INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia

Matej Martinc, Senja Pollak

Jozef Stefan Institute, Ljubljana, Slovenia
matej.martinc@ijs.si, senja.pollak@ijs.si

Abstract

The paper describes a multimodal approach to the automated recognition of Alzheimer's dementia in order to solve the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge at INTERSPEECH 2020. The proposed method exploits available audio and textual data from the benchmark speech dataset to address challenge's two subtasks, a classification task that deals with classifying speech as dementia or healthy control speech and the regression task of determining the mini-mental state examination scores (MMSE) for each speech segment. Our approach is based on evaluating the predictive power of different types of features and on an exhaustive grid search across several feature combinations and different classification algorithms. Results suggest that even though TF-IDF based textual features generally lead to better classification and regression results, specific types of audio and readability features can boost the overall performance of the classification and regression models.

Index Terms: Cognitive Decline Detection, Computational Linguistics, Natural Language Processing, Speech Processing

1. Introduction

Alzheimer's Disease (AD) is the most common underlying cause of dementia, a neurodegenerative disease that leads to behavior and personality changes, such as decline in cognitive abilities and memory loss. AD is age-related and due to recent population trends suggesting large increases in elderly population [1], development of efficient methods for AD early detection and management has become of utmost importance.

The ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge [2] at INTERSPEECH 2020 [3] deals with automatic detection of AD from audio recordings and corresponding transcripts of subjects participating in a picture description task. The challenge defines two subtasks: Subtask 1 is a binary classification, i.e., to determine whether a patient has dementia or not, and SubTask 2 aims to determine the mini-mental state examination scores (MMSE) for each patient, i.e., a regression task.

The related work on AD classification reports accuracies of up to around 80% when best features are selected from a large set of linguistic and audio features [4, 5], or just linguistic features [6]. The accuracy in most cases decreases to below 70% in studies that consider only audio features [7], an exception being a study by Haider et al. [8], where the best accuracy of 78.7% is reported when an active data representation (ADR) feature extraction method is employed. When it comes to the regression task of determining the MMSE, we are aware of just one study that tackled it, reporting a mean absolute error (MAE) of 3.83 [5].

Due to findings from the related work and a relatively small size of the training set (108 training examples), our approach to both tasks was based on an extensive grid search over all possible feature combinations for each of the several pre-chosen classifiers and regressors[1]. These feature sets include several audio features (e.g., MFCC, ADR...) and a diversity of text features, covering different aspects of text transcripts (e.g., semantic features such as unigrams, syntactic features based on universal dependencies, which are in recent natural language processing research replacing the traditional part-of-speech tags and language dependant parsers, and statistical features indicating the readability of the text). The main contributions of this paper are as follows:

- Systematic evaluation of 16 distinct feature sets engineered from the audio signals and text transcripts and an insight into how they can be combined in the most efficient way.
- Deployment of novel universal dependency based features, and additional readability features for automated AD detection (i.e. ARI [9], GFI [10] and SMOG [11]).
- Development of a number of dementia AD classification and regression models with good performance and an available code for all experiments.

2. Methodology

Our core methodology consists of three parts, feature engineering (Section 2.1), choosing the learning algorithms (Section 2.2) and selection of the best feature combinations (Section 2.3).

2.1. Feature engineering

Features employed in the conducted experiments can be roughly divided into four distinct types, **audio features**, **TF-IDF features**, **readability features** and **embeddings**.

2.1.1. Audio features

All audio features were generated from the normalised audio-chunks, i.e., the .wav files extracted from the audio recordings of the AD and non-AD patient's speech after applying voice activity detection [2]. The following feature sets were constructed:

- **Mean MFCC**: means of first 13 mel-frequency cepstral coefficient features averaged across all audio recordings of each patient's speech. Window width of 25 ms and a stride of 10 ms were used in the extraction.
- **ADR**: an active data representation cluster based method for feature extraction [8] employed on Geneva minimalistic acoustic parameter set (**eGeMAPS**) and **MFCC**

[1]Code for the experiments is available under the MIT license at https://github.com/matejMartinc/ADReSSchallenge.

http://dx.doi.org/10.21437/Interspeech.2020-2202

features. Note that in our implementation, the self-organising maps (SOM) [12] clustering was replaced by a more widely used k-means clustering, with k=30.

- **Average duration** of audio recordings of each patient.

