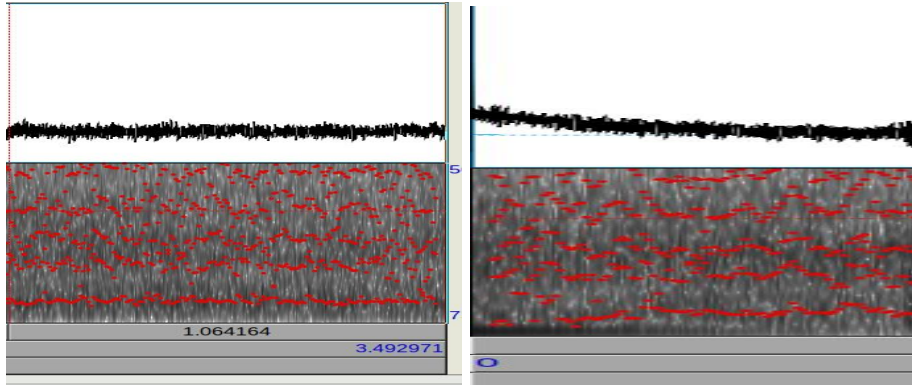
So, following will be the Text annotation of the file Link :

<https://drive.google.com/file/d/1dRgu5l8aK-mfGqUPqgsKSCPdzQYnaEio/view?usp=sharing>



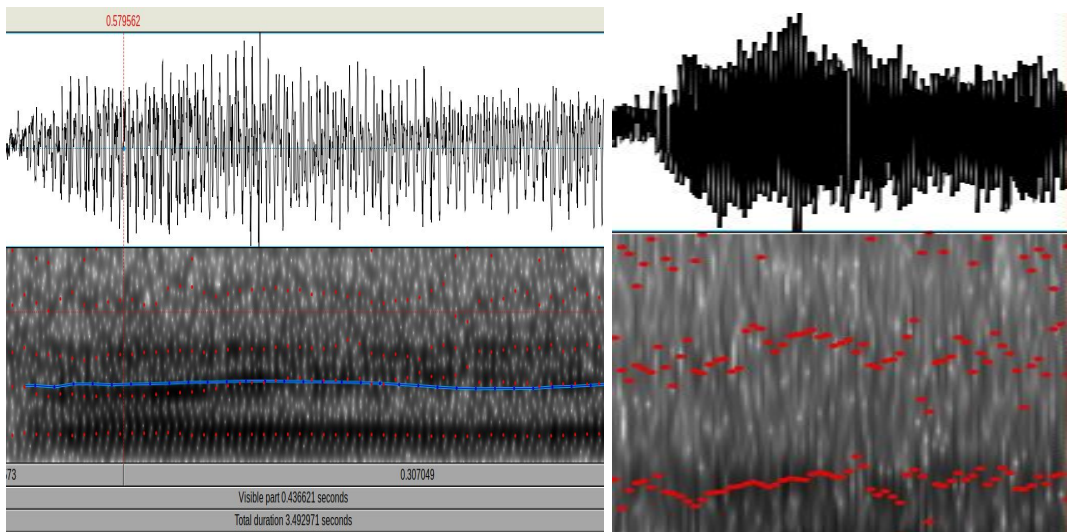
2.Regions of Unvoiced , Voiced , Plosive and Silent are (wrt time axis) :

1.Silent Parts : From  $t=0$  to  $t= 0.517$  secs, and between  $t= 2.42$  to  $t=3.49$  secs, the waveform contains background noise , and there are no harmonics and no amplitude .



2. Voiced Parts (l / aa/) {l pronounced it without a gap , hence heavy overlap }

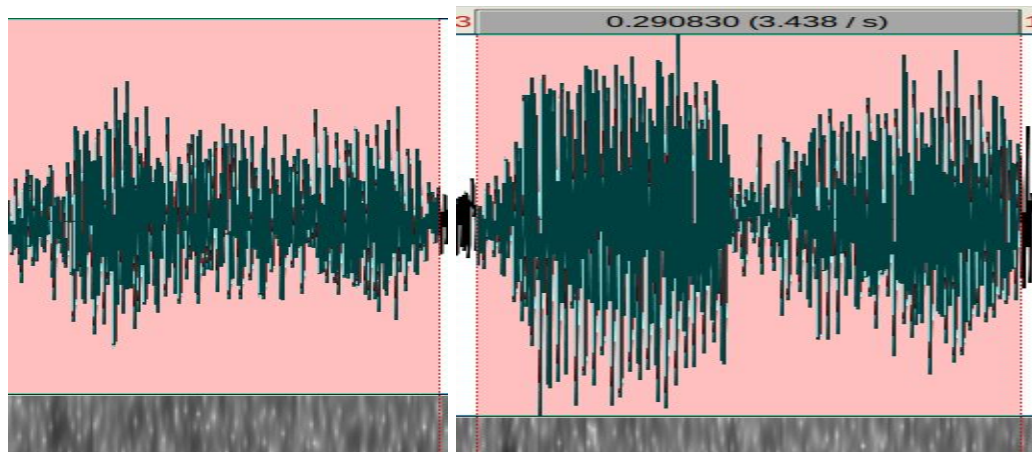
-t=0.597 sec to t= 0.88 secs , We can observe quasi-stationary , and periodicity in these signals.



