ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095825

EARLY DETECTION OF COGNITIVE DECLINE USING VOICE ASSISTANT COMMANDS

Eli Kurtz[1], *Youxiang Zhu*[1], *Tiffany Driesse*[2], *Bang Tran*[1], *John A. Batsis*[2],
Robert M. Roth[3], and *Xiaohui Liang*[1]

[1]Department of Computer Science, University of Massachusetts Boston, MA, USA
[2] Division of Geriatric Medicine, School of Medicine,
University of North Carolina at Chapel Hill, NC, USA
[3] Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

ABSTRACT

Early detection of Alzheimer's Disease and Related Dementias (ADRD) is critical in treating the progression of the disease. Previous studies have shown that ADRD can be detected and classified using machine learning models trained on samples of spontaneous speech. We propose using Voice-Assistant Systems (VAS), e.g., Amazon Alexa, to monitor and collect data from at-risk adults, and we show that this data can be used to achieve functional accuracy in classifying their cognitive status. In this paper, we develop multiple unique feature sets from VAS data that can be used in the training of machine learning models. We then perform multi-class classification, binary classification, and regression using these features on our dataset of older adults with three varying stages of cognitive decline interacting with VAS. Our results show that the VAS data can be used to classify Dementia (DM), Mild Cognitive Impairment (MCI), and Healthy Control (HC) participants with an accuracy up to 74.7%, and classify between HC and MCI with accuracy up to 62.8%.

Index Terms— Voice assistant, early detection, cognitive decline, machine learning, linguistic and acoustic features

1. INTRODUCTION

Alzheimer's disease and related dementias are characterized by progressive degeneration of cognitive function, including the capability of producing coherent speech [1, 2]. Analyzing spontaneous speech for symptoms of Dementia (DM) is recognized as an important frontier in early diagnosis of the disease, as patients usually show a decline in both syntactic and semantic language faculties as the disease progresses [3, 4]. The progression of Dementia occurs over several stages from Healthy Control (HC), to Mild Cognitive Impairment (MCI), to Dementia (DM) with other intermediary stages depending on the staging scheme. An important functionality of an automated detection routine would be the classification of a patient into one of these categories. Classification between HC and MCI we regard as the most important task, as an early diagnosis can allow for proper planning and patient treatment, which can, in turn, lead to better outcomes [5].

The relatively non-invasive and low-cost nature of speech recording makes it a promising target for developing automated detection tools. However, developing such tools requires extensive data of both audio and transcripts. In the ADReSS challenge, spontaneous speech was induced through a picture description task, and transcripts of the participant's speech were produced manually [6]. This method is impractical for an automated detection routine. Conversely, speech induced through interactions with Voice-Assistant Systems (VAS) is spontaneous and ongoing over time: well suited to an automated routine. Those who own a VAS device tend to use it multiple times per day, and a significant proportion of VAS users are adults over the age of 55 [7].

VAS provides downloadable data of the user's speech audio and ASR transcripts. This allows for both acoustic and linguistic features to be examined in combination or individually and removes the high cost of human transcription. Previous studies have shown that models trained on acoustic features alone can achieve an accuracy of 60% in classification tasks [6]. However, in the ADReSS challenge, linguistic features from transcripts were shown to be more effective than acoustic features [6, 8, 9, 10]. Therefore, the capability of analyzing both is an advantage [11, 12].

Our study relies on the transcripts produced by a VAS, which employs a proprietary Automatic Speech Recognition (ASR) algorithm. ASR can introduce errors into the generated transcripts, which are often measured using the Word Error Rate (WER). When recognizing speech from patients with cognitive decline, this error rate might be elevated [13]. We seek to mitigate this effect by using both lexical and semantic features. The evaluation of VAS transcripts in detecting DM is a major contribution of our research.

To investigate the early detection of cognitive decline, we recruited 90 older adult participants to interact with a VAS. The audio and transcripts of speech from participants to the VAS were used to implement early detection models of cognitive decline. The contributions of our paper are three-fold.

First, we develop several novel feature extraction routines that explore lexical and semantic features of transcripts and pre-trained embedding of audios from VAS interactions.

Second, we evaluate these features and their combinations via various machine learning algorithms, which we employ to classify a participant's Alzheimer's status and predict their cognitive scores based on their interactions with a VAS.

