INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Detecting cognitive decline using speech only: The ADReSS$_o$ Challenge

Saturnino Luz[1], *Fasih Haider*[1], *Sofia de la Fuente*[1], *Davida Fromm*[2], *Brian MacWhinney*[2]

[1]Usher Institute, Edinburgh Medical School, The University of Edinburgh, UK
[2]Department of Psychology, Carnegie Mellon University, USA
{S.Luz, fasih.haider, sofia.delafuente}@ed.ac.uk, {fromm, macw}@andrew.cmu.edu

Abstract

Building on the success of the ADReSS Challenge at Interspeech 2020, which attracted the participation of 34 teams from across the world, the ADReSS$_o$ Challenge targets three difficult automatic prediction problems of societal and medical relevance, namely: detection of Alzheimer's Dementia, inference of cognitive testing scores, and prediction of cognitive decline. This paper presents these prediction tasks in detail, describes the datasets used, and reports the results of the baseline classification and regression models we developed for each task. A combination of acoustic and linguistic features extracted directly from audio recordings, without human intervention, yielded a baseline accuracy of 78.87% for the AD classification task, a root mean squared error (RMSE) of 5.28 for prediction of cognitive scores , and 68.75% accuracy ($F_1 = 66.67$) for the cognitive decline prediction task.

Index Terms: Cognitive Decline Detection, Affective Computing, Alzheimer's dementia, computational paralinguistics

1. Introduction

Alzheimer's dementia (AD) is a category of neurodegenerative diseases which entail long-term and usually gradual decrease of cognitive functioning. As the main risk factor for AD is age, it is increasingly prevalent in our ageing society. Due to the severity of the disease, institutions and researchers worldwide are investing considerably on dementia prevention, early detection and disease progression monitoring [1]. There is a need for cost-effective and scalable methods for early detection of AD and prediction of disease progression.

Methods for screening and tracking the progression of dementia traditionally involve cognitive tests such as the Mini-Mental Status Examination (MMSE) [2] and the Montreal Cognitive Assessment (MoCA) [3]. MMSE and MoCA are widely used because, unlike imaging methods, they are cheap, quick to administer and easy to score. Despite its shortcomings in specificity in early stages of dementia, the MMSE is still widely used [4]. The promise of speech technology in comparison to cognitive tests is twofold. First, speech can be collected passively, naturally and continuously throughout the day, gathering increasing data points while burdening neither the participant nor the researcher. Furthermore, the combination of speech technology and machine learning creates opportunities for automatic screening and diagnosis support systems for dementia. These opportunities need to be systematically assessed through common evaluation frameworks.

The ADReSS$_o$ Challenge aims to foster systematic comparison of approaches to the detection of cognitive impairment and decline based on spontaneous speech. As has been pointed out elsewhere [5, 6], the lack of common, standardised datasets and tasks has hindered the benchmarking of the various approaches proposed to date, resulting in a lack of translation of these speech based methods into clinical practice. The ADReSS$_o$ Challenge thus provides a forum for researchers working on approaches to cognitive decline detection based on speech data to test their existing methods or develop novel approaches on a new shared standardised dataset. The approaches that performed best on last year's dataset [5] employed features extracted from manual transcripts which were provided along with the audio data [7, 8]. The best performing method [8] made interesting use of pause and disfluency annotation provided with the transcripts. While this provided interesting insights into the predictive power of these paralinguistic features for detection of cognitive decline, extracting such features, and indeed accurate transcripts from spontaneous speech remains an open research issue. This year's ADReSS$_o$ (Alzheimer's Dementia Recognition through Spontaneous Speech *only*) tasks provide more challenging and improved spontaneous speech datasets, requiring the creation of models straight from speech, without manual transcription, though automatic transcription is encouraged.

