Assignment -1 Report

Name: Bhaskar Bhatt
Roll No: 2022102003

2. Choose one speech application which is of recent interest, explain the speech application in detail and highlight the linguistics part in the application.

Introduction to Prosody: Prosody is what makes our speech sound natural and expressive. It includes the way we use intonation, stress, and rhythm to convey emotions and meaning, adding a human touch to our words.

Application in Speech Synthesis: In speech synthesis, prosody helps computer-generated voices sound more lifelike. By mimicking natural intonation and stress, these systems can speak in a way that feels more engaging and clear.

Application in Speech Recognition: For speech recognition, prosody helps computers understand the subtleties of how we speak. It allows them to pick up on nuances like emphasis and pitch changes, which are crucial for accurately interpreting what we mean.

Linguistic Perspective: Linguistically, prosody is linked to the structure and meaning of language. Different languages have unique prosodic patterns, and knowing these helps create more effective speech technologies.

Conclusion: Prosody makes a big difference in how speech technologies interact with us. It turns robotic voices into ones that feel more human and improves how well systems understand our spoken words.

3. Let s(t) = sin(2πf1 t); 0 ≤ t ≤ τ & sin(2πf2 t); τ < t ≤ 1 Else 0, where τ = 0.5sec; f1 = 100Hz; f2 = 200Hz, then empirically find the τ considering amplitude spectrum of STFT with window size of 20ms.

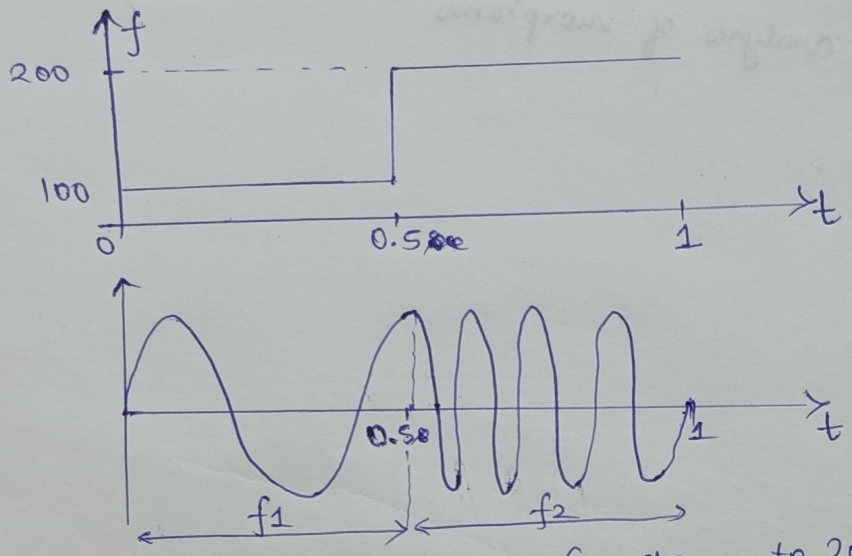
3) $s(t) = \sin(2\pi f_1 t)$ & $\sin(2\pi f_2 t)$

$\tau = 0.5s$, $f_1 = 100Hz$ & $f_2 = 200Hz$

$$s(t) = \begin{cases} \sin(2\pi f_1 t) & [0, \tau] \\ \sin(2\pi f_2 t) & (\tau, 1] \\ 0 & \text{otherwise} \end{cases}$$

