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

Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge

P. A. Pérez-Toro[1,2] ⋆, *S. P. Bayerl*[3] ⋆, *T. Arias-Vergara*[1,2,4], *J. C. Vasquez-Correa*[1,2], *P. Klumpp*[1], *M. Schuster*[4], *E. Nöth*[1], *J. R. Orozco-Arroyave*[1,2], *K. Riedhammer*[3]

[1]Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany
[2]Facultad de Ingeniería. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[3]Technische Hochschule George Simon Ohm, Nürnberg, Germany
[4]Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany

paula.andrea.perez@fau.de, sebastian.bayerl@ieee.org

Abstract

Alzheimer's Disease (AD) results from the progressive loss of neurons in the hippocampus, which affects the capability to produce coherent language. It affects lexical, grammatical, and semantic processes as well as speech fluency. This paper considers the analyses of speech and language for the assessment of AD in the context of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) 2021 challenge. We propose to extract acoustic features such as X-vectors, prosody, and emotional embeddings as well as linguistic features such as perplexity, and word-embeddings. The data consist of speech recordings from AD patients and healthy controls. The transcriptions are obtained using a commercial automatic speech recognition system. We outperform baseline results on the test set, both for the classification and the Mini-Mental State Examination (MMSE) prediction. We achieved a classification accuracy of 80% and an RMSE of 4.56 in the regression. Additionally, we found strong evidence for the influence of the interviewer on classification results. In cross-validation on the training set, we get classification results of 85% accuracy using the combined speech of the interviewer and the participant. Using interviewer speech only we still get an accuracy of 78%. Thus, we provide strong evidence for interviewer influence on classification results.

Index Terms: Alzheimer's Disease, Speech Analysis, Natural Language Processing, Speaker Modeling, Emotional Modeling

1. Introduction

Alzheimer's Disease (AD) is the most prevalent neurodegenerative disease and the most common form of dementia [1]. It is characterized by progressive dementia, neurological degeneration, and death of brain cells. AD symptoms include memory, behavioral, and psychological impairments. The deterioration of cognitive functions also leads to communication deficits, i.e., the capability to produce coherent language [2]. Abnormalities in language production of AD patients are caused by the difficulty to access semantic information intentionally, which affects speech fluency [3]. A standard scale to evaluate the cognitive function of AD patients is the Mini-Mental State Examination (MMSE) [4]. It is a 30-point scale that accounts for language production, immediate memory, naming, and spatial attention. Scores of over 24 indicate normal cognition.

In last year's Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge a dataset comprised of recordings from the Dementia Bank was provided [5]. The challenge included two tasks; classification of dementia and prediction of the MMSE score. In [6], the authors used Term Frequency-Inverse Document Frequency (TF-IDF) and Mel Frequency Cepstral Coefficients (MFCCs). Their best results on the test set were achieved using late fusion, leading to an accuracy of 77.8% for the classification task and a Root Mean Square Error (RMSE) of 4.44 for the prediction task. A finetuning of Bidirectional Encoder Representations from Transformers (BERT) embeddings was performed by [7]. Achieving an accuracy of 83% in the classification task and an RMSE of 4.56 on the test set. The use of state-of-the-art speaker recognition (X-vectors) and word-embeddings (BERT) techniques were employed in [8]. 81% accuracy on the test set was obtained by combining both embeddings. In [9], classical speech paralinguistics features, as well as acoustic and BERT embeddings, achieved accuracies of up to 85% using linguistic and 76% using speech features. For the prediction task linguistic features obtained an RMSE of 4.30, while using speech features achieved an RMSE of 5.92.

Dementia Bank [10] provides a publicly available dataset with recordings from AD patients and Healthy Control (HC) subjects. It has led to a large body of prior work. An n-gram based approach combined with recurrent cells to classify AD patients and HC subjects was proposed in [11]. Fraser et al. use a correlation-based feature ranking technique to select from language, psycho-linguistic, and acoustic features such as energy, periodicity, and vocabulary richness [12].

Research on automatic assessment of AD utilizing data outside of dementia bank can be found in [13] where the authors used an Automatic Speech Recognition (ASR) system to extract speech features and linguistic features from transcripts to discriminate between AD patients and HC subjects as well as people suffering from Mild Cognitive Impairments (MCI). Pérez et al. [14] use articulation, prosody, X-vectors, and state-of-the-art word embeddings to classify genetic and early-onset AD.