The ADReSS$_o$ datasets are carefully matched so as to avoid common biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant (common in longitudinal datasets), variations in audio quality, and imbalances of gender and age distribution. The challenge defines three tasks:

1. an *AD classification task*, where participants were required to produce a model to predict the label (AD or non-AD) for a short speech session. Participants could use the speech signal directly (acoustic features), or attempt to convert the speech into text automatically (ASR) and extract linguistic features from this automatically generated transcript;
2. an *MMSE score regression task*, where participants were asked to create models to infer the patients' MMSE score based solely on speech data; and
3. a *cognitive decline (disease progression) inference task*, for prediction of changes in cognitive status over time, for a given speaker, based on speech data collected at baseline (i.e. the beginning of a cohort study).

These tasks depart from neuropsychological and clinical evaluation approaches that have employed speech and language [9] by focusing on prediction and recognition using spontaneous speech. Spontaneous speech analysis has the potential to enable novel applications for speech technology in longitudinal, unobtrusive monitoring of cognitive health [10], in line with the theme of this year's INTERSPEECH, "Speech Everywhere!".

This paper describes the ADReSS$_o$ dataset and presents baselines for all tasks, including feature extraction procedures and models for AD detection, MMSE score regression and cognitive decline prognosis.

http://dx.doi.org/10.21437/Interspeech.2021-1220

2. Related work

There has been increasing research on speech technology for dementia detection over the last decade. The majority of this research has focused on AD classification, but some of it targets MCI detection as well [6, 11]. These objectives are most closely related with our first task, namely, the AD classification task. Such related research includes the best performing models presented in the ADReSS challenge in 2020. These achieved an 85.45% [7] and 89.6% [8] accuracy in AD classification using acoustic features and text-based features extracted from manual transcripts, respectively. Another approach that builds on manually transcribed text and disfluency annotation, incorporating a time-based representation achieved a maximum 93.75% classification accuracy [12]. Classification based on acoustic features only was also attempted in [7], and obtained 76.85% accuracy with the IS10-Paralinguistics feature set (a low dimensional version of ComParE [13]) and Bag-of-Acoustic-Words (BoAW).

Few works rely exclusively on acoustic features or text features extracted through ASR. One of these achieved a 78.7% accuracy on a subset of the Cookie Theft task of the Pitt dataset, using different comprehensive paralinguistic feature sets and standard machine learning algorithms [14]. Another approach, using the complete Pitt dataset, obtained 68% accuracy with only vocalisation features (i.e. speech-silence patterns) [10]. A classification accuracy of 62.3% was reported in a study that used fully automated ASR features with a different dataset [15].

As regards the second task, regression over MMSE scores, there is less literature available and most of it has been presented in recent workshops [6]. Several of these works used the above mentioned Pitt dataset to extract different linguistic and acoustic features and predict MMSE scores. A recent study captured different levels of cognitive impairment with a multiview embedding and obtained a mean absolute error (MAE) of 3.42 [16]. Another study reported a MAE of 3.1 relying solely on acoustic features (a set of 811 features) [17]. Error scores as low as 2.2 (MAE) have been obtained, but relying on non-spontaneous speech data elicited in semantic verbal fluency (SVF) tasks [18].

Studies addressing disease progression are far less common. Notable in this category is [19], which incorporated a comprehensive set of features into a Bayesian network, reporting a MAE of 3.83 on prediction of MMSE scores across study visits. Two other studies account for disease progression in classification experiments. One study based on the speech features from the ISLE dataset achieved 80.4% accuracy for intra-subject change detection (i.e. distinguishing healthy participants who remained healthy from those who developed cognitive impairment) [20]. The second study used SVF scores to build a machine learning classifier able to predict changes from MCI to AD over a 4-year follow-up, with 84.1% accuracy [21].

3. The ADReSS$_o$ Datasets

We provided two distinct datasets for the ADReSS$_o$ Challenge: (1) a dataset consisting of speech recordings of Alzheimer's patients performing a category (semantic) fluency task [22] at their baseline visit, for prediction of cognitive decline over a two year period, and (2) a set of recordings of picture descriptions produced by cognitively normal subjects and patients with an AD diagnosis, who were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [23, 24].