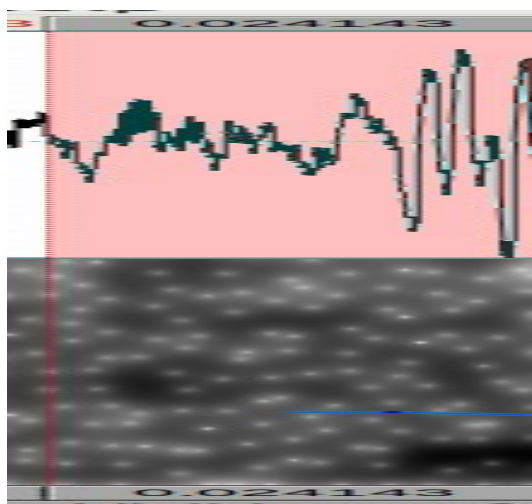
-t=1.04 sec to t= 1.45 secs , -- It's the part where "au-gh-" is pronounced .

On right , we have t=1.31 to t= 1.65 ( t/e/ /r/ /o/) {This part is specifically produced here}

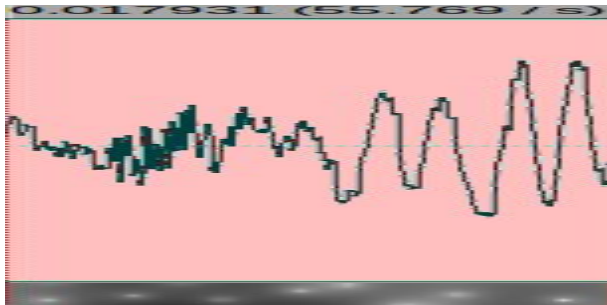




- Plosives : -Here , in my sentence there are 2 instances of plosives , at the beginning of the word - Daughter (/D/) and Padmaja.(/Pa/,/D/)
- Padmaja -- /P/ t=1.5 to t=1.7 (Small window) . {/D/ in Padmaja is not very properly pronounced in my voice sample , hence no plosive was found}

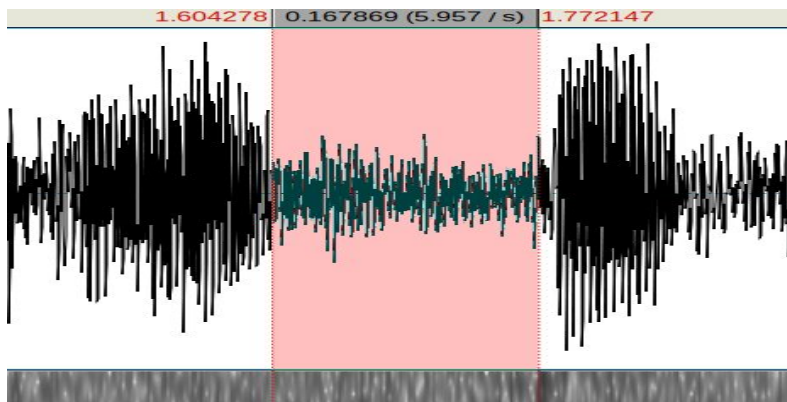


- Daughter - (/d/) -- We can observe heavy density waves at the beginning of the word .  $t = 1.31$  sec to  $t = 1.38$  sec

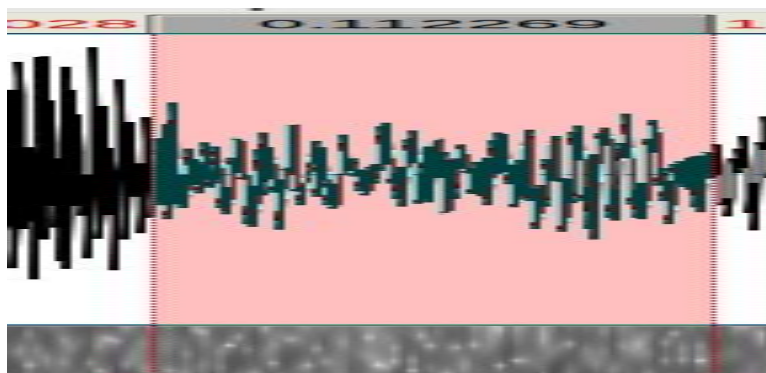


- Unvoiced Parts :

$t = 1.60$  to  $t = 1.75$  sec , it's a continuous part of a speech , but no formants are present ,and high energy spikes at formants. (On zooming they have no regular structure) { It shouldn't be confused as a part of noise / plosive , since its a part of a word }



$t = 1.8$ sec to  $t = 2.0$  sec





### 3.Acoustic-phonetic description of the regions (MOA and POA) :

MoA : Manner of Articulation

PoA : Position of Articulation

/l/ :Vocal folds of vibration .

daughter:(d/au/gh/t/e/r):Unaspirated voiced consonant followed by a diphthong(/ai/),followed by a velar constriction followed by front vowel (e) vocal fold vibration and turbulence at alveolar bridge (r) .

of(/o/f):Back vowel followed by a fricative (f) , which is forced through upper teeth and lower lip (labiodental) .

Padmaja(Pa/da)/ma/ja): A slightly elongated unaspirated bilabial followed by an middle vowel , then an unaspirated consonant (dh) and a nasal constraint (m) followed by voiced unaspirated palatal consonant.

NOTE : Here, da is almost pronounced like 'dh' .

### 4. Time Varying System Characteristics :

/l/:Tongue hump at front position of the vocal tract system and narrow opening of oral cavity.

/a/:Tongue hump at central position of the vocal tract system and wide opening of oral cavity

/m/:Opening of velum and closure at the lips are the system characteristics .

/d/: Formed with the tongue touching or approaching the inner ridge of the gums of the upper front teeth.

