Predicting dementia from spontaneous speech using large language models

Felix Agbavor, Hualou Liang, 2022

Gian Marco Simonazzi

gianmarco.simonazzi@studenti.unipr.it

Masters degree course in Computer Science

2023/24

Introduction

The main objective of the proposed method is the detection of Alzheimer’s disease (AD)

Current diagnosis of dementia are made through:

- brain imaging
- biomarkers detection
- cognitive tests, such as the MMSE (Mini-Mental State Examination)

These medical evaluations are however lengthy and expensive [1,2]

Spontaneous speech has been shown to contain valuable information in AD

Previous works

Mainly focused on speech features:

- Acoustic features: prosody and sound frequency spectrum
- Language features: lexicon, syntax, semantic coherence, sentiment

[2]

Most studies focused on domain-specific methods and knowledge, which may not generalize to various stages of the disease.

AI-driven speech analysis has recently emerged as a viable option for early screening of dementia.

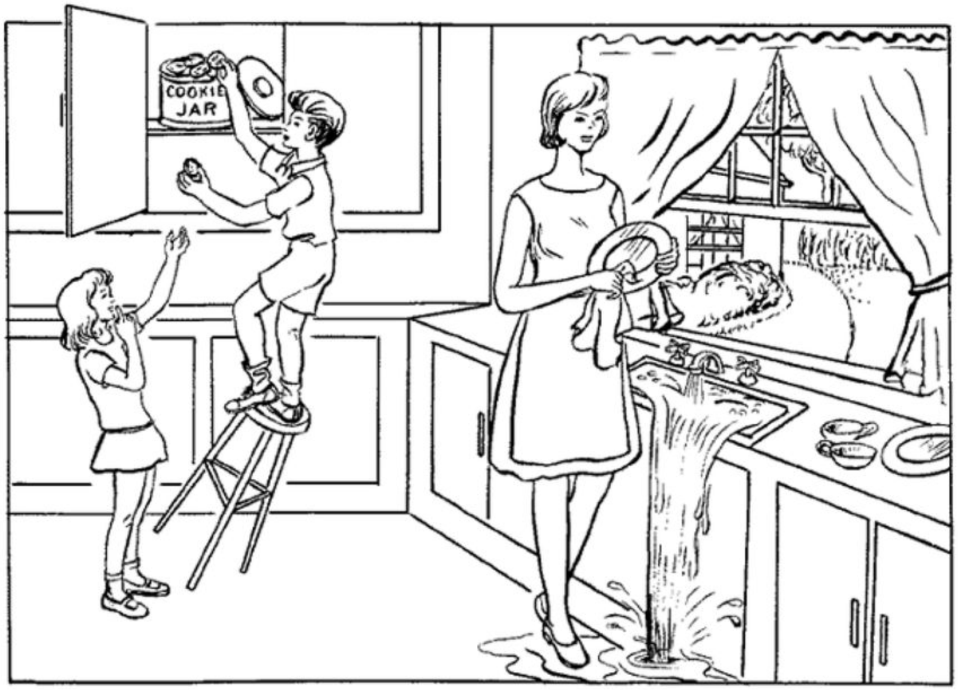
- BERT or other language models [5,6-10]
- ADReSS dataset [3,4]

The dataset

The dataset is provided by ADReSS challenge.

- 237 recordings in total
- Balanced for age and gender
- 70/30 split
- Training set balanced for AD diagnosis
- No transcription provided (ADReSSo)

The subjects are tasked to describe the *Cookie Theft* image from the Boston Diagnostic Aphasia Examination.



Embedding

Method that can encode the semantic value of a word in a vector space.
Words with similar meaning are close in the semantic space.