The recorded data also includes speech from different experimenters who gave instructions to the patients and occasionally interacted with them in short dialogues. No transcripts were provided with either dataset, but segmentation of the recordings into vocalisation sequences with speaker identifiers [25] were made available for optional use. The ADReSS$_o$ challenge's participants were asked to specify whether they made use of these segmentation profiles in their predictive modelling. Recordings were acoustically enhanced with stationary noise removal and audio volume normalisation was applied across all speech segments to control for variation caused by recording conditions such as microphone placement.

The dataset used for AD and MMSE prediction was matched for age and gender so as to minimise risk of bias in the prediction tasks. We matched the data using a propensity score approach [26, 27] implemented in the R package MatchIt [28]. The dataset was matched according to propensity scores defined in terms of the probability of an instance of being treated as AD given covariates age and gender. All standardised mean differences for the age and gender covariates were < 0.001 and all differences for age^2 and two-way interactions between covariates were well below .1, indicating adequate balance. Propensity scores were estimated using a probit regression of the treatment on covariates age and gender as probit generated a better balanced than logistic regression. The matching is summarised in Figure 1, which shows the respective (empirical) quantile-quantile plots for the original and balanced datasets. A plot showing instances near the diagonal indicates good balance. The resulting dataset encompasses 237 audio files. These were split into training and test sets, with 70% of instances allocated to the former and 30% allocated to the latter. These partitions were generated so as to preserve gender and age matching.

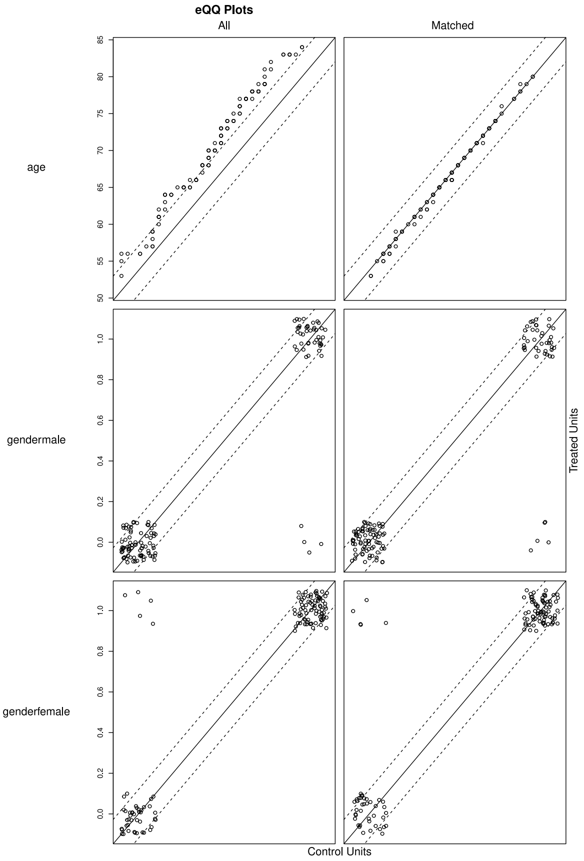

Figure 1: *Quantile-quantile plots for data before (left) and after matching (right) by age and gender.*

The dataset for the disease prognostics task (prediction of cognitive decline) was created from a longitudinal cohort study involving AD patients. The time period for assessment of disease progression spanned the baseline and year-two visits of the patients to the clinic. The task involves classifying patients into 'decline' or 'no-decline' categories, given speech collected at baseline as part of a verbal fluency test. Decline was defined as a difference in MMSE score between baseline and year-two greater than or equal 5 points This dataset has a total of 105 audio recordings split into training and test sets as with the diagnosis dataset (70%/30%).