/au/:Tongue hump is high and it is in back position of the VT system, VT system is narrowly open and cylindrical in shape

/gh/:Complete closure at velum

/t/:Complete closure at dental

/e/:Tongue hump is medium and it is in front position of the VT system, VT system is moderately open

/r/:Partial closure of VT with tongue tip at alveolar ridge (Semi vowel)

/o/:Tongue hump is medium and it is in back position of the VT system, VT system is moderately open and cylindrical in shape

/f/:Complete closure at velum and opening of the nasal cavity.

/P/:Complete closure at velum and opening of the nasal cavity.

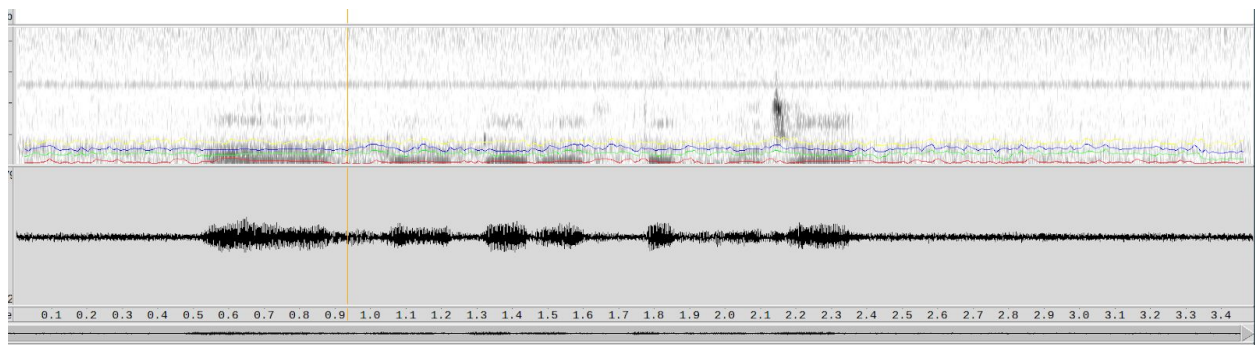
/a/:Tongue hump is low and it is in central position of the vocal tract (VT) system, VT system is widely open

/d/:Release of dental constriction and vocal folds vibration

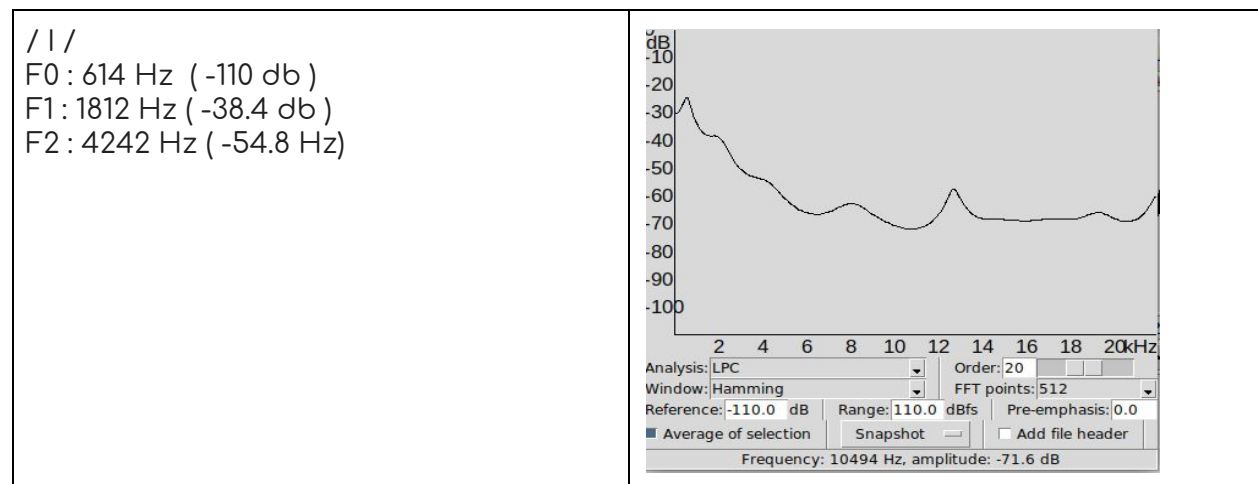
/ma/:Vocal folds vibration, velum is lowered and constriction at lips (little elongated)

/ja/:Release of palatal constriction and vocal folds vibration

5 . After analysing the spectrogram of the waveform (no DC component ,and formants marked ) , we can observe individual phone analysis .

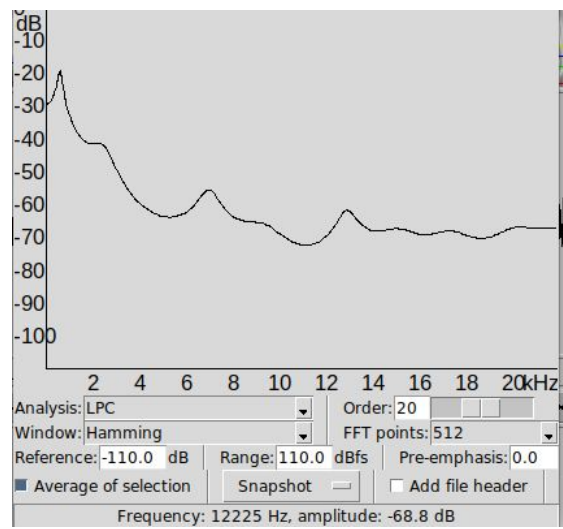


For each syllable , the LP analysis of the specific time duration will be provided beside such that understanding formants and amplitude will be easier :



