Cognitive Impairment Detection Using Audio Analysis and NLP

Kousik Kumar Barnwal
Roll No: 22112056
IIT Roorkee

April 2025

Contents

1 Introduction

Cognitive impairment affects an individual's ability to think, concentrate, and remember. Early detection is critical for timely intervention and support. In this MemoTag assignment, we aim to explore the possibility of identifying cognitive impairments through speech patterns.

We have analyzed **five audio samples**, focusing on speech-related features such as silence duration, pitch variation, and speech rate. These features may indirectly reflect cognitive load, hesitation, and memory recall abilities. By extracting interpretable features and applying unsupervised modeling, we uncover patterns that could differentiate between potentially impaired and non-impaired speakers.

2 Methodology

2.1 Data Collection

Five audio samples were collected from different friends, all speaking the same sentence. The data was anonymized to maintain privacy and ensure unbiased analysis. They recorded from a mobile phone as an .mp3 file, after that, we converted them to .wav files for further audio analysis.

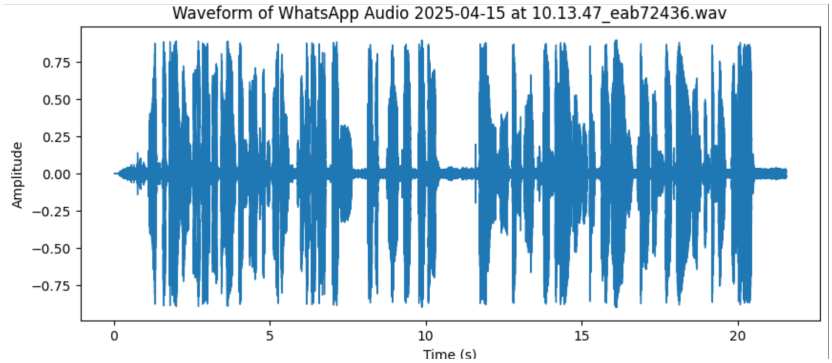

Figure 1: Waveform of an Audio Sample

2.2 Feature Extraction

1. Identification of Pauses:

- **Silent Parts Identification:** Silent regions in the audio files were detected by analyzing segments with little or no speech activity.
- **Silence Ratio:** The **silence ratio** was calculated as the proportion of silent segments relative to the total duration of the audio.

- **R2 Score Dependency:** The **R2 score** was computed to measure the dependency between the current silent segment and the previous non-silent segment, which helps to analyze how pauses are related to prior speech patterns.
- **Other features:** *Number of silent segments*, *Average Silent duration*, and *Maximum Silent duration* are also calculated for each audio file.

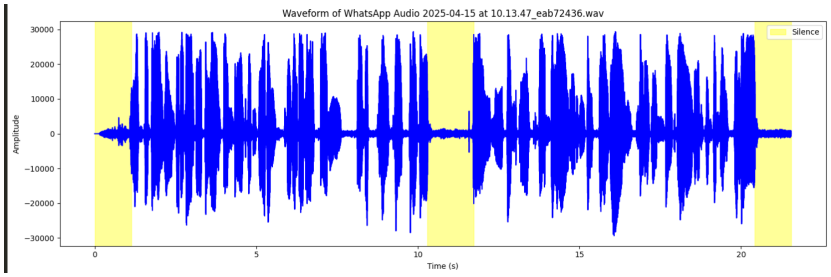

Figure 2: Silence segments in the waveform of the audio sample

2. **Pitch Variability:**
   - **Calculation Method:** Pitch variability was calculated using the `librosa.piptrack` library, which analyzes the frequency variations within the non-silent regions of the audio.
   - **Non-Silent Regions:** Only the non-silent segments of the audio were considered to evaluate the pitch variability, as these are more indicative of cognitive speech patterns.
3. **Speech Recognition and Evaluation :**
   - **Speech Recognition:** Speech recognition was performed using the in-built `SpeechRecognition` Google API, which transcribes the audio input into text.
   - **Recall Score:** The **Recall Score** measures the percentage of words in the original text that were correctly matched in the transcribed text. This includes both exact matches and semantic matches (synonyms), evaluated using WordNet for synonym detection.
   - **Substitution Score:** The **Substitution Score** quantifies how many words from the original text were incorrectly substituted or replaced in the transcribed text. This metric helps identify errors in the recognition process, either through incorrect words or a lack of synonym matching.
   - **Speech Rate:** The **Speech Rate** was calculated by dividing the total number of words spoken by the total duration of the non-silent segments. This metric helps assess the pace at which speech is delivered, which can be indicative of cognitive load or emotional state during the speech.