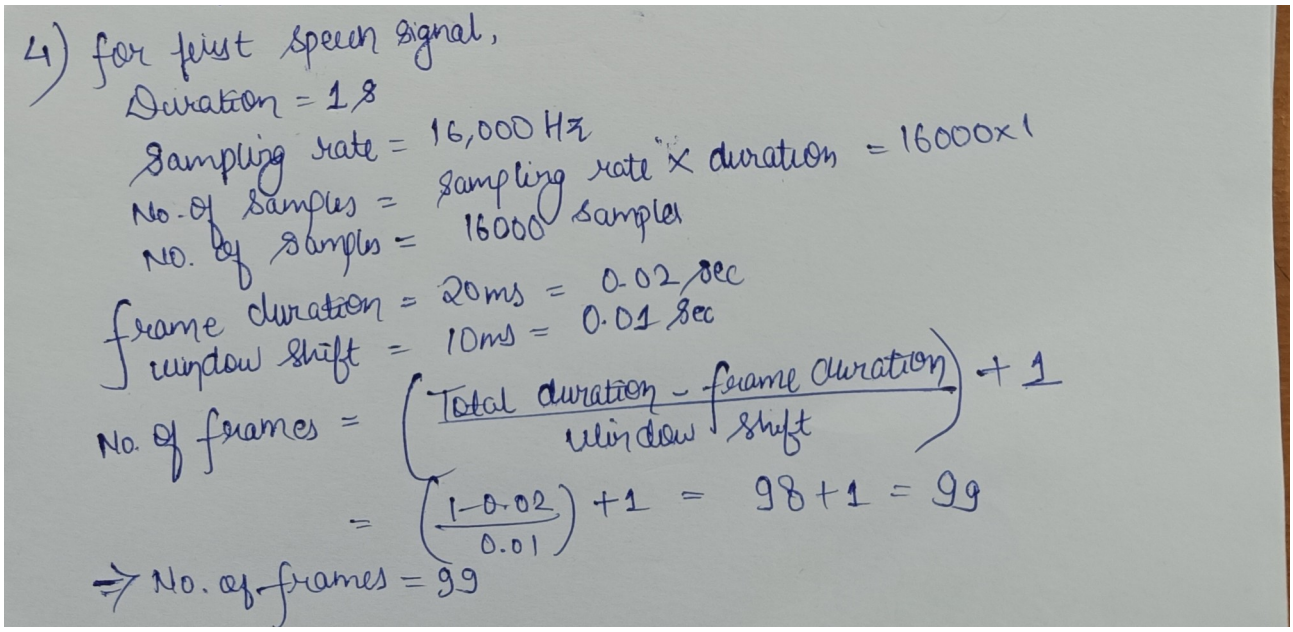
We compute the STFT and analyze the amplitude spectrum of the STFT.

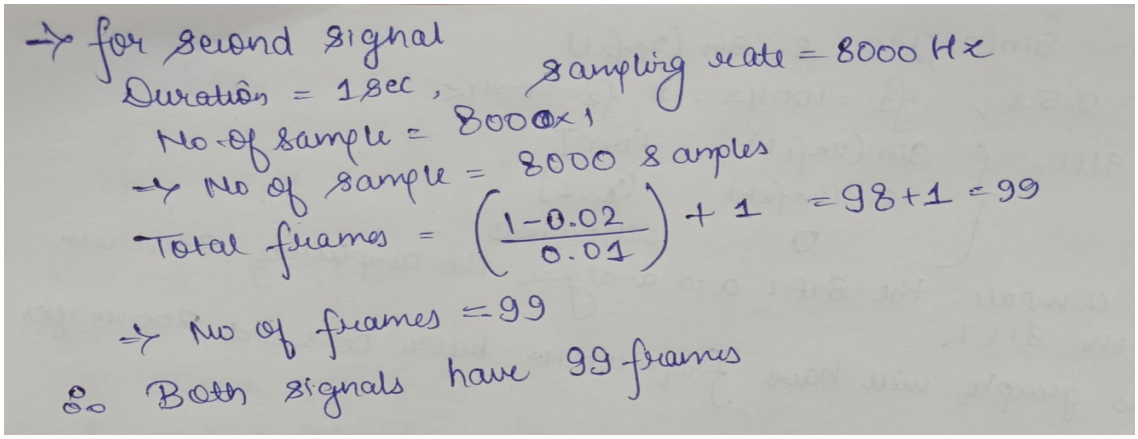
The graph will have $f_{req} = 100$ Hz before 0.5s and 200 Hz after that



The time when dominant freq changes to 200 from 100 is the optimal value of τ.