Lastly, we show that utilizing a VAS for early detection of cognitive decline can allow for robust automated procedures that will be able to identify signs of cognitive decline from voice interactions between the VAS and the participants.

2. VAS DATA

We recruited 30 HC, 30 MCI, and 30 DM patients, to interact with an Alexa Echo device in a remote/in-person setting. Our remote evaluation was offered due to the COVID-19 pandemic. Each participant was given a list of 30 Alexa commands, including "Alexa, what is the weather outside?," "Alexa, remember my daughter's birthday is June first," and "Alexa, add oranges and grapes to my shopping list." These commands were selected due to their popularity among older adults. Our study included four other commands that participants would say in response to Alexa, including "yes" and "pause." This yielded a total of 34 accepted commands. A complete list of commands and the protocol can be found in the previous work [7]. We list the demographics in Table 1: 30 patients with HC (Montreal Cognitive Assessment (MoCA) score ≥ 26), 30 patients with MCI (MoCA score $<$ 26) and 30 patients with dementia (based on clinical evaluations). One DM patient failed to produce commands due to advanced dementia and was excluded, leaving 29 DM participants. The number of HC males is slightly more than HC females, while the numbers of females in both MCI and DM are slightly more than males. This difference was not statistically significant $[\chi^2(1) = 1.16, p = .56]$.

	HC			MCI			DM		
Age	M	F	MoCA	M	F	MoCA	M	F	MoCA
[65, 70)	4	6	27.5	5	4	23.33	4	3	14.86
[70, 75)	10	6	27.19	6	8	23.29	1	3	13.5
[75, 80)	3	1	27.25	3	3	23.33	1	2	18
≥80	0	0	NA	0	1	25	7	9	12.94
Total	17	13	27.3	14	16	23.37	13	17	13.97

Table 1: Demographics. **M**ale, **F**emale, **N**ot **A**vailable.

3. FEATURES

We extracted features from the VAS data, including Basic, Distance, and Acoustic features, and their combinations.

3.1. Basic features

We constructed two rudimentary transcript-based features, called "Basic" features. One basic feature is a binary variable assigned to each prompted command: 1 if a participant produced a command matching the prompted command, and 0 if a participant did not produce a command matching the promoted command. The other basic feature we use is the number of unrecognized commands given by the participant. An unrecognized command is when the VAS system is triggered to record but cannot understand the input audio. In this study, we distinguish *unmatched* commands from *unrecognized* commands, the former being commands that do not exactly match a prompted command but still elicit a response from Alexa, and the latter being unintelligible audio.

3.2. Distance features

We assembled a set of "Distance" features: for each *unmatched* command from a participant, we used both lexical and semantic methods to determine their most likely intended (or "closest") command. For lexical distance, we compared each spoken command to each prompted command and calculated the WER between the two, selecting the lowest WER command-pair as the intended command. Using the WER metric for each command, we assembled a new vector, indicating how well the participant performed each command. If the WER for a spoken command vs. the matched prompted command was high, this indicated that the participant struggled to produce the command. The WER metric, however, does not capture the semantic intent of a user's command. For example, if a participant wanted to quiet the VAS, they might say, "Alexa, be quiet," rather than the prompted command "Alexa, pause." As the goal of this study is to evaluate the eventual use of VAS for in-home monitoring, we wanted to be able to account for varied inputs that might have the same meaning as our prompted commands. To incorporate this intent into our feature set, we employed BERT [14] embeddings. An embedding was generated for each command given by a participant and was matched using cosine similarity to the closest BERT embedding of a prompted command, giving us an idea of what the intent of the participant's utterance was. We used this metric to construct another vector measuring a participant's competency in performing commands. In combination, these features allowed us to extrapolate from an unmatched command and observe which prompted command a participant was likely trying to perform. However, both these Distance features rely on the accuracy of the ASR transcript, which may contain errors.

3.3. Acoustic features

We consider the following two acoustic features:

eGeMAPS: eGeMAPS is an acoustic feature set that was developed for and has been widely used in voice research and computing applications [15]. We extracted the 88 features specified - using the OpenSmile python package and eGeMAPS v02 - from the audio commands.

HuBERT: Transfer learning focuses on storing knowledge gained from an easy-to-obtain large-sized dataset from a general task and applying the knowledge to a downstream task where the data is limited. To extract embeddings from our audio recordings, we adopted HuBERT as a pre-trained model [16]. HuBERT is the state-of-the-art automatic speech recognition model. It was trained on a large-scale dataset LibriSpeech [17] and was fine-tuned on the Librilight splits [18]. We envision the knowledge of the pre-trained HuBERT model will help extract useful features from the audio commands.