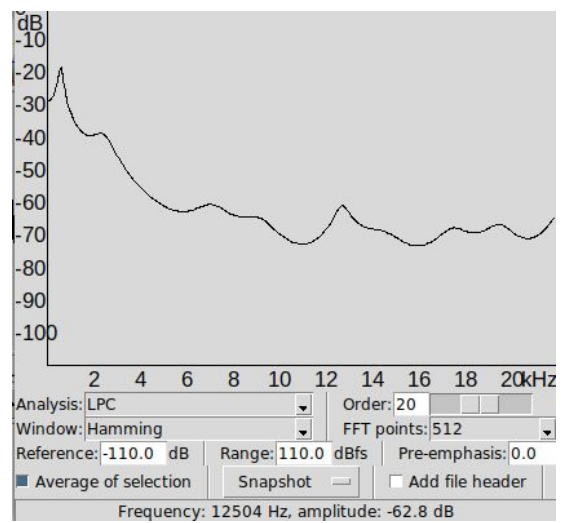
/a/

F0 : 669 Hz , -19.1 db  
F1 : 6992 Hz , -56.2 db  
F2 : 12839 Hz , -63.1 db



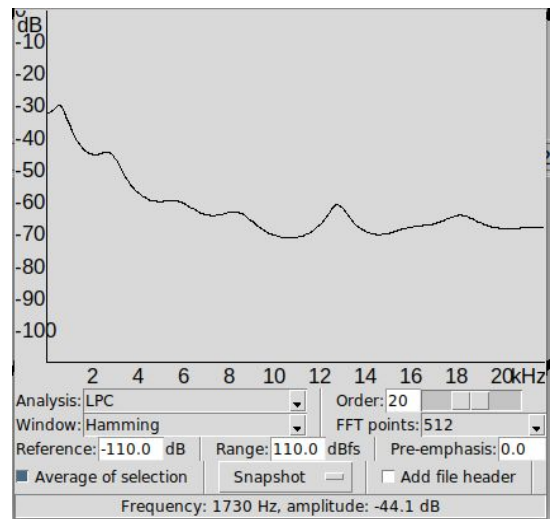
/m/

F0 : 558 Hz , -18.3 db  
F1 : 13005 Hz , -60.9 db



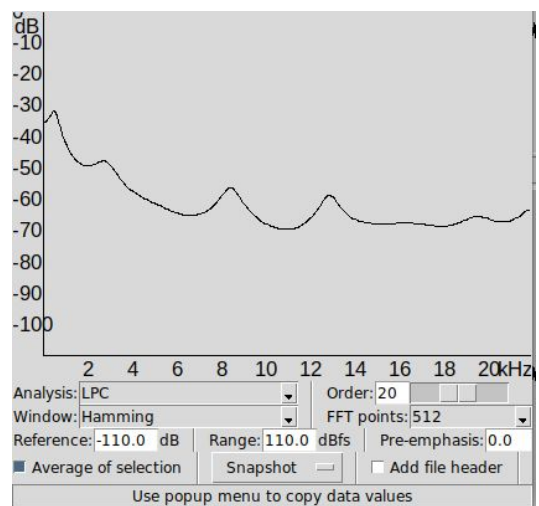
/D/

F0 : 502 Hz , -30.1 db  
F1 : 12839 Hz , -60.0 db  
F2 : 18254 Hz , -64.1 db



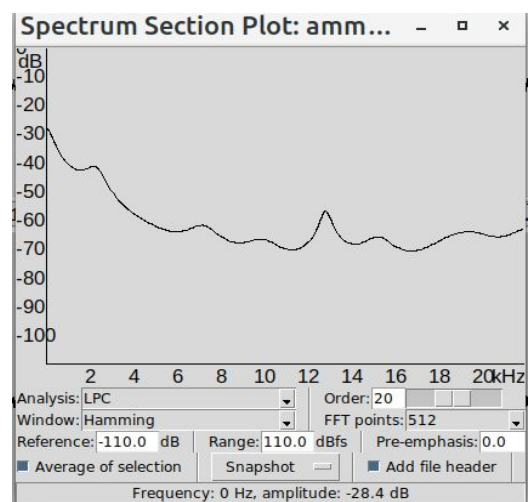
/au/

F0 : 558 Hz , -16.4 db  
F1 : 2567 Hz , -43.0 db  
F2 : 7089 Hz , -60.3 db



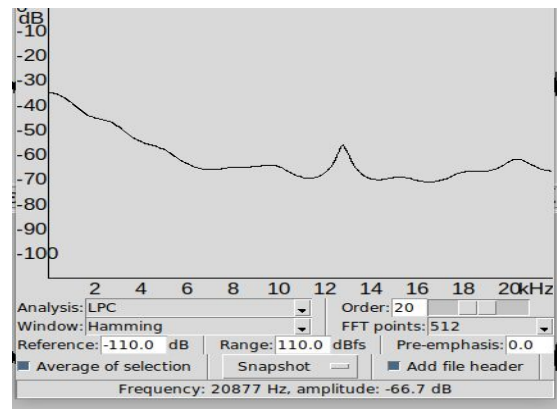
/of/

F0 : 2334 Hz , -41.4 db  
F1 : 12839 Hz , -56.7 db



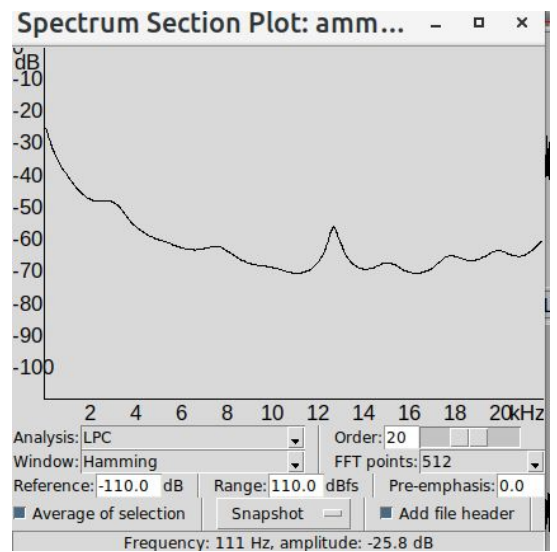
/pa/

F0 : 167 Hz , -34.1db  
F1:12950 Hz , -23.7 db



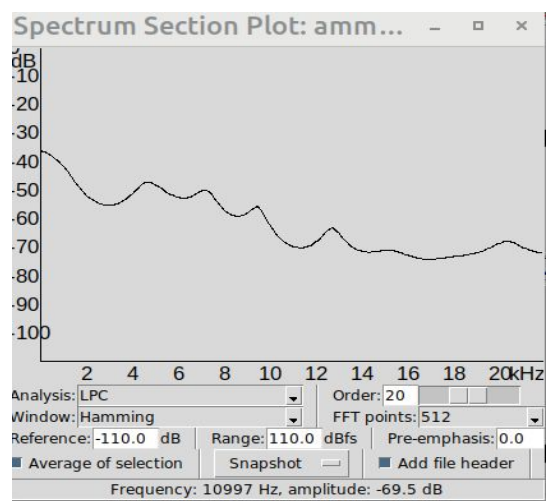
