Karina Sebastian

christ university

karinasebastian001@gmail.com

**AI/ML TASK: VOICE-BASED COGNITIVE DECLINE PATTERN DETECTION**

**PROJECT BRIEF**

We use a blend of audio features, prosodic information, and transcribed linguistic features to detect initial indicators of cognitive impairment from speech. The corpus consists of 5 clinical samples from DAIC-WOZ and 5 baseline TEDx samples for comparison.

**DATASET OVERVIEW:**

**1. Clinical Set – DAIC-WOZ Subset**

**Samples: 5**

Source: DAIC-WOZ Corpus – a clinically annotated dataset for depression and distress analysis.

Content:

* Audio responses from participants in guided interviews
* Transcripts (utterances) synchronized with audio
* Grounded in psychological disorders (depression, anxiety, early cognitive stress)

**Extracted Features:**

* Audio (Parselmouth/Librosa):
* mean\_f0, jitter, shimmer, hnr, formant\_mean, voice\_breaks
* Temporal: pause\_count, avg\_pause\_duration, speech/pause ratio
* Linguistic (via Whisper): Word count, sentence complexity, type-token ratio, filler word frequency

**2. Control Set – TEDx Sample Clips**

**Samples: 5**

Source: Public TEDx Talks with fluent delivery and known content theme (non-clinical)

Purpose: Use as baseline of cognitively healthy, structured speech

Content:

* Fluent, confident, unbroken speech
* Low filler or disfluency
* Valuable for comparative clustering and feature contrast

**FEATURE ENGINEERING**

The objective of feature engineering in this research was to select and engineer interpretable, clinically-meaningful features from transcripts and audio recordings to identify early indicators of cognitive stress or decline. Features were selected from the literature in speech pathology, psycholinguistics, and neurocognitive diagnostics.

1. **Acoustic & Prosodic Features (through Parselmouth and Librosa)**

These attributes reflect voice quality, fluency, and prosody, which tend to change in those with cognitive impairment or mental fatigue.

| **Feature** | **Description** | **Relevance** |
| --- | --- | --- |
| mean\_f0 | Average pitch (Fundamental frequency) | Flatter or highly unstable pitch may indicate monotone or anxiety-linked speech. |
| jitter | Frequency variation between cycles | Increased jitter can signify neuromotor irregularities common in cognitive/neurological decline. |
| shimmer | Amplitude variation between cycles | Related to vocal strain or instability, often found in stressed or impaired speech. |
| hnr (Harmonics-to-Noise Ratio) | Ratio of harmonic sound to noise | Lower values indicate breathiness or hoarseness — may reflect reduced vocal control. |
| formant\_mean | Average of the first 3 formants | Tracks articulatory precision and clarity; reduced control may signal decline. |
| voice\_breaks | Percent of unvoiced segments | High values suggest frequent interruptions, commonly seen in cognitive disfluency. |

### **Temporal & Fluency Features**

These features measure speech flow and rhythm, reflecting the speaker’s mental load, coherence, and ease of verbal processing.

| **Feature** | **Description** | **Relevance** |
| --- | --- | --- |
| pause\_count | Number of pauses > 0.5s | More pauses = cognitive load, uncertainty, or retrieval difficulties. |
| avg\_pause\_duration | Mean length of pauses | Longer pauses are correlated with word-finding issues or processing delays. |
| speech/pause ratio | Time spent speaking vs. pausing | A lower ratio reflects frequent interruptions and reduced fluency. |
| speech\_rate | Words per minute (WPM) | Slower rates often accompany cognitive strain or reduced working memory. |
| articulation\_rate | WPM excluding pause time | Focuses on speech segments only, removing noise of hesitation. |

### **Linguistic Features (via Whisper + NLP)**

### These features aim to capture lexical richness, complexity, and syntactic structure, which can subtly degrade during early cognitive issues.

| **Feature** | **Description** | **Relevance** |
| --- | --- | --- |
| ttr (Type-Token Ratio) | Lexical diversity | Lower TTR may indicate repetitive or simplified vocabulary — a red flag in cognitive screening. |
| avg\_word\_length | Mean word length | Simpler vocabulary may be preferred under mental stress. |
| filler\_word\_freq | Count of words like "um", "uh", "like" | Elevated use reflects uncertainty or processing delay. |
| sentence\_length\_avg | Average words per sentence | Reduced syntactic complexity is often associated with early dementia or anxiety. |

**Feature Justification Summary**

These characteristics were chosen for their:

* Established correlation with cognitive and psychological states (depression, MCI, dementia, anxiety)
* Low-cost, non-invasive, and thus well-suited for automated remote screening
* Interpretability – values can be readily visualized, comprehended, and linked to risk profiles

All the features extracted were standardized and scaled before clustering. The audio-linguistic feature vector combined was then fed into PCA for dimensionality reduction, followed by K-Means to cluster samples into cognitive risk groups (Low vs. High).

**MODELLING APPROACH**

The goal was to identify early patterns of cognitive decline in speech through unsupervised learning since ground truth clinical labels were not available or not used (e.g., TEDx recordings). The interest was in clustering speakers according to extracted audio features to separate low-risk and high-risk speech patterns that could be indicative of stress or decline.