Table 1 describes both datasets. These data, including manual transcrips which were not distributed for the challenge, are now available from DementiaBank [29].

Table 1: *Composition of the datasets.*

	Tasks 1 and 2		Task 3	
	AD	CN	Decline	No decline
Age	69.38 ($sd = 6.9$)	66.06 (6.3)	69.84 (9.3)	70.26 (8.5)
Men	35.2% ($n = 43$)	34.8% (40)	24.0% (6)	47.5% (38)
Women	64.8% (79)	65.2% (75)	76.0% (19)	52.5% (42)
MMSE	17.8 (5.5)	28.9 (1.2)	17.9 (4.6)	20.7 (5.2)
Duration	65.7s (38.6)	61.6s (26.9)	58.2s (16.0)	48.9s (19.5)

4. Data representation

4.1. Acoustic features

We applied a sliding window with a length of 100 ms on the audio files of the dataset with no overlap and extracted *eGeMAPS* features over such frames. The *eGeMAPS* feature set [30] resulted from an attempt to reduce the somewhat unwieldy feature sets above to a basic set of acoustic features based on their potential to detect physiological changes in voice production, as well as theoretical significance and proven usefulness in previous studies. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, totalling 88 features per 100ms frame. We then applied the active data representation method (ADR) [14] to generate a frame level acoustic representation for each audio recording. The ADR method has been used previously to generate large scale time-series data representation. It employs self-organising mapping to cluster the original acoustic features and then computes second-order features over these cluster to extract new features (see [14] for details). Note that this method is entirely automatic in that no speech segmentation or diarisation information is provided to the algorithm.

4.2. Linguistic Features

We used the Google Cloud-based Speech Recogniser to automatically transcribe the audio files. The overall mean word error rate (WER) for these transcripts was 60 (±20.9), computed against manual transcripts using NIST's Sclite tool [31]. A potential explanation for this relatively high WER is the fact that AD speech often involves an imprecise use of language (e.g., ungrammatical sentences), which current ASR systems are not optimised to handle [6]. The ASR transcripts were converted into CHAT format which is compatible with CLAN [32], a set of programs that allows for automatic analysis of a wide range of linguistic and discourse structures. Next, we used the automated MOR function to assign lexical and morphological descriptions to all the words in the transcripts. Then, we used two commands: EVAL which creates a composite profile of 34 measures, and FREQ to compute the Moving Average Type Token Ratio [33]. For comparison, we also applied the same procedure to generate linguistic features from manual transcripts.

5. Diagnosis baseline

5.1. Task 1: AD Classification

The AD classification experiments were performed using five different methods, namely: decision trees (DT, with leaf size optimised through grid search within a range of 1 to 20), nearest neighbour (KNN, where the K parameter is optimised through grid search from 1 to 20), linear discriminant analysis (LDA), Tree Bagger (TB, with 50 trees, and leaf size optimised through grid search from 1 to 20), and support vector machines (SVM), with a linear kernel, box constraint optimised by grid search between 0.1 to 1.0, and a sequential minimal optimisation solver.

The results for accuracy in the AD vs Control (CN) classification task are summarised in Table 2. As indicated in boldface, the best classifier in leave-one-subject-out cross validation (CV) was DT, achieving 78.92% and 72.89% accuracy using acoustic and linguistic features, respectively. On the test set, however, the results were reversed, with linguistic features producing an overall best accuracy of 77.46%, with the SVM classifier. Late fusion of the acoustic and linguistic models improves the accuracy on the test set further to 78.87% (Figure 2, left). Also shown on the table are the results for linguistic features generated from manual transcripts. One can see that manual transcription leads to overall improvements in CV and to a slight punctual improvement on testing. This suggests that even though the ASR WER was high (60), the automatically generated transcripts still contribute appreciably to the models.