3.4. Feature sets

Using our basic, distance, and acoustic features, we assembled three more feature sets as combinations of the existing features: FS1 (Basic + Distance), FS2 (Basic + Distance + eGeMAPS), and FS3 (Basic + Distance + HuBERT).

4. MODELS

Model choices for this research were informed by the models selected for use in the ADReSS Challenge. For the tasks of classification, each feature set in our study was classified using 5 different models: Decision Tree (DT), Linear Discriminant Analysis (LDA), Linear Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Random Forest (RF). In addition to the 5 general machine learning models, we employed TPOT, an auto-Machine-Learning (ML) algorithm, to create uniquely tuned models for each feature set [19]. Because TPOT searches many pipelines for the most optimal ML model, it is necessary to specify a maximum time limit on model searching. TPOT models were each fitted to feature sets for 60 minutes. Despite standardized training time, the models produced by TPOT are arrived at through a randomized search, and are therefore not replicable except with the exact model code [19]. We built multi-class models for DM, MCI, and HC classes. Each class's chance level is 0.33.

We also created three binary classification tasks to predict between DM vs. HC, DM vs. MCI, and MCI vs. HC. With seven feature sets, three classification tasks per feature set, and five models per classification task, we ran a total of 105 binary classifications. Each class's chance level is 0.5.

Using a similar approach, we built regression models to predict an MoCA score. MoCA ranges from 0 - 30, with a lower score indicating more cognitive impairment. Three different regression methods were used: Linear Ridge Regression (LRR), DT, and SVM. We first used these methods with default hyperparameters and then used TPOT to optimize a model for each regression feature set.

5. EVALUATION

Evaluation settings. Evaluation of all classification models was conducted using stratified K-Fold cross-validation (k=10) and 10 repetitions per feature set for a total of 100 scoring trials per model. Model-specific accuracy is reported as the average of these trials. Regression models were evaluated using the same K-Fold cross-validation strategy. We averaged all model-specific trials to report the average Root-Mean-Square Error (RMSE) of MoCA scores.

Evaluation results. Figure 1a displays the average accuracy of our multi-class classification models. We observed a higher accuracy for the Distance features at 62.3% when compared to the Basic features at 57.8%. Further, we can see that FS1 - the combination of Basic and Distance features - produced the highest overall accuracy at 64.2%. This far exceeds the chance level of 33%. Of the default models, RF achieved the highest average accuracy across all feature sets, at 62.4%, and achieved the highest accuracy using FS3, at 73.4%. FS3 includes Basic, Distance, and HuBERT features, confirming the effectiveness of the newly developed Distance features and features from pre-trained models. We show the confusion matrix for this model in Figure 1b, which confirms the model was primarily confused between MCI and HC.

Optimized pipelines generated by TPOT achieve even higher accuracy for each feature set. TPOT models achieved the highest multi-class accuracy at 74.7% using FS1.

Figure 1a also includes the evaluation results of the regression models. The feature set that produced the lowest average RMSE using a default model was FS3 (4.81). FS3 also contained the lowest average RMSE for a default model (LRR, 4.28). TPOT regressors achieved lower RMSE across most models, with the exception of FS3. A TPOT model trained on HuBERT features achieved the overall lowest RMSE of 3.95.

Table 2 displays the average accuracy of our binary classification models. Classifications between DM and HC and between DM and MCI were robust, but confusion between HC and MCI lowered the accuracy, nearly reaching 0.5.

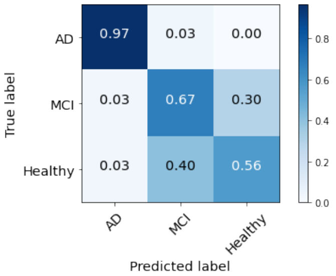
DM vs. HC. FS3 produces the highest accuracy with default models in this task (96.5%), slightly higher than the Basic features (96.2%). The TPOT models (96.7%) marginally outperformed the default models when trained on FS3. High classification accuracy between these two classes is due to the significant difference in matched and recognized commands.

DM vs. MCI. FS3 produces the highest accuracy on default models (97.5%). The Distance features outperformed the Basic, confirming the effectiveness of lexical and semantic features. The TPOT model achieved 98.3% on FS3.