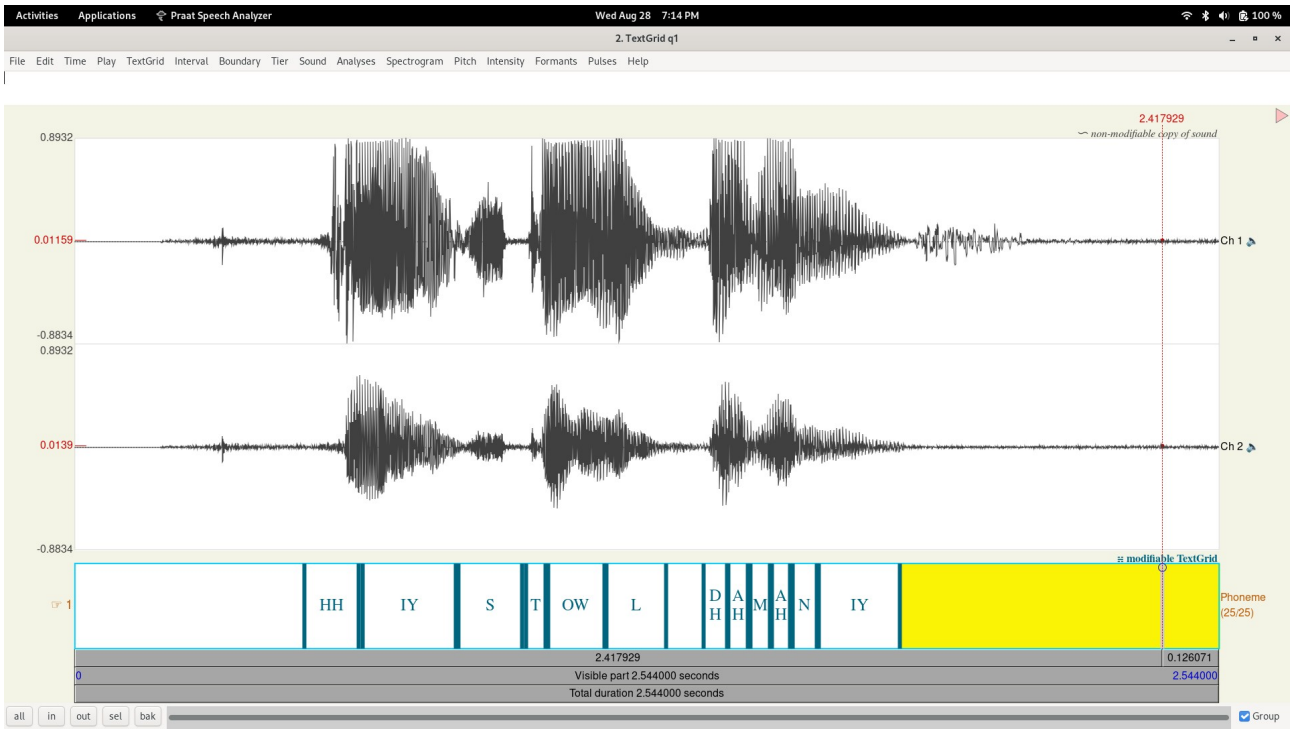
4. Suppose there are two speech signal, each with a duration of 1 second. One speech has sampled at a sampling rate of 16,000 Hz, and the other at a sampling rate of 8,000 Hz. How many samples are there in each speech signal. If we consider a frame duration of 20 ms with window shift of 10 ms, how many frames will be obtained for each speech signal?