/da/

F0 : 111 Hz , - 25 .8 db  
F1: 7647 Hz , -62.3 db



/ja/

F0 : 4800Hz , -47.1 db  
F1: 7863 Hz , -50.0 db  
F2: 9489 Hz , -55 db



Q5. Record your native place which should be as "I am from < native place >".

Ans : Here is the link to the voice sample , stating the following line :

"I am from Vijayawada"

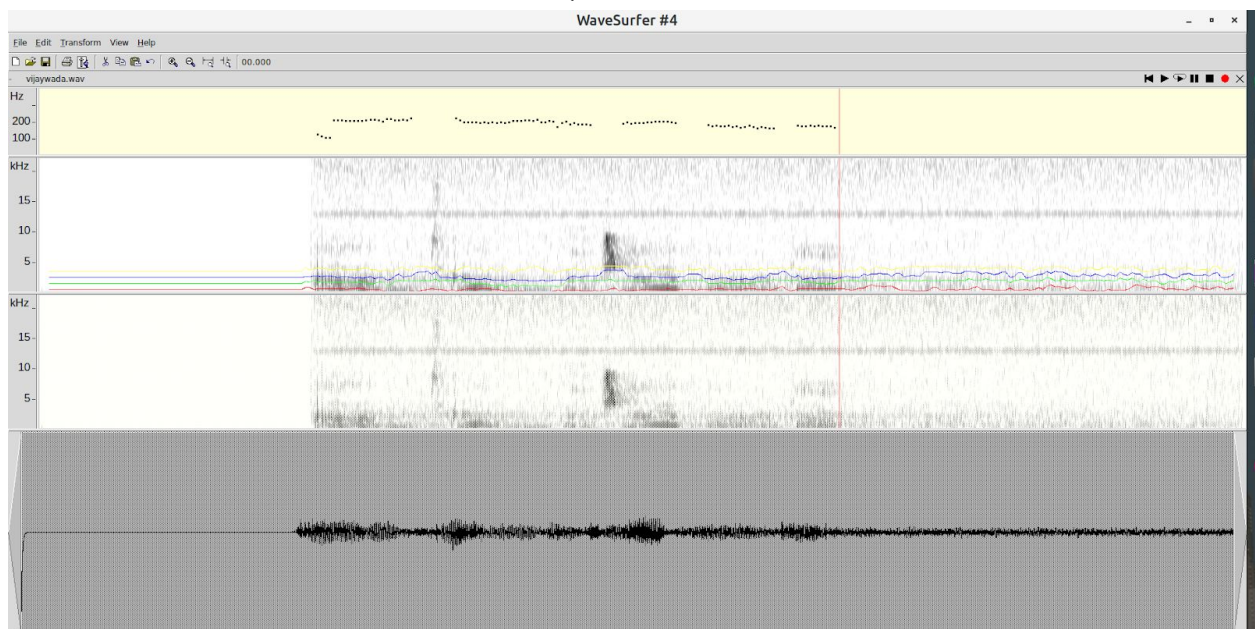
Sample Link :

<https://drive.google.com/file/d/1EPRnJMppHnYpjvinqyRZqXO1zmE2dzJI/view?usp=sharing>

TextGrid Annotation :

<https://drive.google.com/file/d/1dsb7VRhikVt5iHggGc8ADIMCu0zrwiWn/view?usp=sharing>

Q.1 The pane above the waveform is the spectrogram , and DC component was removed from the waveform , to analyse better .