MCI vs. HC. The RF model trained on HuBERT features produced the highest accuracy at 62.8%. HuBERT features also produced the highest average accuracy at 59.4%. This shows the effectiveness of acoustic features from pre-trained models in the early detection of cognitive decline. Our TPOT model (59.2%) trained on HuBERT features performed marginally worse than the average of default HuBERT models. Linguistic features might not help in our evaluation, as most HC and MCI finished all 30 commands correctly. However, we observed that most classifiers produce about the chance level, confirming the difficulties of the early detection.

	Multi-class Classification DM vs. MCI vs. HC							Regression				
Features	DT	LDA	SVM	KNN	RF	mean	TPOT	DT	SVM	LRR	mean	TPOT
Basic	0.606	0.675	0.574	0.517	0.519	**0.578**	0.601	5.34	5.35	4.49	5.06	4.30
Distance	0.715	0.572	0.661	0.506	0.662	**0.623**	0.744	5.03	5.08	4.39	4.83	3.99
Acoustic	0.443	0.483	0.340	0.371	0.529	0.433	0.437	6.31	6.69	6.30	6.43	4.85
HuBERT	0.464	0.619	0.586	0.517	0.618	0.561	0.644	6.13	6.58	4.40	5.70	**3.95**
FS1	0.684	0.685	0.697	0.494	0.652	**0.642**	**0.747**	5.41	5.03	4.34	4.93	4.03
FS2	0.632	0.531	0.340	0.428	0.657	0.518	0.613	5.81	6.68	6.28	6.26	4.58
FS3	0.683	0.560	0.697	0.517	**0.734**	0.638	0.697	5.19	4.95	**4.28**	**4.81**	5.16
mean	0.604	0.589	0.549	0.479	**0.624**			5.60	5.77	4.93		

(a) Evaluation results of multi-class classification and regression models



(b) FS 3, RF model

Fig. 1: Evaluation results of multi-class classification and regression models and confusion matrix

	DM vs. HC						DM vs. MCI						MCI vs. HC					
Feature set	DT	LDA	SVM	KNN	RF	mean	DT	LDA	SVM	KNN	RF	mean	DT	LDA	SVM	KNN	RF	mean
Basic	0.950	0.913	0.983	0.983	0.983	**0.962**	0.917	0.883	0.967	0.900	0.958	0.925	0.405	0.517	0.467	0.400	0.393	0.436
Distance	0.970	0.817	0.967	0.983	0.982	0.944	0.981	0.867	0.983	0.983	0.983	0.959	0.575	0.533	0.500	0.350	0.478	0.487
eGeMAPS	0.781	0.797	0.678	0.710	0.863	0.766	0.800	0.880	0.565	0.610	0.823	0.736	0.373	0.350	0.493	0.483	0.438	0.427
HuBERT	0.722	0.933	0.767	0.750	0.782	0.791	0.728	0.900	0.783	0.797	0.795	0.801	0.593	0.600	0.617	0.533	**0.628**	**0.594**
FS1	0.953	0.847	0.967	0.983	0.983	0.947	0.981	0.863	0.983	0.983	0.983	0.959	0.567	0.467	0.467	0.300	0.468	0.454
FS2	0.958	0.967	0.966	0.797	0.983	0.934	0.979	0.933	0.942	0.727	0.983	0.913	0.370	0.400	0.488	0.533	0.427	0.444
FS3	0.955	0.950	0.967	0.983	0.970	**0.965**	0.979	0.983	0.983	0.983	0.945	**0.975**	0.597	0.550	0.477	0.367	0.615	0.521
mean	0.898	0.889	0.899	0.884	0.935		0.909	0.901	0.887	0.855	0.924		0.497	0.488	0.501	0.424	0.492	

[1] DM vs. HC FS3 TPOT accuracy = **0.967**, [2] DM vs. MCI FS3 TPOT accuracy = **0.983**, [3] MCI vs. HC HuBERT TPOT accuracy = **0.592**.

Table 2: Evaluation results of binary classification models

6. DISCUSSION

Our data from older adults interacting with a VAS has been evaluated for the first time to classify DM, MCI, and HC. Previous evaluations classified MCI vs. HC using a smaller dataset [7]. Major contributions of our paper include showing the high accuracy of classifying DM patients, and confirming the importance of acoustic features in early detection.