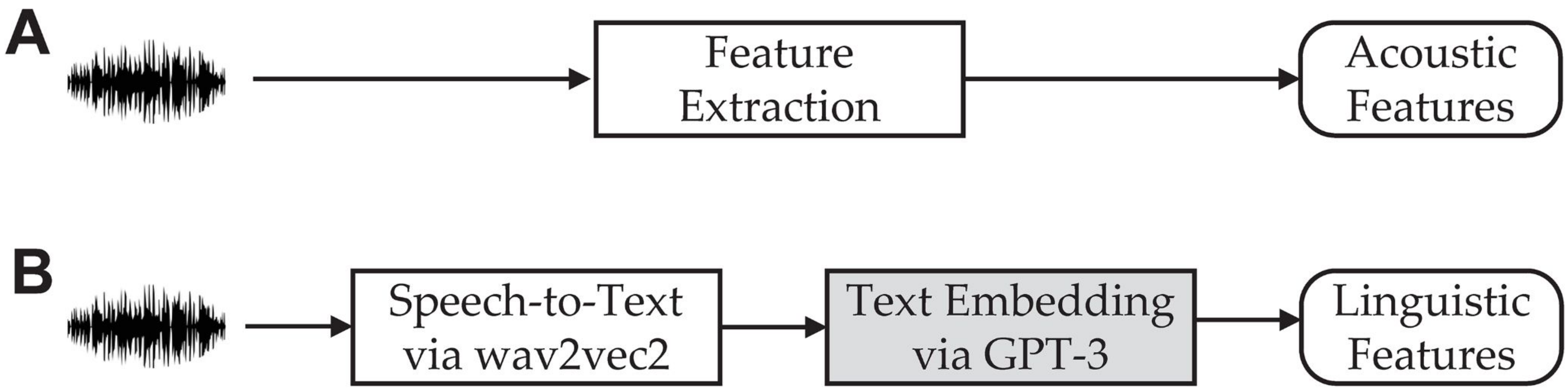
GPT3 is a Large-Language Model developed by OpenAI

- Text understanding and generation capabilities
- Zero-shot learning
- Embeddings
- Fine-tuning for domain-specific tasks

The paper uses the two smaller models available:

- Ada: 1k embedding vector, 300M parameters
- Babbage: 2k embedding vector, 1.2B parameters

The proposed method



https://doi.org/10.1371/journal.pdig.0000168.g001

B is the main method proposed:

1. Automatic speech transcription via wav2vec2
2. Text embedding encoding via GPT-3
3. Classification/Regression via ML model

A is a secondary method used for comparison:

1. Acoustic features extraction (temporal analysis, frequency spectrum, prosody, …)
2. Classification/Regression via ML model

Acoustic-based classifier

All ML models are trained with 10-fold CV

Random Forest, Support Vector Classifiers and Logistic Regression are tested for their performance

	Model	Accuracy	Precision	Recall	F1
10-fold CV	SVC	**0.697 (0.095)**	**0.722 (0.091)**	0.660 (0.120)	0.678 (0.084)
	LR	0.632 (0.120)	0.645 (0.136)	0.656 (0.131)	0.647 (0.121)
	RF	0.668 (0.101)	0.705 (0.156)	**0.704 (0.114)**	**0.686 (0.084)**
Test Set	SVC	0.634	0.657	0.622	0.639
	LR	0.620	0.600	0.618	0.609
	RF	**0.746**	**0.771**	**0.730**	**0.750**

https://doi.org/10.1371/journal.pdig.0000168.t001

Random forest is the best performer in this scenario

Embedding-based classifier

The embeddings are provided by GPT-3.
Both Ada and Babbage results are showed.

	Embeddings	Model	Accuracy	Precision	Recall	F1
10-fold CV	Ada	SVC	0.788 (0.075)	0.798 (0.109)	0.819 (0.098)	0.799 (0.066)
		LR	0.796 (0.107)	0.798 (0.126)	**0.835 (0.129)**	0.808 (0.100)
		RF	0.734 (0.090)	0.738 (0.109)	0.763 (0.149)	0.743 (0.103)
	Babbage	SVC	0.802 (0.054)	0.823 (0.092)	0.804 (0.103)	0.806 (0.053)
		LR	**0.809 (0.112)**	**0.843 (0.148)**	0.811 (0.091)	**0.818 (0.091)**
		RF	0.760 (0.052)	0.780 (0.102)	0.781 (0.110)	0.770 (0.047)
Test Set	Ada	SVC	0.788	0.708	**0.971**	0.819
		LR	0.718	0.653	0.914	0.762
		RF	0.732	0.690	0.829	0.753
	Babbage	SVC	**0.803**	**0.723**	**0.971**	**0.829**
		LR	0.718	0.647	0.943	0.767
		RF	0.761	0.714	0.857	0.779