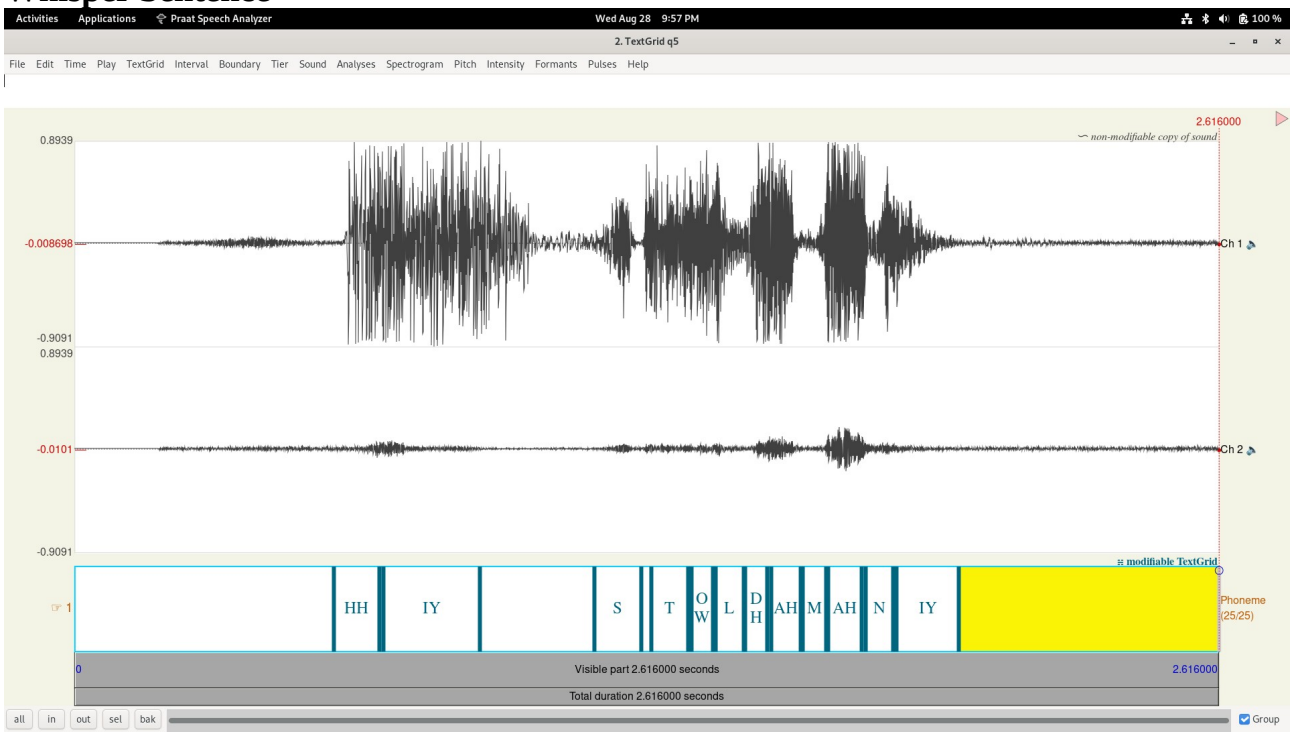


5. Consider the same sentence in Q1 and Record it in whispering and mark the boundaries of phonemes. Highlight and discuss the below based on normal and whisper recordings

Raw Sentence



Whisper Sentence



a) What has not changed from normal to whisper? Is the message preserved in both cases?

The core linguistic elements, such as the sequence of phonemes, are maintained in both normal and whispered speech because the articulatory gestures that create phonemes are largely unchanged. The intended message remains intact, as whispering still communicates the same phonemic and syllabic information, enabling listeners to grasp the spoken content. However, the lack of vocal fold vibration might result in reduced clarity, particularly for phonemes that depend on voicing.

b) What new information is added to whisper to become normal? Is the Voice/speaker information preserved?

In regular speech, the pitch comes from the vibrations of the vocal cords, which aren't present when whispering. This pitch helps communicate things like tone, emotions, and who's speaking. In whispers, a lot of this personal information is missing. Still, some clues about the speaker, such as their speaking style, how they articulate words, and even certain vocal traits, can give hints about their identity.

c) What is the vocal cords state while whispering? What are the other states of vocal cords while speaking normally?

In whispering, the vocal cords are positioned close together but do not vibrate. Instead, air flows through a small gap, creating a turbulent noise without any vocal cord vibrations.

In regular speech, the vocal cord states include:

- **Voiced Sounds:** The vocal cords come together and vibrate, producing sound waves that create pitch.
- **Unvoiced Sounds:** The vocal cords stay apart, and no vibrations occur. Sound is generated by the turbulent airflow through other parts of the vocal tract.
- **Breathy Voice:** The vocal cords are slightly open and vibrate loosely, resulting in a breathy or airy sound.
- **Glottal Stop:** The vocal cords are fully closed, briefly halting the airflow.

d) Can all the words in the English vocabulary be discriminable while whispering? How about "pig" vs. "big"? Why might this not be discriminable?

When whispering, it becomes difficult to distinguish between words that depend on voicing differences, such as "pig" and "big." This is because the main difference in these words is the voicing of the first consonant, which is absent in whispered speech.
For "Pig" versus "Big," the voiced "b" and the unvoiced "p" both lose their voicing when whispered. Without the pitch and voicing information, these sounds start to sound alike.

In normal speech, a spectrogram would show clear differences in the voicing of these consonants. In whispered speech, these differences become less pronounced, making it harder to tell the words apart.

The absence of voicing in whispering leads to similar patterns in the spectrograms of minimal pairs, resulting in fewer distinctions between them.

