Comparing Acoustic-based Approaches for Alzheimer's Disease Detection

Aparna Balagopalan[1,2], Jekaterina Novikova[1]

[1]Winterlight Labs, Toronto, Canada
[2]University of Toronto, Toronto, Canada
`aparna@winterlightlabs.com, jekaterina@winterlightlabs.com`

Abstract

Robust strategies for Alzheimer's disease (AD) detection is important, given the high prevalence of AD. In this paper, we study the performance and generalizability of three approaches for AD detection from speech on the recent ADReSSo challenge dataset:1) using conventional acoustic features 2) using novel pre-trained acoustic embeddings 3) combining acoustic features and embeddings. We find that while feature-based approaches have a higher precision, classification approaches relying on the combination of embeddings and features prove to have a higher, and more balanced performance across multiple metrics of performance. Our best model, using such a combined approach, outperforms the acoustic baseline in the challenge by 2.8%.

Index Terms: Alzheimers disease, ADReSSo, dementia detection, computational paralinguistics

1. Introduction

Alzheimer's disease (AD) is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills. Research into the early assessment of Alzheimer's dementia is becoming increasingly more important, as the proportion of people affected by it grows every year [1]. Changes in cognitive ability due to neurodegeneration associated with AD lead to a progressive decline in memory and language quality [2].

Studies have shown that valuable clinical information indicative of cognition can be obtained from spontaneous speech elicited using pictures [3]. Studies have capitalized on this clinical observation, using speech analysis, natural language processing (NLP), and machine learning (ML) to distinguish between speech from healthy and cognitively impaired participants in datasets of semi-structured speech tasks such as picture description [4, 5, 6]. As such, this approach shows a potential to serve as quick, objective, and non-invasive assessments of an individual's cognitive status. The ADReSSo Challenge [7] aims to generate systematic evidence for these promises towards their clinical implementation.

ADReSSo (Alzheimer's Dementia Recognition through Spontaneous Speech *only*) Challenge is a shared task for the systematic comparison of approaches to the detection of cognitive impairment and decline based on spontaneous speech. In 2021, it focuses mainly on acoustic characteristics of speech, requiring the creation of models straight from speech, without manual transcription.

In this work, we develop ML models to detect AD from speech using picture description data with the demographically-matched ADReSSo challenge speech dataset. Following the previous work on comparing the linguistic approaches to AD detection from speech [8], we compare the following *acoustic-*based approaches to detect AD:

1. **Extracting conventional acoustic features from speech:** with this approach, we extract acoustic features from the audio files for binary AD vs non-AD classification. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection [4, 9].

2. **Using pre-trained deep neural models:** with this approach, we embed the raw audio into representations using the last hidden state of the pre-trained wav2vec 2.0 model [10] and classify audio samples using these embeddings as input to simple classifiers.

3. **A combination of both approaches.** Here, we combine the two approaches and make use of both engineered features and audio representations generated using a pre-trained deep neural model.

In this paper, we evaluate performance of the three approaches on both the ADReSSo train dataset, and on the unseen test set. We find that models based on conventional acoustic features are best suited for the AD screening tasks, as they are able to achieve very high precision, while under performing on the rest of evaluation metrics, such as accuracy, recall and F1 score. For the cases when a more balanced performance is required, a combination of conventional acoustic features and pre-trained deep neural models is the most promising method, as it allows to achieve high performance and generalize well to unseen data.

The main contributions of our paper are as follows:

- We use audio representations extracted from the pre-trained deep neural model wav2vec 2.0 [10] for the task of AD classification from speech. To the best of our knowledge, this method and this model were never applied before for such a classification task.

- We present the model that combines conventional acoustic features with pre-trained representations of speech and outperforms the ADReSSo baseline model by 2.82%.

- We carefully compare three different acoustic-based approaches for AD detection, which allows us to draw more general conclusions on performance and generalizability of acoustic models.