In addition, we also tested predictive power of mean root-mean-square, zero-crossing rate, spectral bandwidth, rolloff and centroid of audio samples, and the ADR feature extraction method on the emobase, ComParE 2013 and Multi-Resolution Cochleagram (MRCG) feature sets, as in [8], but did not use them in further experiments due to bad performance.[2]

2.1.2. TF-IDF features

TF-IDF features, which have been used in previous AD detection studies [13], were generated from the transcriptions of audio recordings[3] by generation of word and character n-gram tokens and employing bag-of-words vectorization and term frequency - inverse document frequency (TF-IDF) weighting on the derived tokens. The following tokens were used in vectorization and TF-IDF weighting:

- **Unigram** tokens, i.e., single words
- **Bigram** tokens, i.e., sequences of two adjacent words
- **Char4gram** tokens, i.e., sequences of four adjacent characters
- **Suffix** tokens, i.e., word suffixes of length 3
- **POS tag** bigrams, i.e., sequences of two adjacent part-of-speech tags
- **Grammmatical dependency (GRA)** features modelling grammatical relations between words in the input text, generated by the organizers of the challenge [2].
- **Universal dependency (UD)** features, i.e., a sequential representations of grammatical relations generated using the Stanford universal dependency parser [14]. For each word in the text, a tuple containing the type of grammatical relation (e.g., a determiner, nominal subject...) and the distance between the word at hand and its related word is generated. Unigrams, bigrams and trigrams of these tuples are used in our experiments.

2.1.3. Embeddings

Since related work reports promising results when word embeddings are used for AD detection [6, 15], we test several doc2vec embedding representations [16], namely *doc2vec text* representations generated from transcript texts, *doc2vec POS tags* representations generated from transcript POS tag sequences, *doc2vec GRA* representations generated from GRA features and *doc2vec UD* representations generated from UD feature sequences. We only use **doc2vec UD** features in further experiments, others were discarded due to bad performance.

2.1.4. Readability features

We experiment with several readability features. The hypothesis is that readability measures capture the complexity of language, which can be related to AD (AD patients display a decrease in the syntactic complexity of language [17] and have trouble in understanding the meaning of more complex words [18]):

- **Gunning fog index (GFI) [10]** was designed to estimate the years of formal education a person needs to understand the text on the first reading. It is calculated as $\text{GFI} = 0.4(\frac{\text{totalWords}}{\text{totalSentences}} + 100\frac{\text{longWords}}{\text{totalSentences}})$, where longWords are words longer than 7 characters.
- **Automated readability index [9] (ARI)** was also designed to return values corresponding to the years of education required to understand the text and is calculated as $\text{ARI} = 4.71(\frac{\text{totalCharacters}}{\text{totalWords}}) + 0.5(\frac{\text{totalWords}}{\text{totalSentences}}) - 21.43$
- **The SMOG grade (Simple Measure of Gobbledygook) [11]** is a readability formula mostly used for checking health messages and is calculated as $\text{SMOG} = 1.0430\sqrt{\text{num3Syllables}\frac{30}{\text{totalSentences}}}3.1291$, where the num3Syllables is the number of words with three or more syllables.
- **Number of unique words (NUW)**, normalized with the number of all words in the transcript.

Besides the readability features above, we also experimented with Flesch reading ease [19], Flesch-Kincaid grade level [19] and Dale-Chall [20] readability formulas, which were not used in further experiments due to bad performance.

2.2. Learning algorithms

Classification experiments were conducted by using four distinct classification algorithms from the Scikit library [21], namely Xgboost [22] (with 50 gradient boosted trees with max depth of 10), Random forest (with 50 trees of max depth of 5), SVM (with linear kernel and 2 box constraint configurations, 10 and 100) and Logistic regression (LogR) (with 2 distinct regularization configurations, 10 and 100). Regression experiments were conducted by using four distinct regression algorithms, namely Xgboost, SVM, Random forest and Linear regression (LinR). For Xgboost, SVM and Random forest same hyperparameters were used as for classification, while for LinR we used default parameters.

2.3. Exploration of feature space and model selection

Our approach is based on the early future-level fusion between different types of audio and textual features and relies on identification of feature combinations with the best synergy effect (see Figure 1). In order to do that, an extensive grid search across 65,535 combinations of 16 different feature sets (i.e., 4 audio, 7 TF-IDF, 1 embeddings and 4 readability feature sets) for each of the learning algorithms was conducted on the train set in a 10-fold cross-validation (CV) setting. For classification, accuracy is used for the performance evaluation, and for regression, root mean square error (RMSE) is used, same as for the official challenge evaluation [2].