https://doi.org/10.1371/journal.pdig.0000168.t002

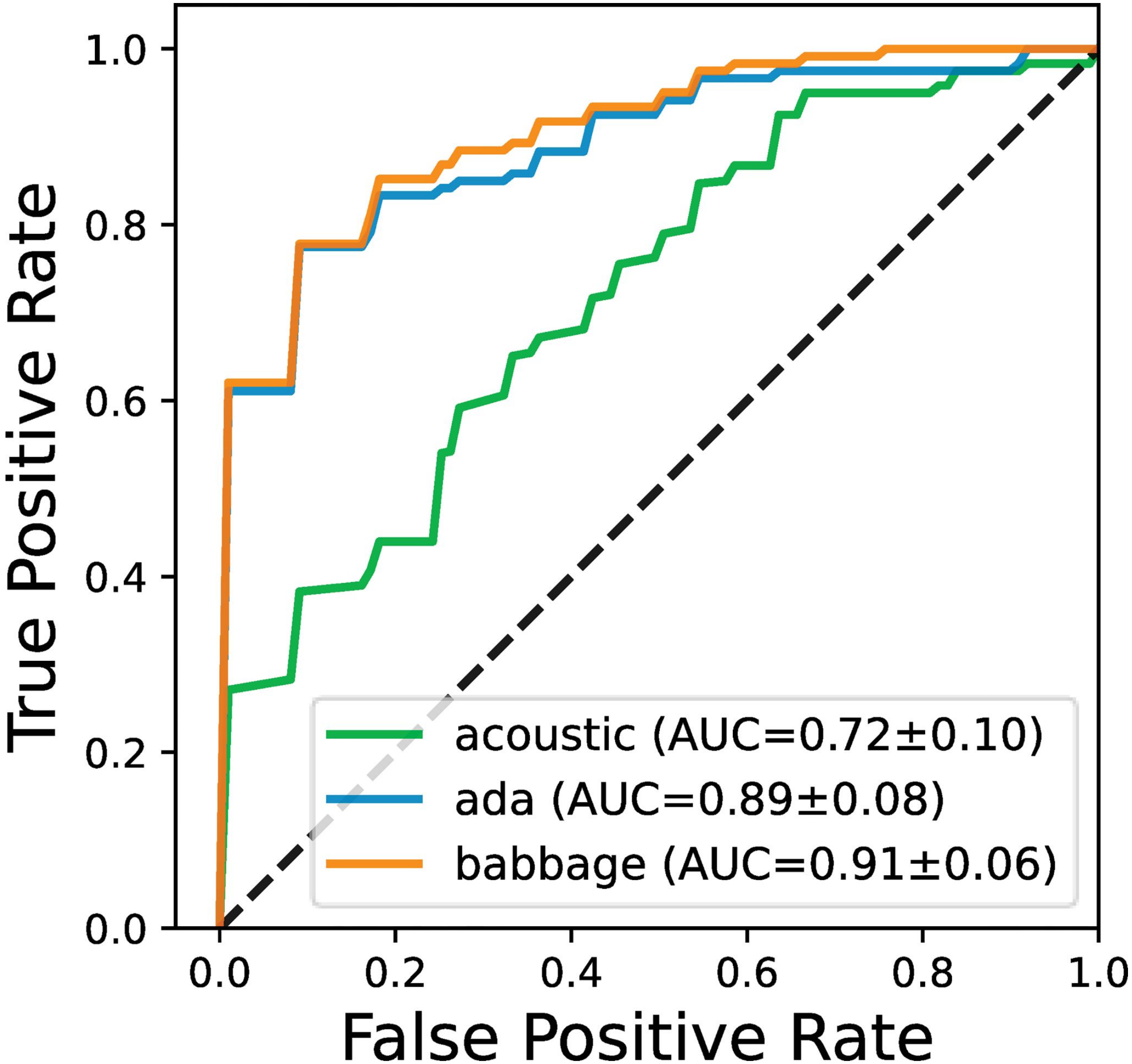
SVC with Babbage is the best performer.
Overall better results than the acoustic-based approach (except for precision).
Impressively high recall is shown.