Spectrogram for the voice sample



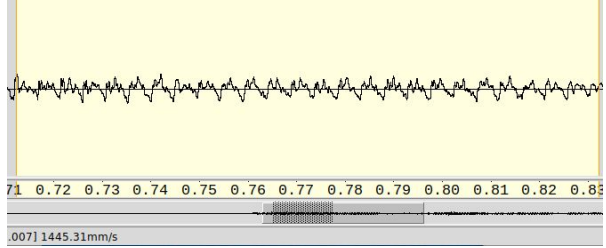
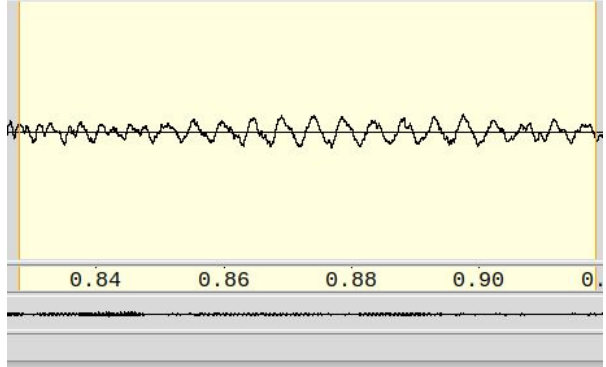
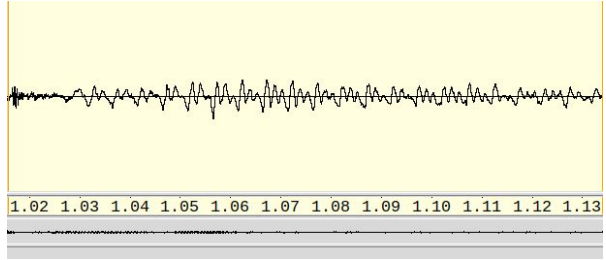
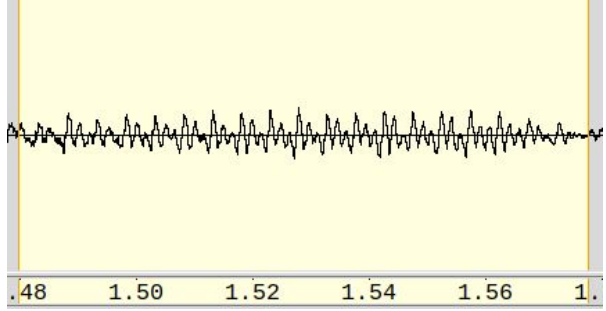
2.

Sentence : l / a-m / f-r-o-m/ /Vi-ja-ya-wa-da/

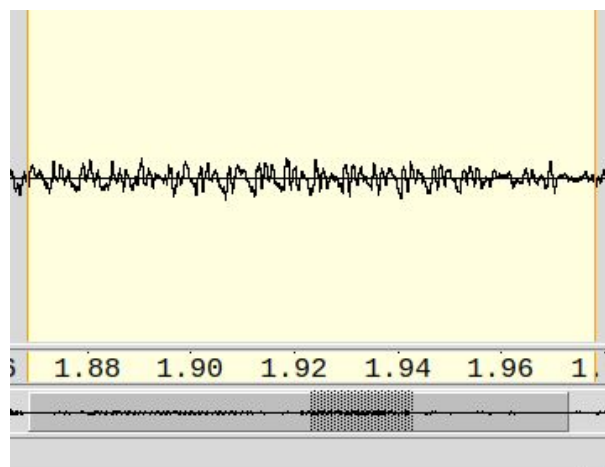
Note : No plosives could be identified ( da is a plosive , but it's unvoiced and subtle)

Voiced Regions :

For the above sentence , the voiced regions will be : /da/ , /m/,/j/ , /r/ , /v/ , / l/ ,etc

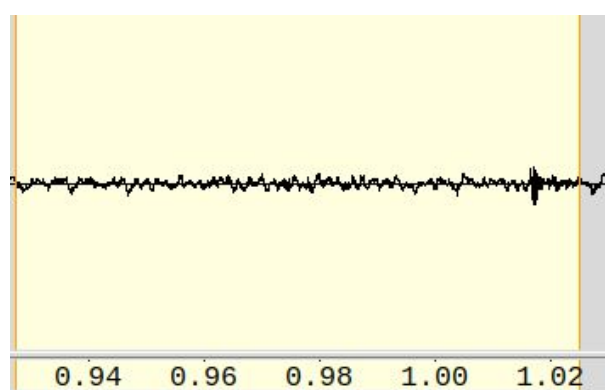
/l/ ( t= 0.706 to t=.830)	
/m/ (t=0.83 to t = 0.91 sec)	
/r/ (t=1.02 to to t=1.13)	
/ya/	

/do/ (t=1.86 to t=1.98 sec)

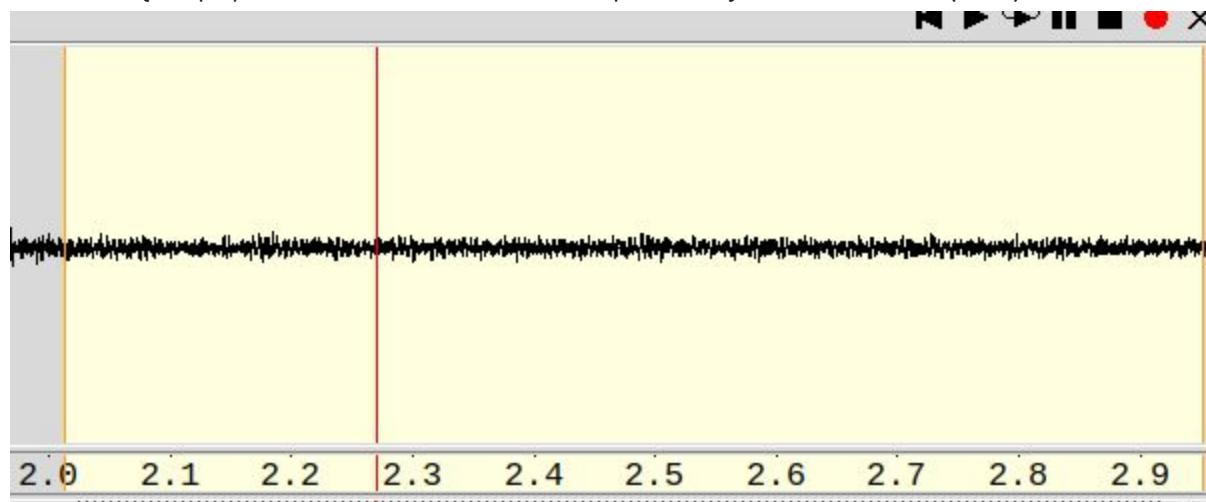


Unvoiced Regions : / f/

/f/ (t=0.94 to t=1.02 secs)