How can a mimicry artist imitate another speaker's voice?

Mimicry artists frequently adjust their vocal tract to match the formants of the person they are imitating. This involves altering the shape of the mouth, lips, and tongue to align with the specific resonant frequencies of the target speaker.

To effectively imitate someone, it’s important to match their pitch, rhythm, and intonation. This requires modifying the tension of the vocal cords and managing breath control.

These artists also pay close attention to the speaking habits of the person they are mimicking, such as their speaking pace, pauses, and other stylistic features.

Additionally, mimicry may involve changing the shape of the nasal cavity, pharynx, and other parts of the vocal tract to replicate the unique qualities of the target voice. By carefully mimicking these acoustic elements, an artist can produce a voice that closely resembles that of another person, despite having a different physical vocal structure.

6. For each of the following sentences, identify the most relevant area of linguistic study (Pragmatics/Discourse, Phonology, Morphology, Phonetics, Syntax or Semantics) and briefly explain your choice.

(a) The word "happiness" is composed of the root "happy" and the suffix "-ness."

- **Area:** Morphology
- **Explanation:** Morphology studies the structure of words and their formation from smaller units known as morphemes. This example illustrates how "happiness" is formed using a root and a suffix, a fundamental concept in morphology.

(b) The word "bank" can signify either the side of a river or a financial institution.

- **Area:** Semantics
- **Explanation:** Semantics focuses on meaning in language. This example shows polysemy, where a single word has multiple, related meanings, which is a key aspect of semantics.

(c) The sentence "She quickly runs." adheres to the standard subject-verb-object structure.

- **Area:** Syntax
- **Explanation:** Syntax examines the rules that govern sentence structure. This example highlights the grammatical arrangement of a sentence, a primary concern of syntax.

(d) The words "cat" and "bat" differ by only one sound.

- **Area:** Phonology
- **Explanation:** Phonology analyzes sound patterns within a language. This example compares two words that differ by just one phoneme, a central concept in phonology.

(e) When someone says, "It's cold in here," they might be implying that you should close the window.

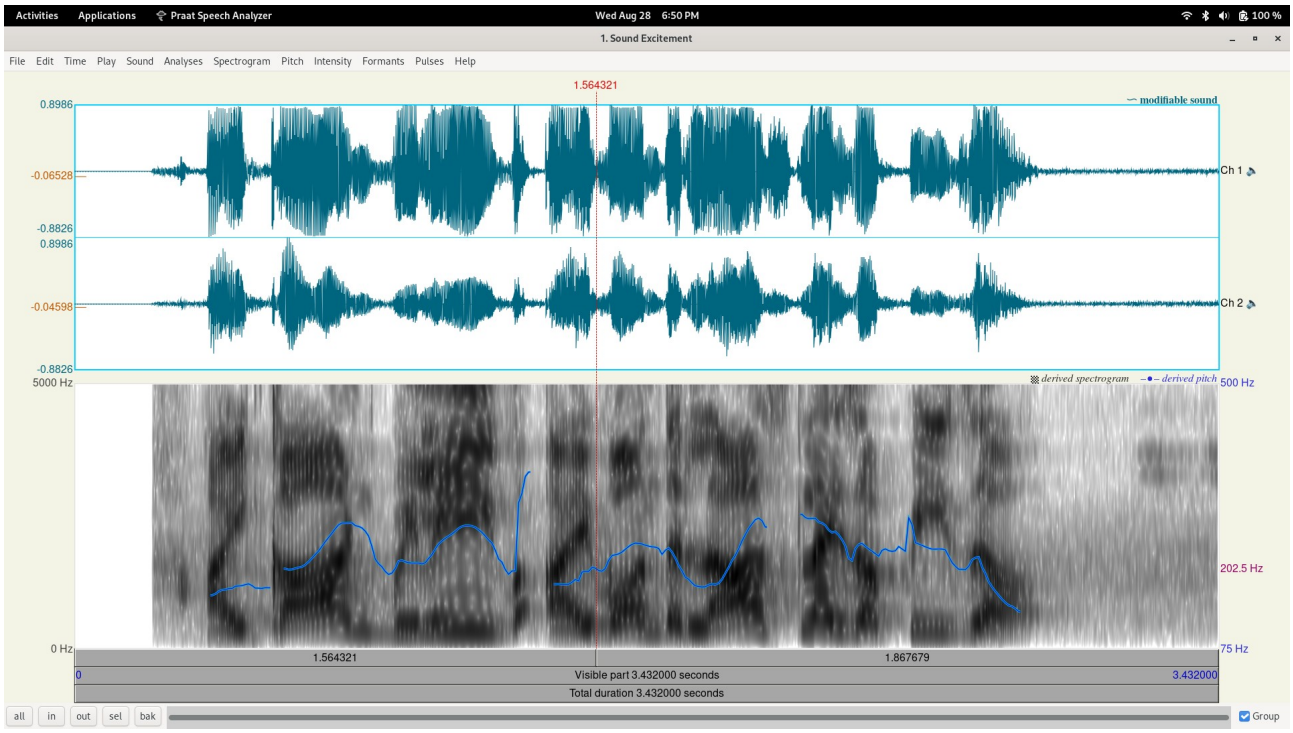
- **Area:** Pragmatics/Discourse
- **Explanation:** Pragmatics explores how context affects meaning. This example demonstrates how a simple statement can carry an implied meaning or request, a crucial idea in pragmatics.