The ADReSS challenge allows for submission of 5 distinct test set prediction tries. Therefore we identify 5 best performing classification models with non-identical predictions on the test set according to the grid search results. Their predictions on the test set are used for a majority vote ensemble, the output of which is used as one of the submissions. The other four submissions are test set predictions of the four best performing classification models. 5 submissions for regression are generated by first identifying 4 best performing regression models that do not

[2]The Logistic regression classifiers leveraging each of these feature sets did not outperform the majority baseline in the 10-fold cross-validation setting on the train set.

[3]Parts of the transcriptions that refer to the interviewer, and not the patient, were not used.

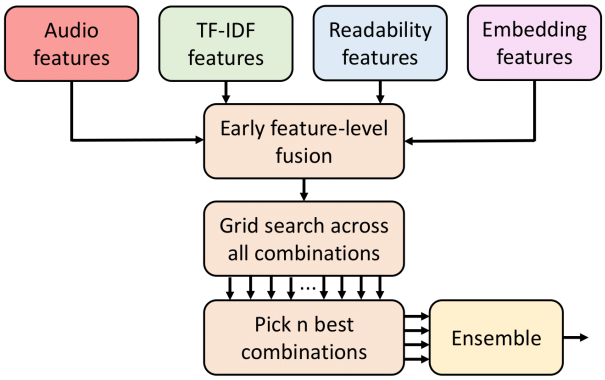


Figure 1: *Exploration of the feature space. Four types of features are combined by a concatenation of feature vectors (i.e., early feature-level fusion) and a grid search across all feature combinations is conducted. The best performing models employing best feature combinations are used for generating predictions on the test set, which are finally used for the ensembling (late prediction-level fusion).*

produce identical predictions on the test set and then calculating the mean of the predicted MMSE scores of these four best performing models in order to produce the fifth submission.

3. Experimental setting

In this Section we quickly overview the dataset and present the experiments conducted and results achieved in the scope of the ADReSS challenge. The Section is divided into three parts, Dataset (Section 3.1) Feature evaluation (Section 3.2) and Experimental results (Section 3.3).

3.1. Dataset

The dataset consists of recordings and transcripts of Cookie Theft picture descriptions by 78 AD and 78 non-AD participants of the Boston Diagnostic Aphasia Exam [23] and is balanced in terms of gender and age. Altogether the dataset contains 4,076 normalized speech segments, on average 24.86 per participant, and one transcript per each participant. It is split into a train set containing 108 examples and the test set containing 48 examples. For details, see [2].

3.2. Feature evaluation

In this experiment we explore the classification and regression performance of distinct feature sets in the 10-fold CV setting on the train set. SVM with box constraint of 10 was used in the feature evaluation experiments. Results for classification are presented in Figure 2. In general, TF-IDF features outperform all other feature types and among them, the best features are Char4grams that by themselves achieve the accuracy of 86.4%. While all TF-IDF feature sets lead to accuracy of about 70% or more, other types of features generally achieve accuracies between 50% and 60%, the only exception being ARI, which achieves accuracy just slightly above 60%. The worst performing feature is another readability measure, GFI, achieving accuracy just slightly above the chance level (51.8%). Among the audio features, the best performing are MFCC features (accuracy of 57.6%) and the worst are ADR features generated on the eGeMAPs (accuracy of 54.7%).

The feature performance on the regression task is somewhat consistent with the performance on the classification task (See Figure 3). TF-IDF features outperform other feature types and Char4grams are again the best features (achieving RMSE of 5.32). Also, ARI is again the best readability feature. On the other hand, MFCC features, which showed the best performance among audio features in the classification setting, are the worst features in the regression setting (achieving RMSE of 8.66). The best performing audio feature is the mean duration of the audio clips.

3.3. Experimental results

Results of the five best performing classification and regression models are presented in Table 1. The best classification accuracy of 77.8% on the official test set was achieved when a LogR model with a regularization strength (C) of 10 was trained on GFI, NUW, Duration, Char4gram, Suffix, POS tag and UD features. The same model also achieved the best accuracy in the CV setting, a much higher accuracy of 92.7%. On the other hand, for regression, the best RMSE score of 4.4388 on the test set was achieved by the SVM model with the box constraint of 10 trained on NUW, Bigram, Char4gram, Suffix, POS tag and GRA features, which performed the worst out of the four best regression models in the CV setting. While the ensemble of models produced the worst classification result on the test set, it ranked as second best on the regression task, although its performance was still much worse than the performance of the best model.