Our contributions are:

- Improvement over baseline results for the AD classification and MMSE prediction tasks of the ADReSSo 2021 Challenge [15] using novel emotional embeddings.
- Evaluation of linguistic and acoustic features, as well as

⋆equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1589

multi-modal fusion-approaches for the classification of AD, using ASR and speaker diarization.

- Providing evidence for the influence of the Interviewer (INV) in the task for automatic assessment of AD

2. Data

The dataset in this work was created by the organizers of the Interspeech ADReSSo 2021 challenge [15]. The dataset consists of 166 recordings (87 AD, 79 HC) for training and 71 for testing. All participants were native English speakers who were asked to describe the cookie theft picture [16]. The recordings are matched for age and gender and have been acoustically enhanced and normalized. In addition to the recordings, speaker segmentation information was provided.

For this study we, obtained transcriptions using a commercial state of the art ASR service[1]. The speech signals were denoised to improve the quality of the recordings using the proposed model in [17].

3. Methods

In this section, we briefly describe the acoustic and linguistic features used in this study. Extraction methods and implementation details can be found in the accompanying repository. [2]

3.1. Acoustic Features

3.1.1. X-vectors

X-vectors are DNN-embeddings that were originally used in speaker recognition and diarization tasks [18, 19]. They have been shown to also work in several paralinguistic tasks such as emotion recognition from speech [20], the detection of Parkinson's disease [21], and AD [22]. [20] describes the influence of emotions on speaker recognition using X-vectors. Their study provides evidence that there is additional information encoded besides speaker information. Their independence of the actual AD training data as well as their robustness to noise and challenging acoustic conditions make X-vectors a good fit for acoustic-only approaches for AD assessment. Another advantage is their ability to map variable-length utterances to fixed-length embeddings.

In our experiments, we use an X-vector system based on a Time Delay Neural Network (TDNN) as proposed in [19]. The TDNN is trained using the Kaldi-toolkit and the VoxCeleb2 corpus [23]. Training and implementation details are described in [19]. The recipe used to train the X-vector system is publicly available.[3] Training of the X-vector system relies on data augmentation to adapt to difficult acoustic conditions, thus making them robust to noise and other channel effects. X-vectors are extracted for every 1.5s window with a minimal segment size of 0.5s. The embeddings are mean normalized and their length is reduced to 200 dimensions using Linear Discriminant Analysis (LDA).

3.1.2. Prosody

The extracted features are based on speech rates and energy and the Fundamental Frequency (F_0) contours, where chunks of 40 ms were taken. The energy contour is computed over the voiced and unvoiced segments. For the F_0 only the voiced segments were considered. The tilt and the mean square error were computed from the contours. From these descriptors, six statistical functionals were computed (mean, standard deviation, kurtosis, skewness, minimum, and maximum) per utterance. Additionally, features based on duration measures considering the voiced and unvoiced segments were also considered. A total of 91 descriptors were extracted.

3.1.3. Voice Activity Detection Features

Duration ratios were extracted using an energy-based Voice Activity Detection (VAD) algorithm. The considered features were defined by; (1) number of pauses per second, (2) number of speech segments per second, (3) ratio between the number of speech segments and pauses, (4) six functionals (mean, standard deviation, kurtosis, skewness, minimum, and maximum) for the duration of the speech segments, and (5) the same six functionals for the duration of the pauses.

3.1.4. Pre-trained model based on the PAD emotional model

The *"Pleasure, Arousal, and Dominance emotional model"* [24] (PAD) leads to represent different emotions in a multidimensional space, where they can be either pleasant-unpleasant (valence), calm-agitated (arousal), or dominant-submissive (dominance). Our approach aims to capture similar aspects related to the emotions, mood, and affective states in AD patients, since the reduced ability of the emotional perception in AD caused by the memory loss may induce the appearance of apathy and depression according to some studies [25, 26]. We trained three models to address three classification problems using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [27]: (1) active vs. passive arousal (accuracy=67%), (2) positive vs. negative valence (accuracy=88%), and (3) strong vs. weak dominance (accuracy=80%).