(f) The sound [t] in "top" is articulated by placing the tongue behind the upper front teeth.

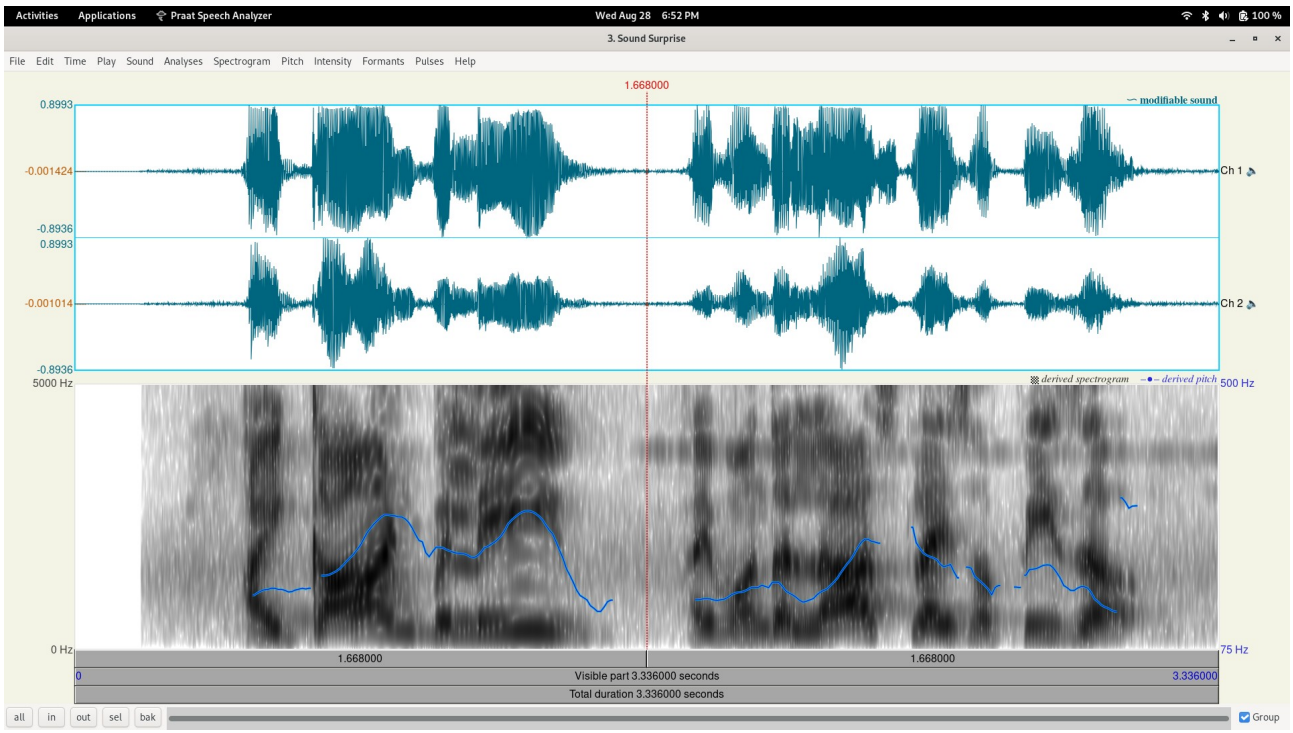
- **Area:** Phonetics
- **Explanation:** Phonetics deals with the physical properties and production of speech sounds. This example describes how a specific sound is articulated, which is a fundamental aspect of phonetics.