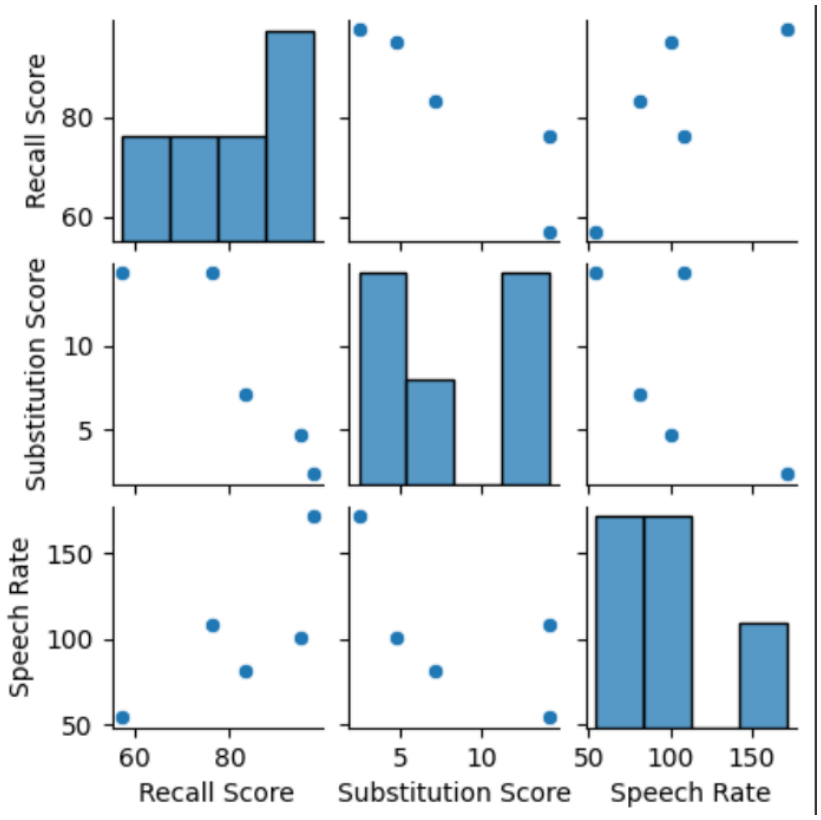

Figure 3: Pairplot of the speech recognition features

You can say here that as the Recall score increases, the substitution score decreases. The speaker is more confident. Also, speech rate increases with the recall score, which indicate the same thing.

2.3 Modelling

The model used for clustering was **Hierarchical Clustering** with *Ward linkage*, which was chosen because it does not require the number of clusters to be predefined. This unsupervised approach progressively merges similar samples and generates a dendrogram, allowing for flexible identification of natural groupings based on feature similarities.

3 Results

We can say from the above dendrogram that files 3, 0, and 2 are very similar to each other, while files 1 and 4 are similar to each other. Basically, there are

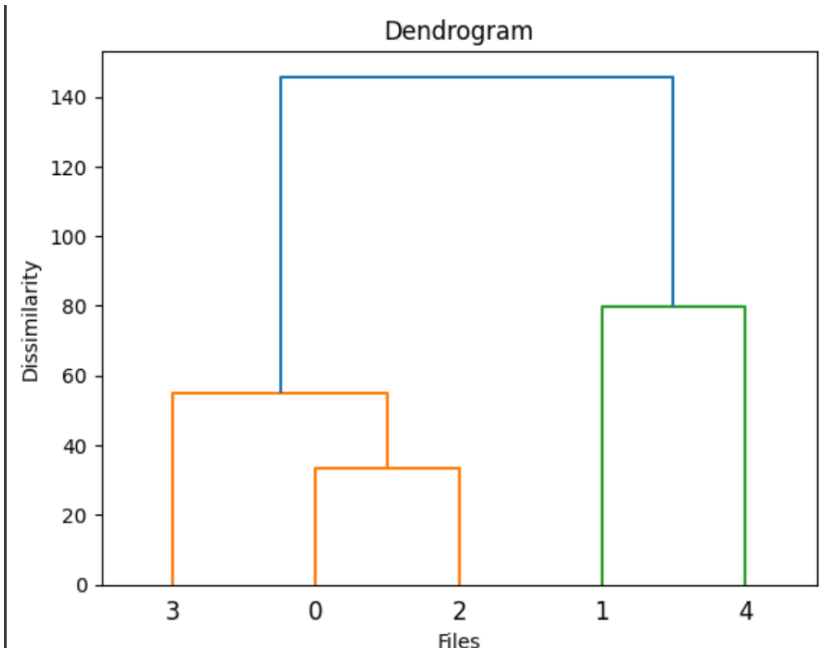

Figure 4: Dendrogram of the datapoints

2 clusters. To determine which clusters are abnormal or at risk, we can get the average of each property of each cluster.

Feature	**Cluster 1**	**Cluster 2**
total_audio_duration	27.5	23.81
silence_ratio	0.27	0.31
num_silence_segments	5	6.5
avg_silence_duration	1.5	1.39
max_silence_duration	2.2	1.76
R2_Pause_Dependency	0.64	0.48
pitch_variation_pause	881.67	954.03
Recall Score	72.22	96.43
Substitution Score	11.91	3.57
Speech Rate	81.59	135.79

Table 1: Comparative analysis of the 2 clustered groups

We can conclude from the above table that cluster 2 performs better than cluster 1 in terms of recall score, word substitution, and speech rate. The patients associated with cluster 2 are files 0, 2, and 3. Other features show comparable values for both clusters.

4 Improvements

To enhance the robustness of the model and make it clinically applicable, several improvements can be considered:

- **Increase Dataset Size:** Currently, only 5 audio samples are used. A larger and more diverse dataset with multiple patients will help improve the model's accuracy and generalizability.
- **Advanced Speech Recognition Systems:** We can implement more advanced speech recognition systems, or even develop a custom system, to better detect hesitations, pauses, and other vocal indicators of cognitive impairment. This will improve the model's ability to recognize subtle signs that are crucial for diagnosis.
- **Real-time Evaluation:** Integrating real-time speech analysis for evaluating cognitive impairments could make the model more applicable in clinical settings, where assessments need to be done quickly.

By implementing these improvements, we can make the system more clinically robust and reliable for detecting cognitive impairments in patients.