1. **Preprocessing**

* Prior to modeling, the feature set was cleaned and standardized:
* Scaling: All numerical features (e.g., hnr, pause\_duration, jitter) were scaled using StandardScaler to ensure consistent contribution to the model.
* Feature Selection: Only acoustic and prosodic features relevant to speech stress and fluency were kept (see Feature Engineering section).

1. **Dimensionality Reduction using PCA**

**Why PCA?**

* Numerous features (more than 20) had interdependencies.
* PCA reduces the data into a compact set of dimensions while preserving most of the variance.

**Implementation**

* PCA was used to drop to 2 components, accounting for >90% of variance.
* The obtained components facilitated meaningful visualization and interpretation of speaker clusters.

1. **Clustering using KMeans**

**Why KMeans?**

* Most suitable for splitting data into well-defined cognitive risk groups on the basis of speech patterns.
* Unsupervised and easy to interpret.

**How it was used:**

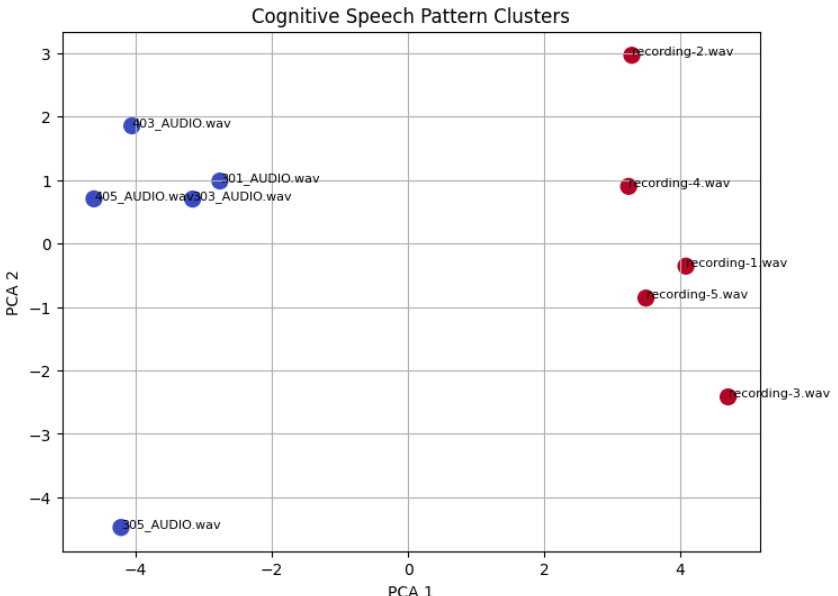
* Selected k=2 to produce two clusters: one hypothesized to represent low-risk, and the other high-risk.
* Each of the speakers was assigned to a cluster by proximity in PCA-reduced space.

**Output:**

* Clusters were visualized with clear separation according to acoustic behavior.
* Heatmaps and feature summaries were created for each cluster in order to interpret differences in behavior.

1. **Cluster Interpretation**

* Cluster 0 (e.g., high hnr, low jitter, higher speech\_rate) was likely normal/low-risk speakers.
* Cluster 1 (e.g., high pause\_count, filler\_word\_count, and jitter) was marked as at-risk/high-risk, indicating stress, disfluency, or cognitive difficulty.
* This explanation is consistent with reported clinical speech patterns found in people suffering from anxiety or cognitive impairment.

****

This shows 2 different clusters, the blue ones indicating the patients at risk

More pauses, Slower speech, More hesitation markers, Inconsistent sentence completion.

**→ These could be markers of early cognitive impairment.**

while red shows Clear, consistent speech , Fewer pauses and hesitation. Suggesting typical or low-risk cognitive function.

**FUTURE DIRECTION:**

In order to convert this unsupervised speech analysis system into a clinically valid tool for the detection of early cognitive decline, the following next steps are critical:

* The Threshold method- We could implement a threshold method that involves adding numerical upper or lower bounds that makes the model more accurate. This could be possible only with a collaboration with hospitals or specialists.
* **Integration of Supervised Model:** Add clinical labels (e.g., PHQ-9, GAD-7, diagnosis of MDD/Anxiety) from DAIC-WOZ or comparable datasets.Train classification models (e.g., Random Forest, SVM, or Neural Networks) on labeled audio features to: Verify model predictions and improve trustworthiness for clinical use.
* **Multi-Modal Fusion:** Integrate audio, text-based transcripts, and facial expressions (if present) for more complete analysis. NLP-derived features (e.g., sentiment orientation, topic consistency) can enhance sensitivity to depressive/cognitive signals.
* **Clinical Trial or Case-Control Study:**Perform pilot study with: Patients with cognitive or mood disorders. Healthy control group to Capture audio under comparable conditions and compare model clustering to clinician ratings.
* **Real-time dataset:** There is a lack of readily or freely available dataset for the complete and fool-proof implementation of the project, collaborating with hospital will be a plus.