Silence : {Empty Parts , with almost 0 amplitude } t=2.0 to t=3.0 (end)



### 3. Manner & Place of Articulation :

/l/ - Vocal folds vibration .

/a-m/ -Vocal folds vibration accompanied with nasal constriction .

/f-r-o-m/- It starts with a fricative sound, where air is forced through the upper teeth and lower lip (labiodental) , followed by (r) narrow opening at the alveolar bridge and then vocal fold vibration with narrowing of the front lips and then with a nasal sound along with lowered velum & closing of lips .

/Vi-ja-ya-wa-da/ - Starts with an elongated bilabial approximant followed by an unaspirated note and release of palatal constriction , and the semi - vowel (y ) and then by an unaspirated voiced with release of alveolar constriction .

### 4. Time Varying System Characteristics :

Describing time varying systems characteristics consists of specifying positions of different articulators and shape of the vocal tract while producing the particular sound unit .

/l/ -Tongue hump at front position of the vocal tract system and narrow opening of the oral cavity.

/a-m/-Tongue hump at the central position of the vocal tract system and wide opening of oral cavity and opening of velum and closure at the lips are the system characteristics .

/from/ :

/f/:(Fricative) Complete closure at velum and opening of the nasal cavity.

/r/:(Semi vowel) Partial closure of VT with tongue tip at alveolar ridge (Semi vowel)

/o/:(Semi vowel) Tongue lies in back position of the oral cavity , and

/m/:(Semi vowel) Partial closure of VT with tongue tip at alveolar ridge (Semi vowel)

/Vijayawada/:

/V/ : It's a voiced fricative labiodental consonant, the lower lip approaches or touches the upper teeth.

/j-a/: Here , the body of the tongue approaches or touches the hard palate of the oral cavity and creates a turbulence in the air stream . It's an approximant , a sound that is produced by bringing one articulator in the vocal tract close to another without, however, causing audible friction . The VT system is moderately open .

/y-a/: Here , the body of the tongue approaches or touches the hard palate of the oral cavity and creates a turbulence in the air stream . It's an

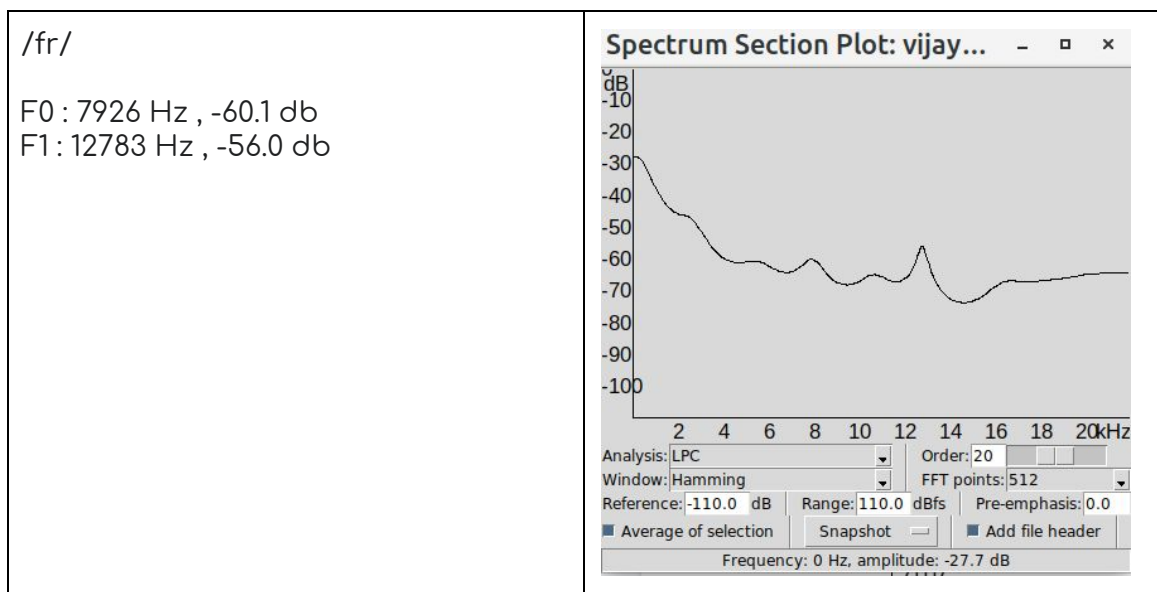
approximant , a sound that is produced by bringing one articulator in the vocal tract close to another without, however, causing audible friction . The VT system is moderately open .

/w-a/: One of the two constrictions that form a 'w' is a bilabial approximant. The other is a velar approximant: the tongue body approaches the soft palate.

/d-a/: In an alveolar consonant, the tongue tip (or less often the tongue blade) approaches or touches the alveolar ridge, the ridge immediately behind the upper teeth. VT is narrowly open .