2. Background

2.1. Extracting conventional acoustic features from speech

Some of the conventional features employed to describe acoustic characteristics of the voice applied to AD detection, include fundamental frequency, jitter and shimmer [11]. In addition to these features, there is also a range of elements properly accompanying linguistic emissions and which constitute signs and clues but are not verbal. These characteristics of speech are called paralinguistic features and have been used to obtain information from the patient by means of the statistics of e.g. Cepstral Mel-Frequency Components (MFCC), among others [9].

Different combinations of these features were used in multiple previous studies when detecting AD from the speech collected via picture description tasks [5, 6, 12, 13]. These works have provided clear evidence on the potential of using simple spoken tasks and conventional acoustic features to automatically assess early dementia and its progression as well as to demonstrate that technology allows automatic detection of AD.

2.2. Using pre-trained deep neural methods

In the recent years, pre-training of deep neural networks has emerged as an effective approach to overcome the problem of data scarcity [14, 15, 16, 10]. The key idea of such a technique, which is also called "transfer learning", is to learn general representations in a setup where substantial amounts of labeled or unlabeled data is available and to leverage the learned robust representations to improve performance on a downstream task for which the amount of data is limited.

In natural language processing, one of the most popular transfer learning models is BERT [15], which trains "contextual embeddings" wherein a representation of a sentence (or transcript) is influenced by the context in which the words occur in sentences. In the field of speech processing, transfer learning models are mainly used for the purpose of automatic speech recognition [10]. We focus on using pre-trained embeddings from a self-supervised audio representation model, wav2vec 2.0 [10], for the task of AD detection from speech. In the wav2vec 2.0 model, audio is encoded via a multi-layer convolutional neural network and then masks spans of the resulting latent speech representations similar to masked language modeling [15]. The latent representations are fed to a Transformer network [17] to build contextualized representations.

For the task of AD detection, transfer learning is particularly useful, as it is difficult and costly to collect labelled data. While several prior works have used pre-trained acoustic embeddings such as x-vectors and i-vectors for AD detection from speech [18, 19], to the best of our knowledge, no works have utilized self-supervised representations of speech for AD detection. Hence, we aim to benchmark a self-supervised representation learning methodology for AD detection.

2.3. A combination of pre-trained and conventional approaches

Incorporating domain-specific external knowledge in neural language representations is a field of research that has been actively explored with both acoustic and linguistic embeddings [20, 21]. However, a large amount of prior work is either focused on linguistic based approaches or using simple acoustic embeddings such as x-vectors [10]. In contrast, we are using a combination of state-of-the art self-supervised techniques for speech representations and combining these with domain-knowledge informed conventional speech features.

3. Methodology

3.1. Dataset

The ADReSSo dataset we use in this work consists of set of speech recordings of picture descriptions produced by cognitively normal (or healthy) subjects and patients with an AD diagnosis, who were asked to describe the Cookie Theft picture [22] from the Boston Diagnostic Aphasia Examination [22]. There are speech samples from 237 participants in total, out of which 166 are in the training set, and 71 in the test

set (70/30 split balanced for demographics). Out of the samples in the training set, 83 were cognitively healthy, while 83 had AD. The prediction dataset used is matched for age and gender so as to minimise risk of bias in the prediction tasks using a propensity score. Along with the recorded speech data, the dataset also included segmentation profiles for optional use. For all our acoustic approaches, we use the full unsegmented audio, following the baseline acoustic approach [7] and prior works [18, 23].

3.2. Feature Extraction

3.2.1. Using conventional acoustic features from speech

We extract 168 acoustic features from the unsegmented speech audio files. Those include several statistics such as mean, variance, kurtosis, etc. of mel-frequency cepstral coefficients (MFCCs), following prior work [4]. We only use samples from the audio component provided with the dataset, and do not perform any Automatic Speech Recognition or linguistic analyses.