ROC curves

Acoustic RF, Ada SVC, Babbage SVC.

The AUC values imply that the best performer is Babbage SVC

The acoustic model in particular performs pretty badly



https://doi.org/10.1371/journal.pdig.0000168.g002

Fine tuned GPT-3 as classifier

GPT-3 models can be fine-tuned to perform specific downstream tasks.

Babbage is fine-tuned with the speech transcripts to perform the current classification task.

	Accuracy	Precision	Recall	F1
10-fold CV	0.797 (0.058)	0.810 (0.127)	0.809 (0.071)	0.797 (0.105)
Test Set	0.803	0.806	0.806	0.806

https://doi.org/10.1371/journal.pdig.0000168.t003

Overall uniform results, better than the acoustic-based approach.
Comparable to the embedding-based approach, with less recall but better precision.

Combined classifier

The last classifier tested in the paper uses both the acoustic features and the text embeddings as inputs.

	Model	Accuracy	Precision	Recall	F1
10-fold CV	SVC	**0.814 (0.115)**	**0.838 (0.133)**	0.802 (0.136)	**0.814 (0.119)**
	LR	0.800 (0.108)	0.831 (0.137)	**0.803 (0.097)**	0.809 (0.093)
	RF	0.731 (0.121)	0.741 (0.141)	0.762 (0.119)	0.745 (0.109)
Test Set	SVC	**0.802**	**0.971**	0.723	**0.829**
	LR	0.676	**0.971**	0.607	0.747
	RF	0.788	0.914	0.727	0.810

https://doi.org/10.1371/journal.pdig.0000168.t004

Very similar F1 score and accuracy with respect to the embedding-based classifier.
Precision is much higher, while recall is much lower.

Comparison to other models

The Babbage embedding SVC is compared to other models trained on the same ADReSSo dataset with 10-fold CV.

	Model	Accuracy	Precision	Recall	F1
GPT-3 Embedding (ours)	SVC	0.803	0.723	0.971	0.829
Pan et al 2021	$BERT_{base}$	0.803	0.862	0.714	0.781
Balagopalan et al 2021	SVC	0.676	0.636	0.800	0.709
Luz et al 2021	SVC	0.789	0.778	0.800	0.789

https://doi.org/10.1371/journal.pdig.0000168.t005

This model boasts a better F1 score and much higher recall, but a worse precision.

ADReSS 2020 best classificator

Baidu USA team [9]

	Precision		Recall		F1		Acc
	non-AD	AD	non-AD	AD	non-AD	AD	
Baseline[6]	0.670	0.600	0.500	0.750	0.570	0.670	0.625
BERT0p	0.742	0.941	0.958	0.667	0.836	0.781	0.813
BERT3p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
BERT6p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE0p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE3p	0.852	0.952	0.958	0.833	0.902	0.889	**0.896**

MMSE Score Regressor

The score ranges between 0 (severest dementia) to 30 (healthy).
A score higher of 26 is not considered dementia.

Like for the classifier, 3 different regression models are tested:

- Support Vector Regression
- Ridge Regression
- Random Forest Regression

The error is measured as RMSE.