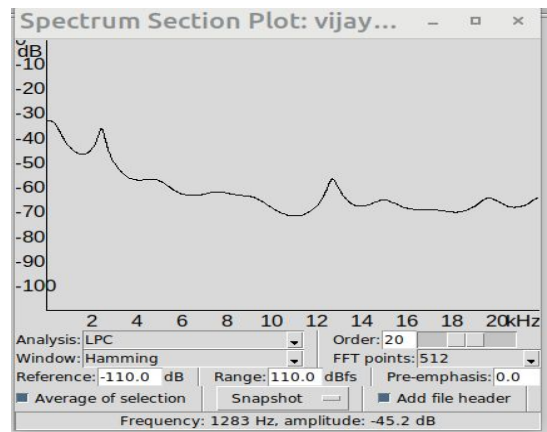
5.

- Spectral Analysis of each component :
  - The spectral analysis for { l , am } will be close to the above Q.4 , 5th part , as the speaker and background settings do not differ drastically.
  - Pls note : Some parts are heavily overlapped , hence separating them was difficult , only the major ones are picked out here . (expect /ya/)



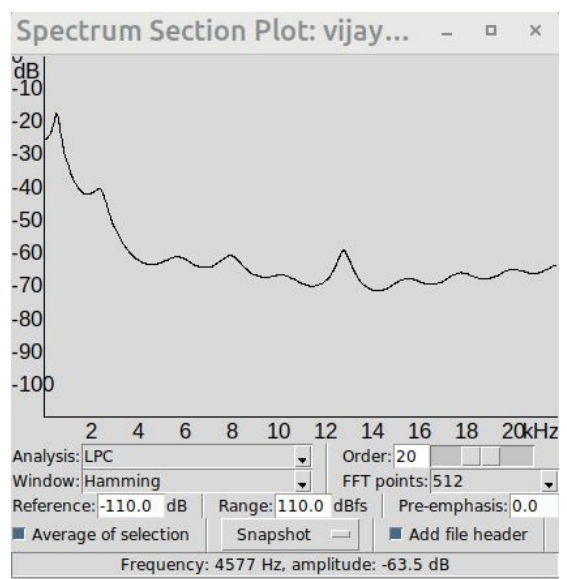
/ o / { Very slight duration }

F0 : 2512 Hz , -36.6 db  
F1 : 12761 Hz , -57.2 db



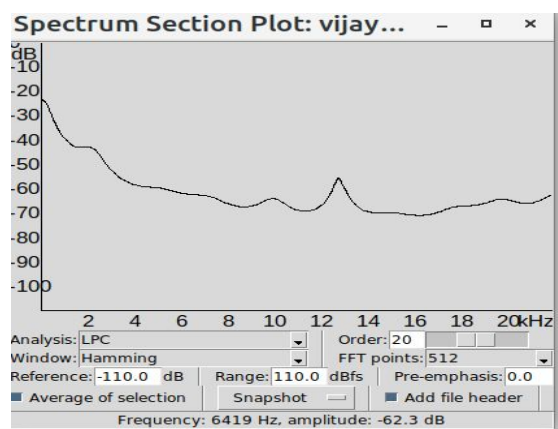
/m/

F0 : 614 Hz , -18.0 db  
F1 : 2456 Hz , -40.4 db  
F2 : 8038 Hz , -60.9 db



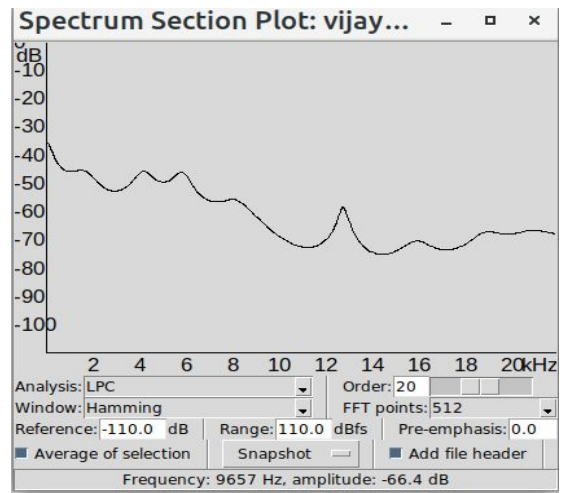
/Vi/

F1 : 12048 Hz , -60 db  
F0 : 10100 Hz , -64.4 db



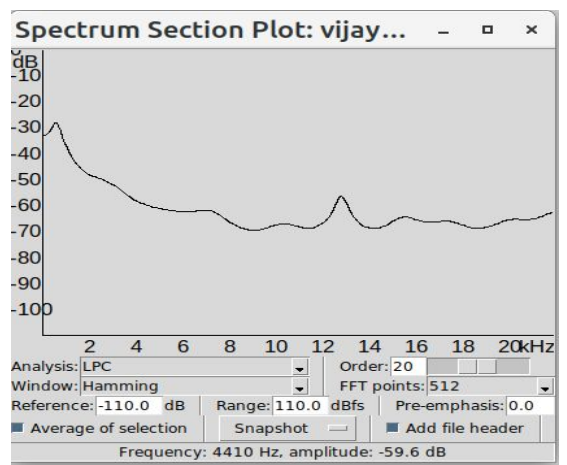
/Ja/

F0 : 1451 Hz , -40.3 db  
F1: 4186 Hz , -46.4 db  
F2: 12389 Hz , -59.9 db



/Wa/

F0 : 669 Hz , -28.3 db  
F1: 12836 Hz , -56.2 db



/Da/

F0 : 12839 Hz , -56.3 db  
(Most significant)