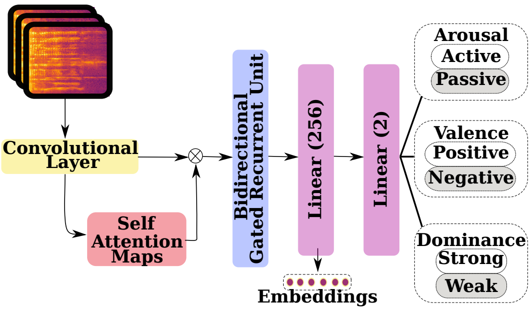

Figure 1: *General architecture of the three pre-trained models based on the PAD emotional model. It consists of four parts: (1) a CNN of 8 filters with a kernel size of (1,3), max pooling (1,2), a batch normalization layer, and a leaky ReLU activation, (2) a self-attention map layer, (3) a Bidirectional GRU (Bi-GRU) of 2 stacked layers with 128 hidden units, a batch normalization layer, and (4) two linear layers with 256 units for the embedding features and 2 units for the classification step.*

Our proposed model (see Figure 1) consists of a multichannel input formed by 3 log-Mel spectrograms with different resolutions (16 ms, 25 ms, and 45 ms) and considering sequences of 500 ms. It aims to model different aspects related to articulation and prosody information by combining Convolutional Neural Networks (CNN) and Gated Recurrent Units

[1] Amazon web services (AWS) transcribe.
[2] https://git.io/JnUJd
[3] https://git.io/JnUUq

(GRU). The output of the embedding layer is used to extract features (transfer knowledge) for the ADReSSo data, with the assumption that some affective patterns can appear in AD [25, 26].

3.2. Linguistic Features

3.2.1. Word-Embeddings

These methods allow the words in a corpus to be represented as lower-dimensional feature vectors to better model the context. BERT and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) are based on the encoder part from the "Transformers" method [28] that maps an input sequence into lower dimensional feature vectors.

BERT consists of a Masked Language Model (MLM), which predicts a small number of words that have been masked out of the input [29]. These models have the advantage of being bidirectional, which considers the use of the previous and the following words. As opposed to BERT, ELECTRA instead of the MLM, performs a pre-training task called Replaced Token Detection (RTD) [30]. Instead of replacing some words with the token *"[Mask]"* as in BERT, RTD corrupts some words with generated incorrect words to discriminate between *"real"* and *"fake"* input words, similar to adversarial models. We considered BERT-Base and ELECTRA-Base models trained with BooksCorpus and the English Wikipedia. The last layer (768 units) is taken as the word-embedding representation in both methods. The mean of the overall word-embeddings is computed for the classification task, while four functionals (mean, standard deviation, skewness, and kurtosis) are computed for the regression task [31].

3.2.2. Perplexity

Perplexity (PPL) is the inverse probability of the test set, normalized by the number of words. PPL is a measure of how well a Language Model (LM) predicts a sample. For a sequence of words $W = w_1, w_2, ..., w_N$, perplexity is computed by

$$PPL(W) = P(w_1 w_2 ... w_N)^{\frac{-1}{N}} \tag{1}$$

Low perplexity indicates that a text can be well predicted by a LM that was trained on a different text, meaning predictable results are considered to be better than randomness. The cookie theft picture can easily be described by a healthy person. This limited task is expected to lead to a small, closed vocabulary and thus to similar n-grams. This is the case if texts of both training and test data are describing what is in the picture similarly and are coherent. As shown by Wankerl et al., AD patients tend to describe the picture in unforeseen ways and divert from the actual task. They stumble frequently and repeat themselves by using different formulations [32]. We, therefore, adopt the n-gram LM-based evaluation of PPL using two LMs, $M_{\text{alzheimer}}$ and M_{control} described in [32]. We acquire two PPL values $p_{\text{alzheimer}}$ and p_{control} as well as their difference $p_{\text{diff}} = p_{\text{alzheimer}} - p_{\text{control}}$ and use those as features in our experiments.

N-gram LMs can be quickly computed and evaluated for small amounts of training data such as the challenge data. Tools for computing n-grams are included in speech recognition toolkits and readily available. In our experiments, we use the popular SRILM toolkit to compute two Bi-gram LMs, $M_{\text{alzheimer}}$ and M_{control} [33]. We account for out of vocabulary (OOV) words by mapping them to a special token and using Witten-Bell smoothing. The resulting model M_{control} has 953 uni-grams and 3875 bi-grams whereas the $M_{\text{alzheimer}}$ model has 906 uni-grams and 3548 bi-grams when computed on the training data.