Acoustic-based

	Model	**RMSE**
10-fold CV	SVR	7.049 (2.355)
	Ridge	**6.768 (1.524)**
	RFR	6.901 (1.534)
Test Set	SVR	6.285
	Ridge	**6.250**
	RFR	6.434

https://doi.org/10.1371/journal.pdig.0000168.t006

Embedded-based Regressor

Like for the classifier, both Ada and Babbage GPT-3 models are tested

	Embeddings	Model	RMSE
10-fold CV	Ada	SVR	6.097 (2.057)
		Ridge	6.058 (1.298)
		RFR	6.300 (1.129)
	Babbage	SVR	5.976 (1.173)
		Ridge	**5.843 (1.037)**
		RFR	6.330 (1.032)
Test Set	Ada	SVR	5.6307
		Ridge	5.8735
		RFR	6.0010
	Babbage	SVR	5.4999
		Ridge	**5.4645**
		RFR	5.8142

https://doi.org/10.1371/journal.pdig.0000168.t007

Ridge regression is best in both the acoustic and the embedding case.
Babbage Ridge is significantly better than Acoustic Ridge

Comparison to other models

No comparisons are given in the paper

ADReSS 2020 best regressors:

Music and Audio Research Group at Seoul National University [8]

Model	Modality	Feature	Classes	Precision	Recall	F1	Accuracy	RMSE
Baseline	Unimodal Network	ComParE	non-AD	0.67	0.50	0.57	0.625	6.14
			AD	0.60	0.75	0.67		
Ours	Unimodal Network	VGGish	non-AD	0.6897	0.8333	0.7547	0.7292	5.0765
			AD	0.7895	0.6250	0.6977		
		Transformer-XL	non-AD	0.8261	0.7917	0.8085	**0.8125**	4.0182
			AD	0.8000	0.8333	0.8163		
	Bimodal Network	VGGish + GLoVE	non-AD	0.7407	0.8333	0.7843	0.7708	4.3301
			AD	0.8095	0.7083	0.7556		
		VGGish + Transformer-XL	non-AD	0.7500	0.7500	0.7500	0.7500	**3.7472**
			AD	0.7500	0.7500	0.7500		
		Ensembled Output	non-AD	0.7586	0.9167	0.8302	**0.8125**	3.7749
			AD	0.8947	0.7083	0.7907		

RMIT University, Australia
Mehran University, Pakistan
[10]

	Accuracy (%)	RMSE
Attempt 1	77.08	4.83
Attempt 2	**85.42**	6.91
Attempt 3	64.58	5.18
Attempt 4	79.17	4.91
Attempt 5	**85.42**	**4.30**
Challenge baseline	62.50	6.15

Conclusions

- Useful new method for AI dementia screening
  - Can be used if false positives are tolerated

- MMSE regression has good but not better results with respect to other methods
  - Multimodal approach not tested

- Insufficient focus on transcription accuracy
  - AD patients can have speech impairments

- Black box problem
  - How is the diagnosis ascertained
  - Cannot replace human doctors

References (1)

[1] Agbavor F, Liang H (2022) Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 1(12): e0000168

[2] Voleti R, Liss JM, Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. IEEE J Sel Top Signal Process. 2019; 14(2):282–98

[3] Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: The ADReSSo Challenge

[4] ADReSS 2020 Challenge Website

[5] de la Fuente Garcia S, Ritchie CW, Luz S. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. J Alzheimers Dis. 2020 Jan 1; 78(4):1547–74

[6] Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. Alzheimers Dement [Internet]. 2022 Jul 7

[7] Balagopalan A, Eyre B, Rudzicz F, Novikova J. To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection [Internet]. arXiv; 2020

References (2)

[8] Koo, J., Lee, J.H., Pyo, J., Jo, Y., Lee, K. (2020) Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition. Proc. Interspeech 2020, 2217-2221, doi: 10.21437/Interspeech.2020-3153

[9] Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., Church, K. (2020) Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. Proc. Interspeech 2020, 2162-2166, doi: 10.21437/Interspeech.2020-2516

[10] Syed, M.S.S., Syed, Z.S., Lech, M., Pirogova, E. (2020) Automated Screening for Alzheimer's Dementia Through Spontaneous Speech. Proc. Interspeech 2020, 2222-2226, doi: 10.21437/Interspeech.2020-3158