4. Discussion

The large discrepancies between the CV and test set classification performances suggest all the models overfitted, since all the models performed worse on the official test set than in the CV setting. The same can be said for four out of five regression models. Overfitting could be to some extent explained with the small size of the train set and might be limited by reducing the number of features. The one exception to the overfitting is the best performing regression model, which achieved a better RMSE score on the test set than in the CV setting. A more thorough error analysis would be required to explain this deviation.

Logistic/linear regression and SVMs with linear kernels proved better than Xgboost and Random forest models for both tasks. Some previous studies [24] suggest that these models work especially well on textual features and this could also explain their good performance on the tasks at hand, where textual TF-IDF features are the best performing features.

Besides the best performing textual features (Char4grams) and POS tags, which appear in most of the best classification and regression feature combinations, GFI and NUW also appear in 5 out of 9 best combinations, which suggests that readability measures add some useful information to the models. Interestingly, UD features only appear in best configurations for classification. When it comes to audio features, the best performing feature for classification appears to be Duration (appearing in 3 out of 5 best combinations) and the best performing feature for regression is MFCC ADR, appearing in 3 out of 4 best combinations. The doc2vec UD embedding features did not appear in any of the best combinations, most likely due to a very small train set which prohibits the successful training of an efficient embedding model.

Overall, our results outperform the baseline by a large margin [2] and are slightly worse than the results reported in the

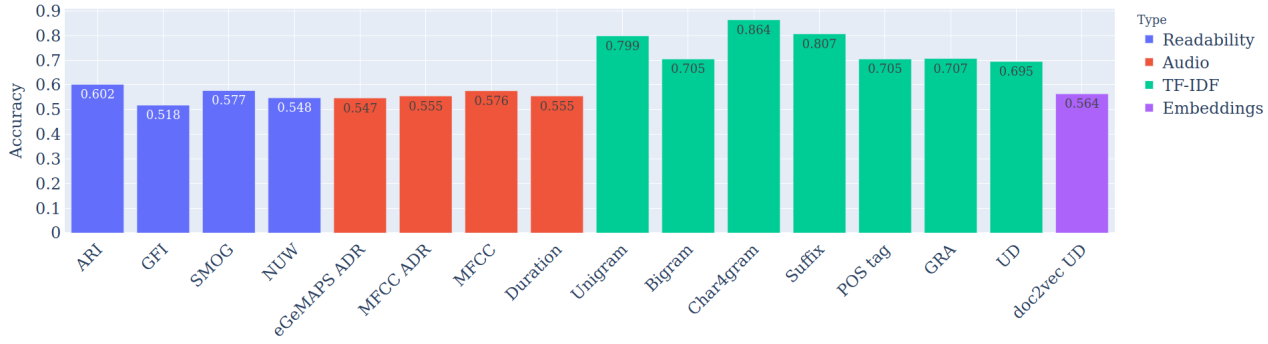


Figure 2: *SVM (with box constraint of 10) classification performance with different features.*

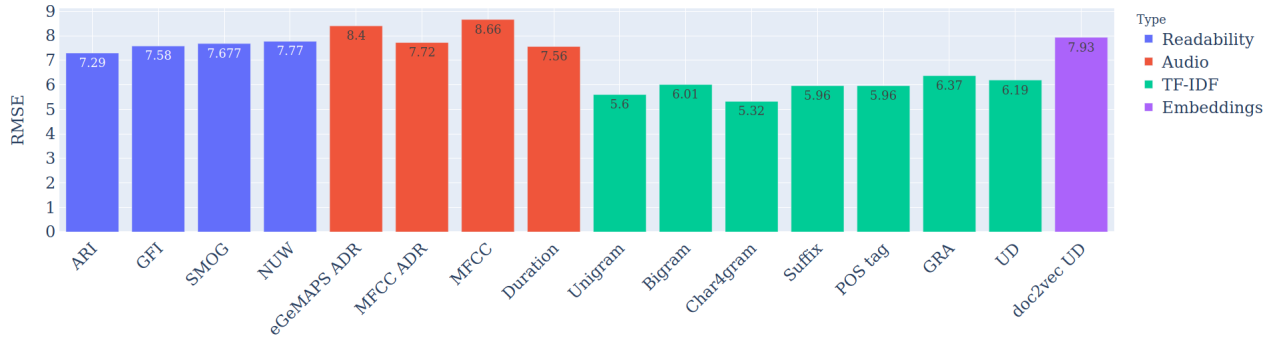


Figure 3: *SVM (with box constraint of 10) regression performance with different features.*