3.3. Optimization, Classification, and Regression

A Radial Basis Function-Support Vector Machine (RBF-SVM) was used as a classifier for the diagnosis task. The optimal parameters of the RBF-SVM were found through a grid search where $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$. The regression task for the prediction of the MMSE was performed by using a Linear Regression (LR) model. Other regressors such as linear Support Vector Regression (SVR) and an RBF-SVR, were discarded since the best performance was obtained using LR for our set of features. The validation for all experiments followed a 5-Fold Cross-Validation (CV) strategy using the training set provided by the challenge. For the classification and regression an early fusion strategy was applied by merging sets of features before performing the classification/regression and making the final decision.

4. Experiments and Results

The experiments consider two tasks in the context of the ADReSSo challenge: (1) Classification of AD/HC and (2) the prediction of the MMSE score. The challenge results for both tasks are presented in Table 1. The baseline results were provided by the challenge and were evaluated considering a Leave One Speaker Out (LOSO) - CV strategy and using the test set. For comparison purposes, we used a CV strategy to evaluate our models only using the train/development set since the test labels were not provided. Similar results for classification using linguistics (acc=81.33%) and acoustics (acc=81.13%) were found. However, the performance increases by 5% points using an early fusion strategy combining acoustics and linguistics. For the prediction task linguistics obtain the most accurate results w.r.t. each modality separately (RMSE=5.14, ρ=0.68). Besides, the combination of acoustics and linguistics improves the prediction with an RMSE of 4.86 and a ρ of 0.72. The reported test results were obtained by submitting our best combinations of features to the challenge. The combination of modalities provides higher results for the classification task, while linguistic features are more accurate in the prediction of the MMSE.

Table 2 shows the results for classification considering the complete recordings (participant and INV together), the segmented recordings for the participant only, and the segmented recordings for the INV only. Experiments considering INV-only speech were performed on a subset of 158 samples, as no labeled INV segments were present in 8 of the samples. The segmentation was performed according to the timestamps provided by the challenge. The results are computed following a CV strategy in the train/development set. In general, the most accurate results are obtained using the unsegmented recordings, where the combination of modalities yields an accuracy of 85.54%. Similar results are achieved considering only the participant and only the INV, which may indicate that the INV is influencing the task in order to better interact with the patient.

The prediction results of the MMSE considering the unsegmented and the segmented recordings are shown in Table 3. Although the best performance is obtained by using the unsegmented recordings, the results are close to those obtained only using the participant's speech. It can indicate that the INV adapts differently to subjects with AD. However, the INV does not seem to directly influence the prediction of the MMSE because she/he cannot intuitively assume the severity of the disease. Unfortunately, the information about how many different INVs were involved was not provided.

Table 1: *ADReSSo challenge results for the classification and prediction task. A, V, and D are the arousal, valence, and dominance embeddings. Prosody is represented by P, BERT by B, ELECTRA by E, perplexity by PPL, and X-vectors by Xvec.*

	Classification					Prediction		
	Features	**F1**	**Acc**	**Sens**	**Spec**	**Features**	**RMSE**	ρ
CV in the training set								
Acoustic	Xvec + D	0.81	81.13	81.01	81.61	P + Xvec + V + D	6.27	0.51
Linguistic	PPL + B	0.81	81.33	83.54	79.31	E	5.14	0.68
Fusion	P + A + D + B	**0.86**	**85.54**	**88.61**	**82.76**	P + Xvec + V + D + B + E	**4.86**	**0.72**
Baseline	Acoustic	–	78.92	–	–	Acoustic	6.88	–
	Linguistic	–	72.89	–	–	Linguistic	5.92	–
Test								
Acoustic	Xvec + D	**0.67**	**67.61**	**75.00**	**60.00**	P + Xvec + V + D	**5.35**	–
Linguistic	PPL + B	**0.78**	**78.87**	**97.22**	**60.00**	PPL + B	**4.56**	–
Fusion	P + A + D + B	**0.80**	**80.28**	**88.89**	**71.43**	P + Xvec + V + D + B + E	**4.79**	–
Baseline	Acoustic	–	64.79	–	–	Acoustic	6.09	–
	Linguistic	–	77.46	–	–	Linguistic	5.28	–
	Late Fusion	0.79	78.87	80.00	77.78	Late Fusion	5.29	0.69

F1: F1-Score. **Acc**: Accuracy. **Sens**: Sensitivity. **Spec**: Specificity. **RMSE**: Root Mean Square Error. **ρ:** Spearman's correlation. Acc, Sens, and Spec are given in [%].

