Speech-Based Detection and Severity Assessment of Alzheimer's Disease

Valentine Ezikeoha, Howard Prioleau, Saurav K. Aryal PhD
AI4PC Lab, Institute for Human Centered Artificial Intelligence

INTRODUCTION

Alzheimer's Disease (AD) affects millions globally and is increasing with aging populations

Current diagnosis methods are Expensive, and Time-consuming:

- Cognitive decline impacts language, fluency, and speech patterns
- Linguistic + acoustic features correlate with MMSE cognitive scores

Prior work:

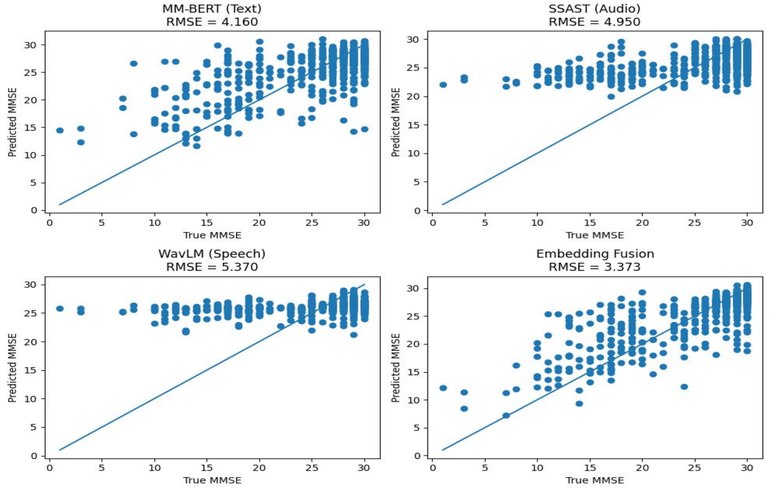
In prior work, we achieved state-of-the-art using text features, and showed that linguistic features outperform audio alone

The gap between Audio-only and multilingual robustness remain underexplored

Our goal was to build a multimodal (audio + text), multilingual system for, AD detection, Severity prediction, Cross-dataset generalization

METHODS

We evaluate four complementary approaches for speech-based Alzheimer's detection: **SSAST, WavLM, mmBERT, and a multimodal fusion model**, each capturing distinct aspects of cognitive decline. SSAST and WavLM learn acoustic representations), while speech transcripts are generated via ASR and encoded with **mmBERT** to capture linguistic features such as syntax and semantic coherence. These representations are combined in a **fusion model** integrating audio and text embeddings.



METHODS

Speech recordings are processed through parallel audio and text pipelines to capture complementary information. Audio signals are encoded using **SSAST and WavLM** to learn acoustic representations such as prosody, pauses, and articulation.

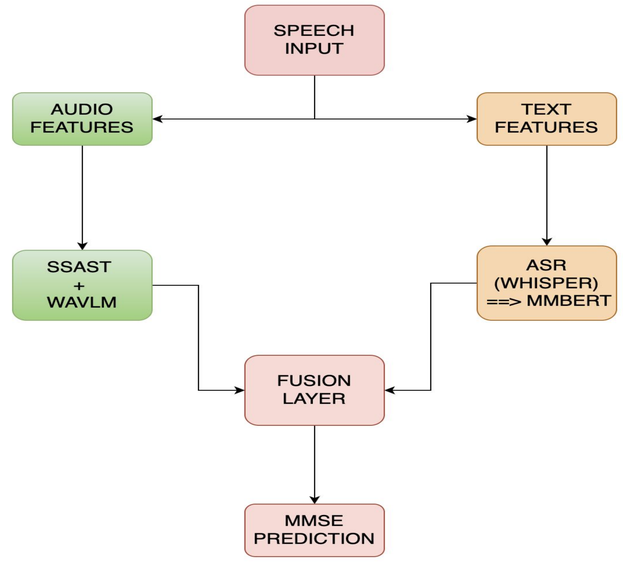
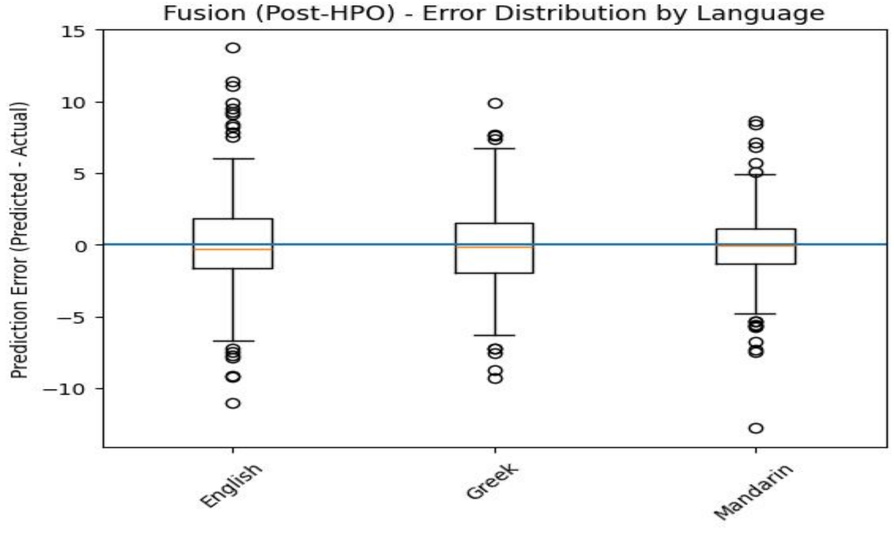


Figure: Multimodal pipeline combining audio (SSAST, WavLM) and text (mmBERT via ASR) representations through fusion for Alzheimer's detection and MMSE prediction

RESULTS

Post-HPO models demonstrate significantly improved prediction stability, with tighter residual distributions and fewer extreme errors. The **fusion model** achieves the most consistent performance, benefiting from the integration of both acoustic(SSAST, WavLM) and linguistic (mmBERT) features



RESULTS

HPO significantly reduces prediction error across all modeling approaches.

Model	RMSE FUSION PRE-HPO	RMSE FUSION POST-HPO
ADReSS	5.150584	4.296281
ADReSS-M	4.423297	3.020264
ADReSSo_DIAG	3.756449	2.900144
TAUKADIAL	2.382247	2.319332
ALL	3.65816	2.913521

CONCLUSION

We developed a multimodal framework for Alzheimer's detection using SSAST, WavLM, and mmBERT, integrating both acoustic and linguistic features through a fusion model. Our results show that hyperparameter optimization significantly improves model performance..

The fusion approach achieves the most stable and reliable predictions, demonstrating the value of combining multiple modalities.

Multimodal fusion combined with hyperparameter optimization provides the most accurate and robust approach for speech-based Alzheimer's detection.

REFERENCES

[1]Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. INTERSPEECH.
[2]Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021). Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. INTERSPEECH.
[3]Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. arXiv
[4]Chen, S., Wang, C., Chen, Z., et al. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Speech Processing. IEEE JSTSP.
[5]Aryal, S. K., Shah, U., Burge, L., & Washington, G. (2023). From Predicting MMSE Scores to Classifying Alzheimer's Disease Detection and Severity. CCSC

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714 and an Amazon Research Award. The work is solely the responsibility of the authors and does not necessarily represent the official view of the sponsors.