Table 1: *Results of the Cross validation (CV) and official test set experiments in terms of accuracy and RMSE.*

Classification			
Feature set	Model	CV score	Test set score
GFI,NUW,Duration,Character 4-grams,Suffixes,POS tag,UD	LogR (C=100)	**0.927**	**0.7708**
Duration,Character 4-grams,Suffixes,POS tag,UD	SVM (C=10)	0.918	0.7500
NUW,Duration,Unigram,Suffixes,POS tag,UD	LogR (C=10)	0.917	0.7500
GFI,Duration,Unigram,Bigram,Suffixes,POS tag,UD	SVM (C=10)	0.908	0.7500
duration,Unigram,Bigram,Suffixes,POS tag,UD	LogR (C=10)	0.907	/
Ensemble	/	/	0.7292
Regression			
Feature set	Model	CV score	Test set score
GFI,NUW,MFCC ADR,Bigram,Character 4-grams,Suffixes,POS tag	LinR	**5.008**	5.1878
GFI,NUW,MFCC ADR,Character,4-grams,Suffixes,POS tag	LinR	5.021	5.4312
GFI,MFCC ADR,Character 4-grams,Suffixes,POS tag	LinR	5.032	5.4483
NUW,Bigram,Character 4-grams,Suffixes,POS tag,GRA	SVM (C=10)	0.505	**4.4388**
Ensemble	/	/	5.0574

related work [4, 5], which have been achieved on a much larger and also unbalanced DementiaBank's Pitt corpus [25].

5. Conclusions

In this paper we have presented a multimodal approach to the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge. The proposed method relies on a feature-level fusion between different feature types and an extensive grid search across all feature combinations, and exploits both audio and textual data for the automatic detection of Alzheimer's dementia.

The results suggest that a multimodal approach leads to better performance than unimodal approaches but also suggest caution about using many different features due to the overfitting risk. Besides testing new features (e.g., clinical features such as concept counts), our future work will therefore be focused on reducing the number of features in order to avoid overfitting, while still sustaining the predictive performance of the classification and regression models.

6. Acknowledgements

The authors acknowledge the financial support from the project SAAM – Supporting Active Ageing through Multimodal coaching (grant agreement no. 769661).

7. References

[1] W. H. Organization *et al.*, "Mental health action plan 2013-2020," 2013.

[2] 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

[3] *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 2020.

[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] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[6] J. S. Guerrero-Cristancho, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Word-embeddings and grammar features to detect language disorders in alzheimer's disease patients," *TecnoLógicas*, vol. 23, no. 47, pp. 63–75, 2020.

[7] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[8] 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*, 2019.

[9] E. A. Smith and R. Senter, "Automated readability index," *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pp. 1–14, 1967.

[10] R. Gunning, *The technique of clear writing*. McGraw-Hill, New York, 1952.

[11] G. H. Mc Laughlin, "Smog grading - a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[12] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[13] J. Bullard, C. O. Alm, X. Liu, Q. Yu, and R. A. Proano, "Towards early dementia detection: fusing linguistic and non-linguistic clinical data," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 12–22.

[14] M.-C. De Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. D. Manning, "Universal stanford dependencies: A cross-linguistic typology." in *LREC*, vol. 14, 2014, pp. 4585–4592.

[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] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[17] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[18] P. Scheltens, "100 questions and answers about alzheimer's disease," 2004.

[19] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida, 1975.

[20] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[21] 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.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

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

[24] F. Rangel, P. Rosso, M. Potthast, and B. Stein, "Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter," *Working Notes Papers of the CLEF*, pp. 1613–0073, 2017.

[25] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "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.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease

Jiahong Yuan[1], *Yuchen Bian*[1], *Xingyu Cai*[1], *Jiaji Huang*[1], *Zheng Ye*[2], *Kenneth Church*[1]

[1]Baidu Research, USA
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, China

{jiahongyuan, yuchenbian, xingyucai, huangjiaji, kennethchurch}@baidu.com, yez@ion.ac.cn

Abstract

Disfluencies and language problems in Alzheimer's Disease can be naturally modeled by fine-tuning Transformer-based pre-trained language models such as BERT and ERNIE. Using this method, we achieved 89.6% accuracy on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, a considerable improvement over the baseline of 75.0%, established by the organizers of the challenge. The best accuracy was obtained with ERNIE, plus an encoding of pauses. Robustness is a challenge for large models and small training sets. Ensemble over many runs of BERT/ERNIE fine-tuning reduced variance and improved accuracy. We found that *um* was used much less frequently in Alzheimer's speech, compared to *uh*. We discussed this interesting finding from linguistic and cognitive perspectives.