Table 2: *Cross-validation classification results on the training set. Using the combined speech of interviewer and participant, and either participant or interviewer speech.*

	Features	**F1**	**Acc**	**Sens**	**Spec**
Complete Recording					
Acoustic	Xvec + D	0.81	81.13	81.01	79.31
Linguistic	PPL + B	0.81	81.33	83.54	79.31
Fusion	P + A + D + B	0.86	85.54	88.61	82.76
Participant Only Speech					
Acoustics	P + A	0.75	74.70	65.82	82.76
Linguistics	PPL + B + E	0.78	78.31	77.22	79.31
Fusion	A + V + E	0.82	82.53	83.54	81.61
Interviewer Only Speech					
Acoustics	VAD+P+V	0.78	77.78	74.24	80.46
Linguistics	B+E	0.71	70.59	71.21	70.12
Fusion	Xvec+V+E	0.77	76.47	72.72	79.31

F1: F1-Score. **Acc**: Accuracy. **Sens**: Sensitivity. **Spec**: Specificity. Acc, Sens, and Spec are given in [%].

Table 3: *Cross-validation MMSE results on the training set. Using the combined speech of interviewer and participant, and either participant or interviewer speech.*

	Features	**RMSE**	ρ
Complete Recording			
Acoustic	P + Xvec + V + D	6.27	0.51
Linguistic	PPL + B	5.16	0.68
Fusion	P + Xvec + V + D + B + E	4.86	0.72
Participant Only Speech			
Acoustic	VAD + P + Xvec + A + V	6.40	0.42
Linguistic	E	5.14	0.68
Fusion	VAD + Xvec + V + E	4.87	0.70
Interviewer Only Speech			
Acoustic	VAD + Xvec+D	6.02	0.51
Linguistic	PPL	10.89	0.37
Fusion	Xvec + V + D + B	5.65	0.53

RMSE: Root Mean Square Error. **ρ:** Spearman's correlation.

5. Discussion and Conclusions

This study proposed and experimentally evaluated a methodology for the automatic assessment and classification of AD. Our method leverages ASR and a combination of classical as well as state-of-the-art speaker- and word-embeddings in multimodal classification and regression models. While we could show that we can improve over the baseline on both the cross-validation and the test set (see Table 1), we do not consider these as the main findings of our study.

To us, the main finding of our study is evidence of the influence of the INV on the results of the classification task. We observed that classification results improved whenever the speech of the INV is involved. At the same time not hurting the performance in the MMSE prediction and by itself still getting very close to baseline results using INV speech only. This may happen since INVs intuitively adapt their behavior to better communicate/interact with the AD patient. Some studies reveal that therapists, physicians, health care providers, and caregivers use different interaction strategies to enhance communication [34, 35, 36]. However, in the case of the data provided, we do not know whether the INV knows about the participant's condition beforehand. A variable of INV behavior is the number of interactions (INV labeled segments) and their duration per sample. We performed a Kruskal-Wallis ($p \ll 0.05$) test to compare the INV duration and the number of INV labeled segments between HCs and AD patients. This leads to the rejection of the null-hypothesis in both cases, i.e., the behavior of the INV is distinctively different when talking to either an HC or an AD patient.

It is important to be mindful w.r.t. the results and to further investigate our observation. The best result in the CV for INV-only speech could be achieved using acoustic features only. This implies that there might be something in the acoustic conditions that adds a bias to the dataset. This could mean that we in part classify acoustic conditions rather than AD.

We suggest further research into two directions. Checking the dataset for inherent bias in acoustic conditions, while at the same time exploring other features, fusion techniques, and data modeling methods. To check the validity of the proposed methods, other datasets need to be used.

6. Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 766287, and partially funded by the project PI 2019-24110 from University of Antioquia. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. Sebastian Bayerl is supported by the Bayerisches Wissenschaftsforum (BayWISS). Additionally, we thank the company Pratech Group S.A.S. for their support in the development of this study.