Table 2: *Task1: AD classification accuracy on CV and test data, for fully automatic acoustic and ASR features. Best results shown in boldface. Performance for features from manual transcription (Transcript) are shown in italics for comparison.*

		LDA	DT	SVM	TB	KNN	mean (sd)
CV	Acoustic	62.65	**78.92**	69.28	65.06	65.06	68.19 (6.4)
	ASR	72.29	**72.89**	72.89	75.90	65.06	71.81 (4.0)
	Transcript	*80.12*	*77.71*	*80.72*	*76.51*	*69.28*	*76.87 (4.6)*
Test	Acoustic	50.70	60.56	**64.79**	63.38	53.52	58.59 (6.2)
	ASR	76.06	74.65	**77.46**	73.24	59.15	72.11 (7.4)
	Transcript	*76.06*	*67.61*	*78.87*	*66.20*	*60.56*	*69.86 (7.5)*

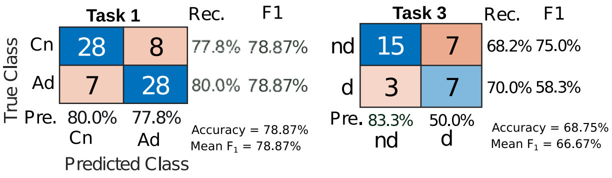

Figure 2: *Late (decision) fusion of the best results of acoustic and linguistic models for Task 1 (left) and Task 3 (right). Precision (Pre) , recall (Rec), accuracy (Accu) and mean F_1 scores are shown on the margins.*

5.2. Task 2: MMSE prediction

For this task we used five types of regression models: linear regression (LR), DT, with leaf size 20 and CART algorithm, support vector regression (SVR, with a radial basis function kernel, box constraint of 0.1 and sequential minimal optimisation solver), Random Forest regression ensembles (RF), and Gaussian process regression (GP, with a squared exponential kernel). The regression methods are implemented in MATLAB [34].

Table 3: *Task2: MMSE score prediction error scores (RMSE) for CV and test data. Results for manual transcripts in italics.*

		LR	DT	SVR	RF	GP	mean (sd)
	Acoustic	6.88	6.88	6.96	7.89	6.71	7.06 (0.47)
CV	ASR	6.65	5.92	6.42	7.02	6.50	6.50 (0.40)
	Transcript	*5.77*	*6.20*	*5.75*	*6.94*	*5.52*	*6.04 (0.56)*
	Acoustic	6.23	6.47	**6.09**	8.18	6.81	6.75 (0.84)
Test	ASR	5.87	6.24	**5.28**	6.94	5.43	5.95 (0.67)
	Transcript	*4.49*	*6.06*	*4.65*	*6.07*	*4.35*	*5.12 (0.87)*

The results are summarised in Table 3. As with classification, DT regression outperformed the other models in CV, with ASR linguistic features outperforming acoustic ADR features. This trend persisted in the test set, with linguistic features producing a minimum RMSE of 5.28 in a SVR model. We then fused the best results of linguistic and acoustic features and took a weighted mean, finding the weights through grid search on the validation results, which resulted in an improvement (6.37) on the validation dataset. We then used the same weights to fuse the test results and obtained an RMSE of 5.29 ($r = 0.69$). For this task, manual transcripts produced substantial improvements (a 14% error reduction on average, with as much as 18% improvement for the best model).

6. Prognosis baseline (Task 3)

We tested the same classification methods used in Task 1 for the task of identifying patients who went on to exhibit cognitive decline within two years of the baseline visit in which the speech samples used in our models were taken. The acoustic and linguistic features were generated as described in Section 4. The results of this prediction task are summarised in Table 4. As the classes for this task are imbalanced we report average F_1 results rather than accuracy, Once again DT performed best on CV, but the F_1 results for the test set was considerably lower, reaching only 66.67% for linguistic and 61.02% for acoustic features.