Index Terms: Alzheimer's disease, disfluency, BERT, ERNIE, ensemble

1. Introduction

Alzheimer's disease (AD) involves a progressive degeneration of brain cells that is irreversible [1]. Therefore, early diagnosis and intervention is essential. One of the first signs of the disease is deterioration in language and speech production [2]. Case studies of the writings of the British Novelist Iris Murdoch indicated that lexical and syntactic changes occurred in the early stage of her AD [3]. Similarly, a study of President Ronald Regan's non-scripted news conferences found decreases in unique words and increases in conversational fillers and non-specific nouns well before his diagnosis of AD [4].

It is desirable to use language and speech for AD detection [5]. The ADReSS challenge of INTERSPEECH 2020 is "to define a shared task through which different approaches to AD detection, based on spontaneous speech, could be compared" [6]. This paper describes our effort for the shared task.

1.1. Studies of speech and language in AD and AD detection

There is an extensive literature on the characteristics of language and speech production in people with AD at various stages of the disease. Summaries of the studies can be found in [7, 8, 9]. Language impairment in AD is most evident in lexical, semantic, and pragmatic aspects. For example, people with AD often produce semantically "empty" words (e.g., *thing*, *stuff*) [10], use fewer information-bearing nouns and especially verbs [11], and their discourse appears to be disorganized [12]. Other aspects (syntax, phonology, and articulation) are believed to be relatively well preserved until late stages of the disease [13], though this conclusion is controversial [14, 15].

Many language problems cause disfluency in connected speech. Disfluencies are also common in normal spontaneous speech [16]. There are various types of disfluencies such as repetitions, false starts, repairs, filled and unfilled pauses. The phonetic consequence of speech disfluency has been well studied [17]. English has two common filled pauses, *uh* and *um*. There is a debate in the literature as to whether *uh* and *um* are intentionally produced by speakers [18, 19]. From sociolinguistic point of view, women and younger people tend to use more *um* vs. *uh* than men and older people [20, 21]. It has also been reported that autistic children use *um* less frequently than normal children [22, 23], and that *um* occurs less frequently and is shorter during lying compared to truth-telling [24]. It will be interesting to examine whether the use of *uh* and *um* in AD speech is different from normal speech. We did a preliminary investigation on this question, which is reported in Section 2. Although disfluencies are a part of normal speech, there is a boundary between normal and abnormal disfluencies. The boundary resides in a high dimensional space, determined by many interrelated factors such as pauses, repetitions, linguistic errors, discourse incoherence, etc. Classification of AD and normal speech requires a model that can capture these factors.

There is a considerable literature on AD detection from continuous speech [25, 26]. This literature considers a wide variety of features and machine learning techniques. [27] used 370 acoustic and linguistic features to train logistic regression models for classifying AD and normal speech. [28] found that acoustic and linguistic features were about equally effective for AD classification, but the combination of the two performed better than either by itself. Neural network models such as Convolutional Neural Networks and Long Short-Term Memory (LSTM) have also been employed for the task [29, 30, 31], and very promising results have been reported. However, it is difficult to compare these different approaches, because of the lack of standardized training and test data sets. One objective of the ADReSS challenge is to overcome this obstacle [6].

1.2. Pre-trained LMs and Self-attention

Modern pre-trained language models such as BERT [32] and ERNIE [33] were trained on extremely large corpora. These models appear to capture a wide range of linguistic facts including lexical knowledge, phonology, syntax, semantics and pragmatics. Recent literature is reporting considerable success on a variety of benchmark tasks with BERT and BERT-like models.[1] We expect that the language characteristics of AD can also be captured by the pre-trained language models when fine-tuned to the task of AD classification.