3.2.2. Using embeddings from pre-trained deep neural models for audio representation

In order to create audio representations using this approach, we make use of the huggingface[1] implementation of the wav2vec 2.0 [10] base model *wav2vec2-base-960h*. This base model is pretrained and fine-tuned on 960 hours of Librispeech on 16kHz sampled speech audio. We first represent each unsegmented audio file as a waveform with librosa[2]. We then tokenize waveforms using *Wav2Vec2Tokenizer* and if necessary, divide them into smaller chunks (with the maximum size of 320000 in our case) to fit into memory, afterwards we feed them into the wav2vec 2.0 model. The last hidden state of the model is used as an embedded representation of audio. When tokenized waveforms are divided into several chunks, the mean value of all chunks is computed to generate the final embedding.

3.2.3. A combination of both approaches

In this approach, we study if engineered feature and pre-trained embeddings can augment each other by concatenating the representations at audio sample level. We combine representations from pre-trained models and conventional acoustic features by simply concatenating them. Hence, the dimension of the overall representation is the sum of the individual dimensions (936-dimensional representation vector overall).

3.3. Evaluation Methods

We evaluate performance primarily using accuracy scores, since all train/test sets are known to be balanced. We also report precision, recall, and F1 with respect to the positive class (AD).

Cross-validation on ADReSSo train set: We use two cross-validation strategies in our work – leave-one-subject-out (LOSO) CV and 10-fold CV. We report evaluation metrics with both these strategies for all models for direct comparison between approaches and with challenge baseline.

Predictions on ADReSSo test set: We generate predictions from the top-5 performing classifiers – where these classifiers are selected on the basis of highest LOSO-CV train accuracy, and challenge test predictions are generated from models trained on the last LOSO train split. We report performance on the challenge test set, as obtained from the challenge organizers.

---

[1]https://huggingface.co/models
[2]https://librosa.org/

Table 1: *10-fold CV and LOSO-CV results averaged across all the folds on the ADReSSo train set. Bold indicates the best performing approach for each model, bold+italics indicate the best overall performance for the metric.*

Model	Accuracy		Precision		Recall		F1	
	10-fold CV	LOSO CV	10-fold CV	LOSO CV	10-fold CV	LOSO CV	10-fold CV	LOSO CV
LR-feat	0.6084	0.6386	0.6774	**0.7213**	0.4828	0.5057	0.5638	0.5946
LR-embed	**0.6867**	**0.6747**	0.6882	0.6737	*0.7356*	*0.7356*	**0.7111**	0.7033
LR-combo	0.6807	0.6687	**0.7024**	0.6951	0.6782	0.6552	0.6901	0.6746
SVM-feat	0.6265	0.6566	*0.8571*	*0.9167*	0.3448	0.3793	0.4918	0.5366
SVM-embed	0.6687	0.6566	0.6818	0.6705	**0.6897**	**0.6782**	0.6857	**0.6743**
SVM-combo	*0.6928*	**0.6807**	0.7308	0.7297	0.6552	0.6207	**0.6909**	0.6708
NN-feat	0.6084	0.6265	0.6833	0.6923	0.4713	0.5172	0.5578	0.5921
NN-embed	**0.6747**	0.6566	**0.6774**	0.6705	**0.7241**	0.6782	**0.7000**	0.6743
NN-combo	0.6506	*0.6928*	0.6706	**0.7000**	0.6552	**0.7241**	0.6628	*0.7119*
DT-feat	0.5964	0.5843	0.6190	0.6071	0.5977	0.5862	0.6082	0.5965
DT-embed	0.6446	0.6807	0.6591	0.6809	0.6667	*0.7356*	0.6629	0.7072
DT-combo	**0.6506**	**0.6867**	**0.6593**	**0.6882**	**0.6897**	*0.7356*	**0.6742**	**0.7111**

Table 2: *AD detection results on unseen, held-out ADReSS test set. Bold indicates the best result.*

