Project Pitch: Multimodal Dementia Identification with Audio and Video Data

Royce Salah
College of Computing
Georgia Institute of Technology
San Jose, CA, United States of America
https://orcid.org/0009-0004-7415-7083

Abstract—**This project seeks to develop a multimodal machine learning tool to assess cognitive decline in dementia using spontaneous speech, acoustic features, and visual behavioral signals. Drawing from data sets such as DementiaNet, WLS, and Dem@Care, the system will extract linguistic, audio, and visual markers, such as lexical diversity, pitch variability, and gaze instability. Using multimodal fusion, the model will attempt to predict the severity and probability of dementia presence.**

I. Introduction

This project lies within the intersection of linguistics, psychology, and artificial intelligence to model cognitive decline in individuals with dementia. It focuses on linguistic and visual markers such as language diversity, semantic coherence, and facial cues, all features tied to semantic and episodic memory which can help identify signs of cognitive deterioration. The computational model will leverage machine learning models across audio, text, and video modalities to capture subtle behavioral and linguistic indicators. By combining these signals with multimodal fusion, the goal is to develop a more sensitive and holistic approach to assessing dementia risk and supporting early-stage detection.

II. Topic of Research

A. Research Question

How effectively can linguistic, acoustic, and visual behavioral features from speech and video recordings predict cognitive decline severity and dementia symptoms in older adults?

B. Research Interest

My interest in this project is deeply personal. Both my grandfather and father were diagnosed with dementia-related conditions — Alzheimer's and Lewy Body dementia, respectively. Witnessing first-hand the progression of these diseases made me aware of how subtle, early-stage cognitive changes can go unnoticed until it is already too late. These experiences have driven me to explore how emerging machine learning methods might help identify these complex and highly nuanced symptoms earlier and more objectively.

From a technical perspective, I am particularly interested in the challenge of attempting to capture these highly ambiguous signals such as semantic drift in speech or shifts in facial cues by using machine learning. I want to understand how cognitive science principles around memory, language degradation, and perceptual attention can be captured through models trained on real data. This project represents an opportunity to integrate state-of-the-art ML with human cognitive processes in a way that could eventually support clinicians, caregivers, and patients.

C. Research Importance

Dementia is a progressive and underdiagnosed condition, especially in its early stages when brief cognitive assessments may overlook subtle yet meaningful behavioral and cognitive cues. This project aims to identify and model early indicators such as changes in speech, facial expression, and attention as part of a more sensitive and objective tool for detecting cognitive decline.

The work also aligns with the growing industry of modern health technologies and virtual assistants. As wearable technology, biomarker sensors, and passive listeners become more capable and widespread, passively collected speech and video data offer an avenue for capturing fluctuations in cognitive function frequently influenced by factors like sleep, medication, or stress and which could go unnoticed in traditional point-in-time assessments.

III. Initial Research Design and Outcomes

The proposed system will take a combination of data sources, including speech and video recordings. These recordings, sourced from Dem@Care [1] and DementiaNet [2], will be fed through a pipeline of data processing and several models to identify key indicators such as facial landmark detection, gaze, vocal features, and nuances in semantic drift. The Wisconsin Longitudinal Study (WLS) additionally provides important longitudinal data such as diagnosis stage, time relative to diagnosis, and basic demographic information such as age and gender which can help lower the margin of error of the prediction. These inputs will be processed through a fusion pipeline to generate a dementia severity prediction score as a single output. This score is intended to quantify the degree of cognitive decline and support early detection, ongoing monitoring, and more informed clinical assessments.

A. Core Cognitive Concepts and Research Approach

This project draws on a few core cognitive science concepts to inform its approach. Semantic memory degradation is a core manifestation of dementia, showing its signs with reduced vocabulary diversity, difficulty with retrieval, and decreased topic coherence in conversational speech. Additionally, principles of semantic drift and cognitive load estimation provide a solid framework for detecting how text and audio features may implicate disruptions in the memory retrieval process. Visual data integration is also central, as patients exhibiting signs of dementia can exhibit trailing gaze patterns, lowered responsiveness, and altered facial motion, indicating broader cognitive shifts.

These principles will be further researched through a literature review on the topics of linguistic patterns and neurodegenerative disease research. As noted, this research and development should ultimately culminate with the development of a multimodal predictive computational tool, which integrates features extracted from audio, text, and video sources. The system will experiment with both early and late fusion strategies to combine these data sources into a unified pipeline.

IV. Appendix

TABLE I
Individual Project Timeline and Task Breakdown

Week	Task	Task Description	Hours	Done (Y/N)
2–3	1	Create the template task list.	0.5	Y
2–3	2	Choose research question.	0.25	Y
2–3	3	Prepare and send data request emails.	0.75	Y
2–3	4	Draft ML pipeline workflows.	2	Y
2–3	5	Prepare project pitch.	5	Y
PROJECT PITCH DUE				
4–6	6	Research feature extraction tools.	10	N
4–6	7	Clean audio data.	5	N
4–6	8	Build audio feature extraction pipeline.	5	N
4–6	9	Transcribe audio to extract textual cues.	5	N
4–6	10	Use pretrained models to assess lexical diversity	5	N
4–6	11	Clean visual data.	5	N
4–6	12	Build visual cue extraction pipeline.	5	N
4–6	13	Develop baseline multimodal fusion model.	10	N
4–6	14	Prepare for midpoint check-in.	5	N
OPTIONAL MIDPOINT CHECK-IN				
7–9	15	Evaluate model.	5	N
7–9	16	Finetune prototyped model.	15	N
7–9	17	Prepare final report.	12	N
7–9	18	Final presentation prep.	5	N
FINAL REPORT DUE				
10	19	Prepare presentation.	5	N
10	20	Conduct final presentation.	3	N
FINAL PRESENTATION DUE				
Total			108.5 hrs	

References

[1] Herd, Pamela, Deborah Carr, and Carol Roan. 2014. “Cohort Profile: Wisconsin Longitudinal Study (WLS).” International Journal of Epidemiology 43:34-41 PMCID: PMC3937969.

[2] Gite, Shreyas. 2023. *DementiaNet: Longitudinal dataset of speech from public figures with and without dementia*. GitHub repository. Available: https://github.com/shreyasgite/dementianet