7. Record yourself saying the following sentence: ”I can’t believe I won the last slice of pizza!” with four different emotions of your choice, extract pitch contour using praat software and comment on the pitch variation for each.

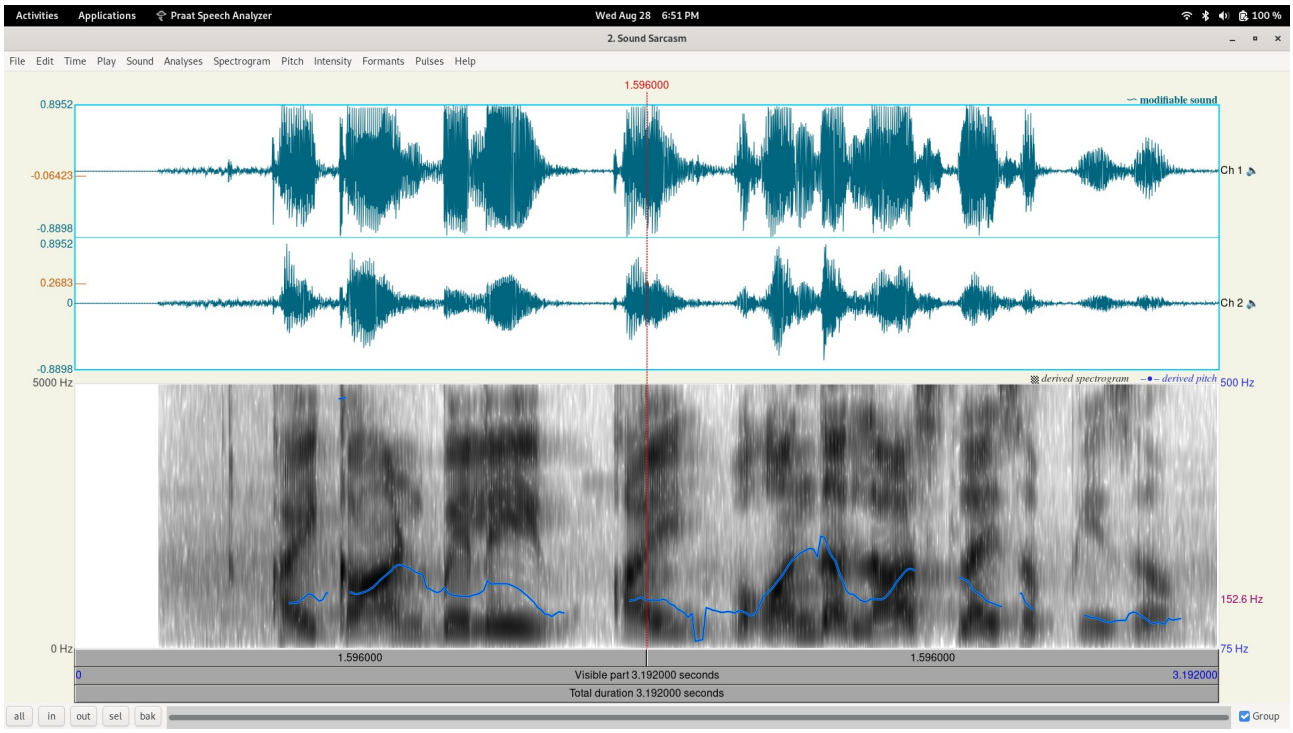
Excitement



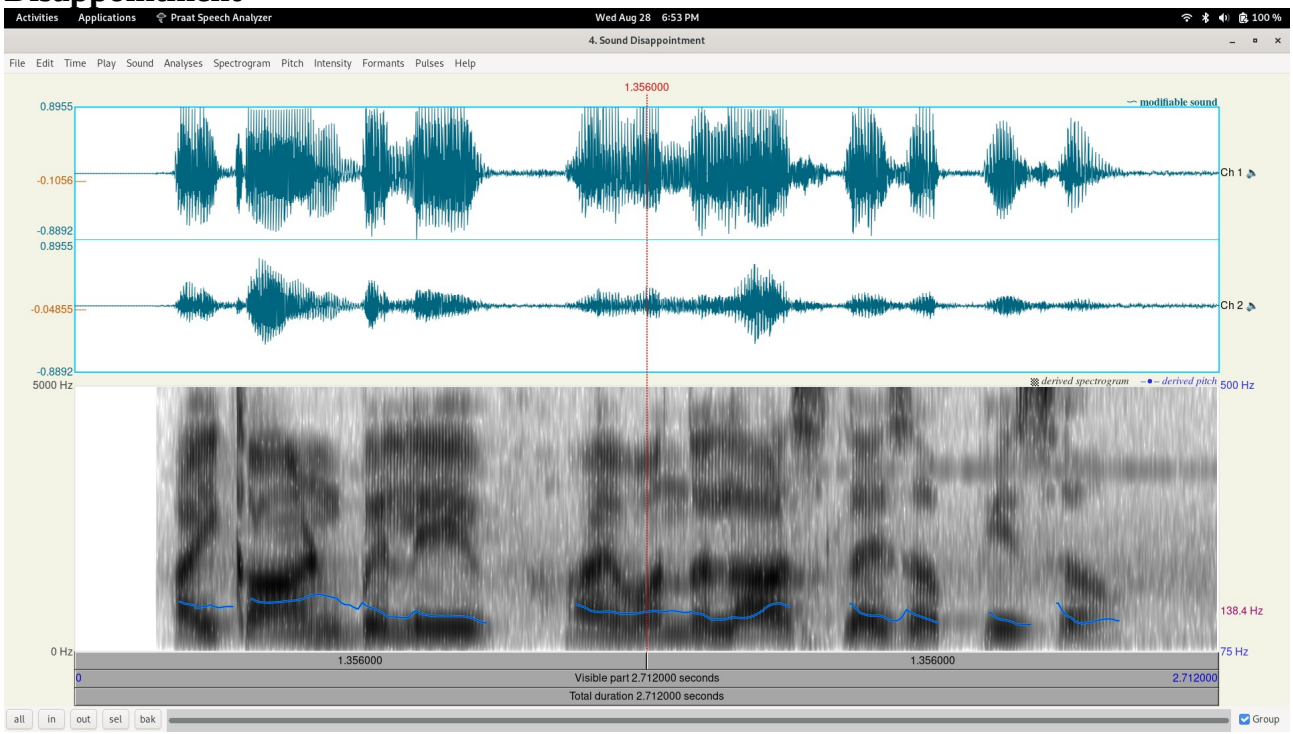
Surprise



Sarcasm



Disappointment



1. **Excitement**:
    - **Pitch pattern**: Upward rise followed by a slight drop
    - **Notes:** The pitch begins high and sharply rises on "can't" and "won," reflecting excitement. It then slightly dips towards the sentence's end. The pitch spans a broad range, typically from 200-350 Hz in male voices or 300-450 Hz in female voices.
2. **Surprise**:
    - **Pitch pattern**: Rapid increase followed by a slow decrease
    - **Notes**: The pitch starts low and spikes dramatically on "can't," peaking there. It then gradually lowers through the rest of the sentence. The pitch covers a wide range, often with a difference of 150-200 Hz between the lowest and highest points.
3. **Sarcasm**:
    - **Pitch pattern**: Overemphasized rises and dips
    - **Notes**: The pitch pattern features exaggerated fluctuations, with noticeable peaks on "can't" and "last," followed by steep drops. The overall pitch range is narrower than in excitement or surprise, generally within a 100 Hz range, but with frequent shifts.
4. **Disappointment**:
    - **Pitch pattern**: Mostly flat with a slight downward slope
    - **Notes**: The pitch begins low and remains fairly level throughout, with a minor downward slope. There’s a slight rise on "won" for emphasis, but the range is narrow, typically within a 50-75 Hz span.

General observations:

- Excitement and surprise show the broadest pitch ranges and most significant variations.
- Sarcasm involves more frequent pitch shifts but within a more confined range.
- Disappointment exhibits the flattest pitch contour and the narrowest range.
- The word "can't" consistently displays noticeable pitch movement across all emotions, highlighting its role in expressing the intended emotion.