Model	Accuracy	Precision	Recall	F1
Acoustic baseline	0.6479			
SVM-feat	0.6479	**0.9167**	0.3143	0.4681
LR-embed	0.6056	0.6000	0.6000	0.6000
DT-embed	0.5775	0.5714	0.5714	0.5714
SVM-combo	**0.6761**	0.6364	**0.8000**	**0.7089**
LR-combo	0.6056	0.6000	0.6000	0.6000

3.4. Experiments

3.4.1. Using conventional acoustic features from speech

We classify acoustic features (see Section 3.2.1) extracted at sample-level with several conventional linear and non-linear ML models : Logistic regression (LR-feat), Support Vector Machines (SVM-feat), Neural Network (NN-feat), and Decision Tree (DT-feat). We perform feature selection by choosing top-k number of features, based on ANOVA F-value between label/features, where k was set to 10 based on LOSO-CV. All model hyper-parameters were set to their default values as on the scikit-learn [24] implementation for each of these: logistic regression is trained with L2-penalty, SVM-feat is trained with a radial basis function kernel with kernel coefficient 0.001, and regularization parameter set to 1, NN-feat used has 1 layer with 10 units, and DT-feat with minimum samples per split set to 2.

3.4.2. Using embeddings from pre-trained deep neural models for audio representation

In order to leverage information encoded by pre-trained audio representation models, we extract embeddings from a pre-trained audio model, wav2vec 2.0, as a representation for each audio sample (see Section 3.2.2). We then classify these by training: Logistic regression (LR-embed), Support Vector Machines (SVM-embed), Neural Network (NN-embed), and Decision Tree (DT-embed) with default hyper-parameters.

3.4.3. Combining conventional features and embeddings from pre-trained deep neural models

Hence, a single representation with length equal to the sum of features and the embedding-dimension is obtained (see Section 3.2.3). We then classify these by training the following models on these concatenated representations : Logistic regression (LR-combo), Support Vector Machines (SVM-combo), Neural Network (NN-combo), and Decision Tree (DT-combo). All model hyper-parameters were set to their default values as on the scikit-learn implementation for each of these (i.e., same as in previous sections).

4. Results

4.1. AD detection results on the ADReSSo train set

The results of both LOSO and 10-fold CV evaluation of classification models' performance show that the conventional features-based approach consistently under performs in terms of accuracy, recall and F1 score (see Table 1). The feature-based approach, however, outperforms all the other approaches in terms of precision with the SVM model. With the NN and DT models, feature-based approach reaches high precision levels, although is not the best performing NN models (see Table 1).

Embedding-based and combo approaches compete for best performance in terms of accuracy, recall and F1 score. With logistic regression, embedding-based approach outperforms all the other approaches both in terms of accuracy, recall and F1. With the neural net model, the type of cross-validation determines whether it is the embedding-based or the combo approach that performs the best. With decision trees, however, the combination of conventional features and embeddings of speech perform the best in all the cases, even in terms of precision.

4.2. AD detection results on the unseen ADReSSo test set

Five models were selected based on their performance on the train set – magnitude as well as well as stability of accuracy in the two modes of CV – to submit to the ADReSSo challenge - SVM-feat to represent the feature-based approach, LR-embed and DT-embed to represent the embedding-based approach, and the LR-combo and SVM-combo to represent the combined approach. Out of these five models, SVM-feat had the highest precision (see Table 2). The SVM-combo model achieved the best performance in terms of accuracy, recall and F1, and was able to beat the baseline acoustic model by 2.82%.

5. Discussion

5.1. Classification performance

The models employing the feature-based approach are performing the best on the train data in terms of precision, when evaluated using LOSO CV. Both the highest achieved precision level (91.67% with SVM-feat) and the second-best result (72.13%