The most important decision boundary for our models is between MCI and HC. The confusion matrix shown in Figure 1b indicates that the primary source of error in our multi-class classification models comes from the incorrect classification of MCI and HC. Our binary classification models also struggle in distinguishing MCI and HC. One of the reasons is the design of our 30-command task, where participants were given prompted commands. We observed that both MCI and HC could correctly finish the task easily by reading the commands. We conclude that acoustic features will be more important in this task than transcript-based features. Accordingly, we observed that the acoustic features from the pre-trained HuBERT model achieved the highest accuracy in binary classification at 59.4% and confirmed the importance of acoustic features in the early detection of cognitive decline. We are currently experimenting with a new 30-intent task in which participants will not be provided with prompted commands but keywords to fulfill intents. In addition, we are collecting longitudinal VAS data from participants' home, where participants will speak commands based on their free will.

Regression results are encouraging, as the RMSE scores are low. However, while MoCA is useful in helping to identify HC, it does not necessarily indicate a diagnosis of either DM or MCI. Some participants had MoCA scores of 23 and yet were suffering from DM, while others who had MCI had MoCA scores of as low as 17. Our regression models might accurately predict the MoCA score, but this score might not directly confer DM or MCI.

7. CONCLUSION

In this paper, we evaluate the ability of a 30-command task using VAS for early detection of cognitive decline. Our features and models perform extremely well in classifying DM patients from MCI and HC. Our multi-class classifier achieved 74.7% accuracy. In classifying MCI and HC patients, our feature sets achieved a maximum accuracy of 62.8%. While MCI and HC both successfully finish the task with quality transcripts, acoustic features of their speech extracted via the pre-trained HuBERT model play an important role in the early detection of cognitive decline. These results are promising and indicate that with further experiments and development, VAS could become an effective, passive, and low-cost tool to monitor patients for early signs of cognitive decline.

8. ACKNOWLEDGEMENT

This research is funded by the US National Institutes of Health National Institute on Aging (R01AG067416). We thank Dr. David Kotz and Dr. Brian MacWhinney for their valuable comments and suggestions. More information about the project can be found at cogvox.org.

9. REFERENCES

[1] Tom Bschor, Klaus-Peter Kühl, and Friedel M Reischies, "Spontaneous speech of patients with dementia of the alzheimer type and mild cognitive impairment," *International psychogeriatrics*, vol. 13, no. 3, pp. 289–298, 2001.

[2] Michael A DeTure and Dennis W Dickson, "The neuropathological diagnosis of alzheimer's disease," *Molecular neurodegeneration*, vol. 14, no. 1, pp. 1–18, 2019.

[3] Blanka Klimova, Petra Maresova, Martin Valis, Jakub Hort, and Kamil Kuca, "Alzheimer's disease and language impairments: social intervention and medical treatment," *Clinical interventions in aging*, vol. 10, pp. 1401, 2015.

[4] Flavio Bertini, Davide Allevi, Gianluca Lutero, Danilo Montesi, and Laura Calzà, "Automatic speech classifier for mild cognitive impairment and early dementia," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–11, 2021.

[5] Gary W Small, "Early diagnosis of alzheimer's disease: update on combining genetic and brain-imaging measures," *Dialogues in clinical neuroscience*, 2022.

[6] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: the adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[7] Xiaohui Liang, John A Batsis, Youxiang Zhu, Tiffany M Driesse, Robert M Roth, David Kotz, and Brian MacWhinney, "Evaluating voice-assistant commands for dementia detection," *Computer Speech & Language*, vol. 72, pp. 101297, 2022.

[8] Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al., "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[9] Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee, "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," *arXiv preprint arXiv:2009.04070*, 2020.

[10] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth, "Domain-aware intermediate pretraining for dementia detection with limited data," *Proc. Interspeech 2022*, pp. 2183–2187, 2022.

[11] Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection.," in *Interspeech*, 2021, pp. 3790–3794.

[12] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in computer science*, p. 22, 2021.

[13] Richard Cave and Steven Bloch, "The use of speech recognition technology by people living with amyotrophic lateral sclerosis: a scoping review," *Disability and Rehabilitation: Assistive Technology*, pp. 1–13, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[16] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: How much can a bad teacher benefit asr pretraining?," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[19] Trang T Le, Weixuan Fu, and Jason H Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.