7. References

[1] M. J. Prince, *World Alzheimer Report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015.

[2] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[3] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[4] M. F. Folstein *et al.*, "The mini-mental state examination," *Archives of general psychiatry*, vol. 40, no. 7, pp. 812–812, 1983.

[5] S. Luz *et al.*, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2571

[6] M. Martinc *et al.*, "Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.

[7] A. Balagopalan *et al.*, "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection," *Proc. Interspeech 2020*, 2020.

[8] R. Pappagari *et al.*, "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[9] M. S. S. Syed *et al.*, "Automated Screening for Alzheimer's Dementia through Spontaneous Speech," *Proc. Interspeech 2020*, pp. 1–5, 2020.

[10] J. T. Becker *et al.*, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[11] J. Fritsch *et al.*, "Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models," in *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5841–5845.

[12] K. C. Fraser *et al.*, "Detecting late-life depression in Alzheimer's disease through analysis of speech and language," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 1–11.

[13] G. Gosztolya *et al.*, "Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[14] P. A. Pérez-Toro *et al.*, "Acoustic and Linguistic Analyses to Assess Early-Onset and Genetic Alzheimer's Disease," in *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (IN PRESS)*, 2021, pp. 1–5.

[15] S. Luz *et al.*, "Detecting cognitive decline using speech only: The ADReSSo Challenge," in *Submitted to Interspeech 2021*, 2021. [Online]. Available: https://edin.ac/31eWsjp

[16] H. Goodglass *et al.*, "Cookie Theft picture," *Boston diagnostic aphasia examination. Philadelphia, PA: Lea & Febiger*, 1983.

[17] H. Schröter *et al.*, "CLC: Complex Linear Coding for the DNS 2020 Challenge," *arXiv preprint arXiv:2006.13077*, 2020.

[18] D. Snyder *et al.*, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[19] G. Sell *et al.*, "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge," in *Proc. Interspeech 2018*, 2018, pp. 2808–2812. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1893

[20] R. Pappagari *et al.*, "X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition," in *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7169–7173.

[21] L. Moro-Velazquez *et al.*, "Using X-Vectors to Automatically Detect Parkinson's Disease from Speech," in *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 1155–1159.

[22] R. Haulcy *et al.*, "Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech," *Frontiers in Psychology*, vol. 11, p. 3833, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137

[23] J. S. Chung *et al.*, "VoxCeleb2: Deep Speaker Recognition," *arXiv:1806.05622 [cs, eess]*, Jun. 2018, arXiv: 1806.05622. [Online]. Available: http://arxiv.org/abs/1806.05622

[24] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[25] J. D. Henry *et al.*, "Emotion experience, expression, and regulation in Alzheimer's disease," *Psychology and aging*, vol. 24, no. 1, p. 252, 2009.

[26] M. S. Goodkind *et al.*, "Emotion regulation deficits in frontotemporal lobar degeneration and Alzheimer's disease," *Psychology and aging*, vol. 25, no. 1, p. 30, 2010.

[27] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[28] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[29] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[30] K. Clark *et al.*, "ELECTRA: Pre-training Text Encoders as Discriminators Rather than Generators," *arXiv preprint arXiv:2003.10555*, 2020.

[31] P. A. Perez-Toro, "PauPerezT/WEBERT: Word Embeddings using BERT," https://doi.org/10.5281/zenodo.3964244, Jul. 2020.

[32] S. Wankerl *et al.*, "An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language," in *Proc. Interspeech 2017*, 2017, pp. 3162–3166. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1572

[33] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[34] M. Egan *et al.*, "Methods to enhance verbal communication between individuals with Alzheimer's disease and their formal and informal caregivers: a systematic review," *International Journal of Alzheimer's Disease*, vol. 2010, 2010.

[35] K. L. Schmidt *et al.*, "Verbal communication among Alzheimer's disease patients, their caregivers, and primary care physicians during primary care office visits," *Patient education and counseling*, vol. 77, no. 2, pp. 197–201, 2009.

[36] J. B. Orange *et al.*, "Alzheimer's disease and other dementias: Implications for physician communication," *Clinics in geriatric medicine*, vol. 16, no. 1, pp. 153–173, 2000.