[1]https://gluebenchmark.com

http://dx.doi.org/10.21437/Interspeech.2020-2516

BERT and BERT-like models are based on the Transformer architecture [34]. These models use self-attention to capture associations among words. Each attention head operates on the elements in a sequence (e.g., words in the transcript for a subject), and computes a new sequence of the weighed sum of (transformed) input elements. There are various versions of BERT and ERNIE. There is a base model with 12 layers and 12 attention heads for each layer, as well as a larger model with 24 layers and 16 attention heads for each layer. Conceptually the self-attention mechanism can naturally model many language problems in AD mentioned in Section 1.1, including repetitions of words and phrases, use of particular words (and classes of words), as well as pauses. We proposed a method to encode pauses in a word sequence to enable BERT-like models to take advantage of disfluencies involving pauses, described in Section 3.1.

Previous studies have found that when fine tuning BERT for downstream tasks with a small data set, the model has a high variance in performance. Even with the same hyperparameter values, distinct random seeds can lead to substantially different results. [35] conducted a large-scale study on this issue. They fine-tuned BERT hundreds of times while varying only the random seeds, and found that the best-found model significantly outperformed previous reported results using the same model. In this situation, using just one final model for prediction is risky given the variance in performance during training. We propose an ensembling method to address this concern.

2. Data and analysis

2.1. Data

The data consists of speech recordings and transcripts of descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [36]. Transcripts were annotated using the CHAT coding system [37]. We only used word transcripts, the morphological and syntactic annotations in the transcripts were not used in our experiments.

The training set contains 108 speakers, and the test set contains 48 speakers. In each data set, half of the speakers are people with AD and half are non-AD (healthy control subjects). Both data sets were provided by the challenge. The organizers also provided speech segments extracted from the recordings using a simple voice detection algorithm, but no transcripts were available for the speech segments. We didn't use these speech segments. Our experiments were based on the entire recordings and transcripts.

2.2. Processing transcripts and forced alignment

The transcripts in the data sets were annotated in the CHAT format, which can be conveniently created and analyzed using CLAN [37]. For example: "the [x 3] bench [: stool]." In this example, [x 3] indicates that the word 'the' was repeated three times, [: stool] indicates that the preceding word, "bench" (which was actually produced), refers to stool. Details of the transcription format can be found in [37].

For the purpose of forced alignment and fine tuning, we converted the transcripts into words and tokens that represent what were actually produced in speech. 'w [x n]' were replaced by repetitions of w for n times, punctuation marks and various comments annotated between '[]' were removed. Symbols such as (.), (..), (...), <, >, / and xxx were also removed.

The processed transcripts were forced aligned with speech recordings using the Penn Phonetics Lab Forced Aligner [38].

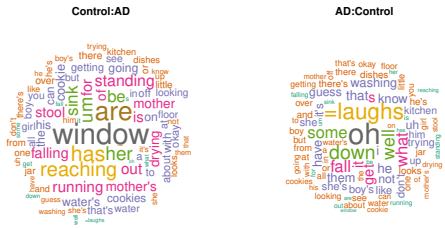


Figure 1: *The word cloud on the left highlights words that are more common among control subjects than AD; the word cloud on the right highlights words that are more common among AD than control.*

Table 1: *Subjects with AD say uh more often, and um less often.*

	uh	*um*
Control (non-AD)	130	51
Dementia (AD)	183	20

The aligner used a special model 'sp' to identify between-word pauses. After forced alignment, the speech segments that belong to the interviewer were excluded. The pauses at the beginning and the end of the recordings were also excluded. Only the subjects' speech, including pauses in turn-taking between the interviewer and the subject, were used.

2.3. Word frequency and *uh/um*

From the training data set, we calculated word frequencies for the Control and AD groups respectively. Words that appear 10 or more times in both groups are shown in the word clouds in Figure 1. The following words are at least two times more frequent in AD than in Control: *oh* (4.33), *=laughs* (laughter, 3.18), *down* (2.66), *well* (2.42), *some* (2.2), *what* (2.16), *fall* (2.15). And the words that are at least two times more frequent in Control than in AD are: *window* (4.4), *are* (3.83), *has* (3.0), *reaching* (2.8), *her* (2.62), *um* (2.55), *sink* (2.3), *be* (2.21), *standing* (2.06).

Compared to controls, subjects with AD used relatively more laughter and semantically "empty" words such as *oh*, *well*, and *some*, and fewer present particles (*-ing* verbs). This is consistent with the literature as discussed in Section 1.1. Table 1 shows an interesting difference for filled pauses. The subjects with AD used more *uh* than the control subjects, but their use of *um* was much less frequent.

2.4. Unfilled pauses

Durations of pauses were calculated from forced alignment. Pauses under 50 ms were excluded, as well as pauses in the interviewer's speech. We binned the remaining pauses by duration as shown in Figure 2. Subjects with AD have more pauses in every group, but the difference between subjects with AD and non-AD is particularly noticeable for longer pauses.