Table 3: *Difference between test performance and performance of performance of the model evaluated using 10-fold CV (i.e. w/ 10-fold) and LOSO CV (i.e. w/ LOSO). Positive results indicate that test performance was higher. Bold indicates the model that outperforms the acoustic baseline.*

| | Accuracy | | Precision | | Recall | | F1 | |
Model	w/ 10-fold	w/ LOSO	w/ 10-fold	w/ LOSO	w/ 10-fold	w/ LOSO	w/ 10-fold	w/ LOSO
SVM-feat	2.14%	-0.87%	5.96%	0.00%	-3.05%	-6.50%	-2.37%	-6.85%
LR-embed	-8.11%	-6.91%	-8.82%	-7.37%	-13.56%	-13.56%	-11.11%	-10.33%
DT-embed	-6.71%	-10.32%	-8.77%	-10.95%	-9.53%	-16.42%	-9.15%	-13.58%
SVM-combo	-1.67%	-0.46%	-9.44%	-9.33%	14.48%	17.93%	1.80%	3.81%
LR-combo	-7.51%	-6.31%	-10.24%	-9.51%	-7.82%	-5.52%	-9.01%	-7.46%

Table 4: *Difference between LOSO and 10-fold CV classification performance. Positive results indicate that LOSO CV performance was higher than that of 10-fold CV.*

Model	Accuracy	Precision	Recall	F1
LR-feat	3.02%	4.39%	2.29%	3.08%
SVM-feat	3.01%	5.96%	3.45%	4.48%
NN-feat	1.81%	0.90%	4.59%	3.43%
DT-feat	-1.21%	-1.19%	-1.15%	-1.17%
LR-embed	-1.20%	-1.45%	0.00%	-0.78%
SVM-embed	-1.21%	-1.13%	-1.15%	-1.14%
NN-embed	-1.81%	-0.69%	-4.59%	-2.57%
DT-embed	3.61%	2.18%	6.89%	4.43%
LR-combo	-1.20%	-0.73%	-2.30%	-1.55%
SVM-combo	-1.21%	-0.11%	-3.45%	-2.01%
NN-combo	4.22%	2.94%	6.89%	4.91%
DT-combo	3.61%	2.89%	4.59%	3.69%

with LR-feat) are achieved by the models that make use of conventional acoustic features. The test results reinforce the precision capability of the feature-based model, with the SVM-feat model achieving more than 28% higher precision on the unseen test set than any other model. However, feature-based models perform very poorly in terms of recall. As such, feature-based models could be a good candidate in a real-life deployment of AD screening models, when high precision is much more important than recall or overall accuracy.

There is no clear difference in train performance between the embedding-based and combo approaches. Combo models tend to perform better when evaluated using LOSO CV method, while embedding-based models often outperform combo models when evaluated using 10-fold CV. The test results show that the SVM-combo model has the best result, but more detailed analysis of generalizability of the three approaches is necessary.

5.2. Generalizability

In order to assess generalizability of the models, we first calculate the difference in performance of the five submitted models on the test and the train sets (see Table 3).

Feature-based model (SVM-feat) has no gap between test and train performance in terms of precision, and even surpasses the 10-fold CV precision on the test set. This model also achieves the highest precision levels both on the train and test sets, which allows us to conclude that the feature-based approach is potentially the best candidate for models when precision is the most important metric. This is not surprising, as these acoustic features were extracted based on several prior works on the speech characterization of patients with AD [5, 6, 12, 4].

Both embedding-based models (LR-embed and DT-embed) have a substantial gap between the test performance and the cross-validated performance on the train set (6.7-13.6% differ-

ence, with the test performance always being lower). The combination approaches on average, however, result in a smaller gap, especially in terms of accuracy, recall and F1 score. The SVM-combo model even performs better on the test set, when evaluated using recall and F1 metrics. In terms of accuracy, the SVM-combo model shows only 0.46% lower performance on the test set than on the train set, when cross-validated using the LOSO method. Such a slight difference confirms that the combination approach in general is a more generalizable method to train the AD detection system, comparing to using the embedding-based representations of speech alone.