Table 4: *Task3: cognitive decline progression results (mean of $F_1 Score$) for leave-one-subject-out CV and test data.*

		LDA	DT	SVM	TB	KNN	mean (sd)
	Acoustic	59.89	84.94	55.64	63.85	65.92	66.05 (11.27)
Val	ASR	55.19	76.52	45.24	63.10	55.25	59.06 (11.64)
	Acoustic	**61.02**	53.62	40.74	40.74	38.46	46.91 (9.89)
test	ASR	54.29	**66.67**	40.74	56.56	39.62	51.58 (11.41)

As before, we fused the predictions of the best models for each feature type, hoping that the diversity of models might improve classification. The confusion matrix for the fusion model is shown in Figure 2, right. This time, however, decision fusion did not yield any improvement in accuracy, although sensitivity (recall) improved from 40% to 70% for patients whose cognitive function declined.

7. Discussion

The AD classification baseline yielded a maximum accuracy of 78.87% on the test set, through the fusion of models based on linguistic and acoustic features. Despite the fact that the ASR transcripts had relatively high WER, linguistic features contributed considerably to the predictions. The overall baseline results for this task are in fact comparable to results obtained for similar picture description data using manual transcripts (see Section 2). We further tested this observation by generating language models out of manual transcripts, and verified that the accuracy improvements by those models was slight. The good performance of the classifiers on ASR data is somewhat puzzling. We speculate that as WER varies widely across the recordings, ASR quality itself might have been detected by the models. As deterioration in speech quality (low loudness and intelligibility) correlates positively with AD and negatively with the ASR performance, poor performance might be indirectly providing an AD predictor. This warrants further investigation. Finally, DT classifiers performed well on the CV experiments, but accuracy decreased on the test set, indicating probable overfitting. Overall, however, all models proved fairly robust.

A similar picture was observed in the MMSE regression task. ASR generated linguistic features contributed appreciably to the prediction, despite transcription errors. In this case, however, late fusion only improved the RMSE score in CV; the test set RMSE remained practically unchanged. Also of note is the fact that on this task manual transcription would have substantially reduced RMSE, showing that fully automatic processing still faces challenges in predicting subtler cognitive differences.

The prognosis task proved to be the most difficult. The CV results varied considerably among models, specially the linguistic models ($sd = 11.64$). The test set results were also varied, reaching a maximum F_1 score of 66.67%, even when the best model predictions were fused. Although the acoustic features produced the best classification results in CV ($F_1 = 66.05\%$ vs 59.06% for linguistic features), these results were not born out by test set evaluation, suggesting that the acoustic features made the classifiers more prone to overfitting. It is possible that this could be mitigated by training the acoustic feature extractor (ADR) on a larger set of off-task recordings (data augmentation) and fine tuning the resulting model on the ADReSS$_o$ data.

8. Conclusions

The ADReSS$_o$ Challenge is the first shared task to target cognitive status prediction using raw, non-annotated a non-transcribed speech, and to address prediction of changes in cognition over time. We believe this moves the speech processing and machine learning methods one step closer to the real-world of clinical applications. A limitation the AD classification and the MMSE regression tasks share with most approaches to the use of these methods in dementia research is that they provide little insight into disease progression. This has been identified as the main issue hindering translation of these technologies into clinical practice [6] and, hence, preclinical modelling emerges as clear avenue for future research [35]. However, these tasks remain relevant in application scenarios involving automatic cognitive status monitoring, in combination with wearable and ambient technology. The addition of the progression task should open avenues for relevance also in more traditional clinical contexts.

9. Acknowledgements

Work funded by the European Union's Horizon 2020 programme, under grant agreements 769661 (SAAM) and 825153 (EMBEDDIA). SG is supported by the Medical Res Council.

10. References

[1] K. Ritchie, I. Carrière, L. Su, J. T. O'Brien, S. Lovestone, K. Wells, and C. W. Ritchie, "The midlife cognitive profiles of adults at high risk of late-onset Alzheimer's disease: The PREVENT study," *Alzheimer's & Dementia*, vol. 13, no. 10, pp. 1089–1097, 2017.