We also aim to understand whether 10-fold or LOSO CV is the best approach to evaluate the model in order to achieve the result that is close to the expected test performance. For this, we additionally compare performance of each model between LOSO and 10-fold CV (see Table 4). With the feature-based approach, LOSO CV results are usually higher than those of 10-fold CV. As these models are most suitable for precision-focused cases (as we discussed above) and given test precision does not differ much from the LOSO CV results, the best solution would be to use LOSO CV evaluation when training these models on the data similar in size to the ADReSSo dataset.

With embedding-based and combo models, 10-fold CV mostly tends to outperform LOSO CV. However, test performance is in most cases closer to the LOSO CV performance. So with these approaches, similarly as with the feature-based approach, relying on the LOSO CV would be a more reliable strategy, when data is similar to the ADReSSo dataset. Note that these findings are dependant on the speech representation and modelling strategies as well as dataset domain in our study.

6. Conclusions and Future Work

In this paper we study the performance of conventional acoustic feature-based and pre-trained embedding-based classification approaches on Alzheimer's Disease detection from speech on the ADReSSo challenge dataset. We observe that feature-based approaches have a higher precision in general, and hence might be well-suited for screening AD via speech analysis. However, a more balanced performance across multiple-metrics is achieved by embedding-based approaches both while cross-validating on the train set and when tested on the unseen test set. Finally, we observe that a representation combining embeddings and conventional features outperforms both individual approaches, and attains an accuracy 2.8% higher that the best acoustic baseline in the challenge. With our careful comparisons, we hope to contribute to principled evaluations of the performance and generalizability of classification strategies on the important task of Alzheimer's Disease detection from speech. In future work, we will focus on different strategies to combine and fine-tune embeddings and feature-based models.

7. References

[1] S. Cahill, "WHO's global action plan on the public health response to dementia: some challenges and opportunities," 2020.

[2] C. Reitz, C. Brayne, and R. Mayeux, "Epidemiology of Alzheimer disease," *Nature Reviews Neurology*, vol. 7, no. 3, pp. 137–152, 2011.

[3] J. C. Borod, H. Goodglass, and E. Kaplan, "Normative data on the Boston diagnostic aphasia examination, parietal lobe battery, and the Boston naming test," *Journal of Clinical and Experimental Neuropsychology*, vol. 2, no. 3, pp. 209–215, 1980.

[4] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[5] A. Balagopalan, J. Novikova, F. Rudzicz, and M. Ghassemi, "The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech," in *NeurIPS Workshop on Machine Learning for Health ML4H*, 2018.

[6] Z. Zhu, J. Novikova, and F. Rudzicz, "Detecting cognitive impairments by agreeing on interpretations of linguistic features," 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)*, 2019, pp. 1431–1441.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," *medRxiv*, 2021.

[8] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection," *Proc. Interspeech 2020*, pp. 2167–2171, 2020.

[9] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554–9564, 2015.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[11] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert systems with applications*, vol. 150, p. 113213, 2020.

[12] Z. Zhu, J. Novikova, and F. Rudzicz, "Semi-supervised classification by reaching consensus among modalities," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language IRASL*, 2018.

[13] A. Balagopalan, K. Shkaruta, and J. Novikova, "Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020, pp. 159–164.

[14] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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: Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[18] A. Pompili, T. Rolland, and A. Abad, "The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge," *Proc. Interspeech 2020*, pp. 2202–2206, 2020.

[19] Y. Hauptman, R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher, "Identifying Distinctive Acoustic and Spectral Features in Parkinson's Disease," in *Interspeech*, 2019, pp. 2498–2502.

[20] A. Balagopalan and J. Novikova, "Augmenting BERT Carefully with Underrepresented Linguistic Features," in *NeurIPS Workshop on Machine Learning for Health ML4H*, 2020.

[21] Y. Cai and X. Wan, "Multi-Domain Sentiment Classification Based on Domain-Aware Embedding and Attention," in *IJCAI*, 2019, pp. 4904–4910.

[22] H. Goodglass and E. Kaplan, *Boston diagnostic aphasia examination booklet*. Lea & Febiger, 1983.

[23] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity."

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.