[2] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "“mini-mental state”: a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[3] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[4] I. Arevalo-Rodriguez, N. Smailagic, M. R. i Figuls, A. Ciapponi, E. Sanchez-Perez, A. Giannakou, O. L. Pedraza, X. B. Cosp, and S. Cullum, "Mini-mental state examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)," *Cochrane Database of Systematic Reviews*, no. 3, 2015.

[5] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[6] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.

[7] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated Screening for Alzheimer's Dementia Through Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 2222–2226.

[8] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease," in *Proc. Interspeech 2020*, 2020, pp. 2162–2166.

[9] V. Taler and N. A. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: A comparative review," *Journal of Clinical and Experimental Neuropsychology*, vol. 30, no. 5, pp. 501–556, 2008.

[10] S. Luz, "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *Computer Based Medical Systems*. IEEE Press, 2017, pp. 45–46.

[11] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic alzheimer's disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.

[12] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 299, 2021.

[13] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Procs. of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2010, pp. 2794–2797.

[14] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[15] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting Signs of Dementia Using Word Vector Representations." in *Interspeech*, 2018, pp. 1893–1897.

[16] C. Pou-Prom and F. Rudzicz, "Learning multiview embeddings for assessing dementia," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2812–2817.

[17] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting alzheimer's disease severity in longitudinal acoustic data," in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, 2017, pp. 57–61.

[18] N. Linz, J. Tröger, J. Alexandersson, M. Wolters, A. König, and P. Robert, "Predicting dementia screening and staging scores from semantic verbal fluency performance," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 719–728.

[19] M. Yancheva, K. C. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 134–139.

[20] J. Weiner and T. Schultz, "Detection of Intra-Personal Development of Cognitive Impairment From Conversational Speech," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5.

[21] D. G. Clark, P. M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R. M. Dukes, P. Kapur, T. P. DeRamus, and L. G. Apostolova, "Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 2, pp. 113–122, 2016.

[22] A. L. Benton, "Differential behavioral effects in frontal lobe disease," *Neuropsychologia*, vol. 6, no. 1, pp. 53–60, 1968.

[23] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease," *Archives of Neurology*, vol. 51, no. 6, p. 585, 1994.

[24] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination – Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[25] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," in *Resources for processing of linguistic, paralinguistic and extra-linguistic data from people with various forms of cognitive impairment*, D. Kokkinakis, Ed. ELRA, May 2018, pp. 35–42.

[26] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 04 1983.

[27] D. B. Rubin, "Matching to remove bias in observational studies," *Biometrics*, vol. 29, no. 1, pp. 159–183, 1973.

[28] D. Ho, K. Imai, G. King, and E. A. Stuart, "Matchit: Nonparametric preprocessing for parametric causal inference," *Journal of Statistical Software, Articles*, vol. 42, no. 8, pp. 1–28, 2011. [Online]. Available: https://www.jstatsoft.org/v042/i08

[29] DementiaBank, https://dementia.talkbank.org/, acessed June 2021.

[30] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[31] NIST, "SCTK, the NIST scoring toolki," https://github.com/usnistgov/SCTK, 2021, accessed 28-3-21.

[32] B. MacWhinney, "Tools for analyzing talk part 2: The CLAN program," 2017, pittsburgh, PA: Carnegie Mellon University. [Online]. Available: http://talkbank.org/manuals/CLAN.pdf

[33] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of quantitative linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[34] MATLAB, *version 9.6 (R2019a)*. Natick, Massachusetts: The MathWorks Inc., 2019.

[35] S. de la Fuente, C. Ritchie, and S. Luz, "Protocol for a conversation-based analysis study: Prevent-ED investigates dialogue features that may help predict dementia onset in later life," *BMJ Open*, vol. 9, no. 3, 2019.