3. BERT and ERNIE Fine-tuning

3.1. Input and Hyperparameters

Pre-trained BERT and ERNIE models were fine-turned for the AD classification task. Each of the $N = 108$ training speakers is considered a data point. The input to the model consists

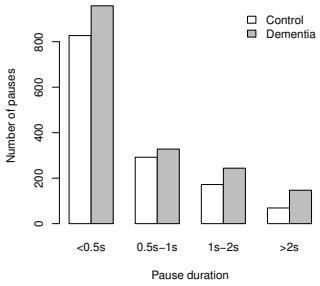


Figure 2: *Subjects with AD have more pauses (in all duration bins).*

of a sequence of words from the processed transcript for every speaker (as described in Section 2.2). The output is the class of the speaker, 0 for Control and 1 for AD.

We also encoded pauses in the input word sequence. We grouped pauses into three bins: short (under 0.5 sec); medium (0.5-2 sec); and long (over 2 sec). The three bins of pauses are coded using three punctuations ",", ".", and "...", respectively. Because all punctuations were removed from the processed transcripts, these inserted punctuations only represent pauses. Two examples of the input text are given below:

1. S136 (AD): *well your , sink is being run over , the . water , the stool the kid's standing on , is , falling and he's getting , cookies from a jar , the ... lady's washing ... dishes . the ... girl's reaching for a cookie ... could , there , be . more , i don't . think so .*
2. S062 (non-AD): *well there's a kid , stealing cookies from the cookie jar and his stool's about to topple over his , his sister's . asking for one the ... cookie jar is open of course the cupboard's open . the , mother's drying dishes the sink is overflowing . there are some , dishes on the side board . window's open i don't ... know , what else you want , there are curtains in the window i don't know if there's any .*

We used Bert-for-Sequence-Classification[2] for fine tuning. We tried both "bert-base-uncased" and "bert-large-uncased," and found slightly better performance with the larger model. The following hyperparameters (slightly tuned) were chosen: learning rate = 2e-5, batch size = 4, epochs = 8, max input length of 256 (sufficient to cover most cases). The standard default tokenizer was used (with an instruction not to split "..."). Two special tokens, [CLS] and [SEP], were added to the beginning and the end of each input.

ERNIE fine-tuning started with the "ERNIE-large" pretrained model (24 layers with 16 attention heads per layer). We used the default tokenizer, and the following hyperparameters: learning rate = 2e-5, batch size = 8, epochs = 20 and max input length of 256.

3.2. Ensemble Reduces Variance in LOO Accuracy

When conducting LOO (leave-one-out) cross-validation on the training set, large differences in accuracy across runs were observed, as illustrated in Figure 3. The black lines in Figure 3 were computed over 50 runs of BERT3p (top) and 50 runs of ERNIE0p and 50 runs of ERNIE3p (bottom). 0p indicates that no pause was encoded, and 3p indicates that three lengths of

[2] https://github.com/huggingface/transformers

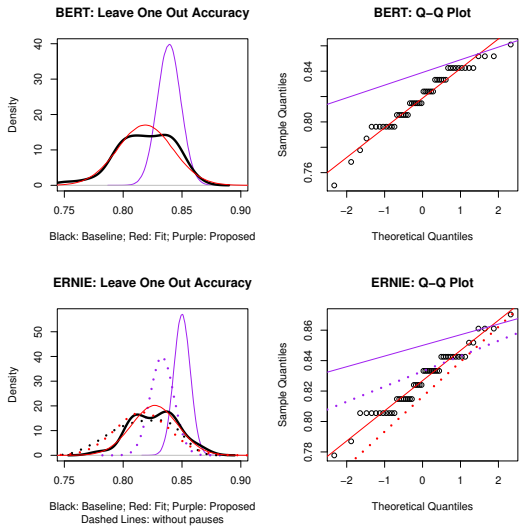


Figure 3: *We computed 50 estimates of leave-one-out (LOO) accuracy for BERT with pauses (top) and ERNIE with and without pauses (bottom). There is a wide variance in both cases (black). The proposed ensemble method (purple) improves the mean and reduces variance. Pauses are useful. Solid lines (with pauses) are better than dashed lines (without pauses).*

pauses were encoded. Each run reports a leave-one-out (LOO) accuracy. Everything was the same across runs except for random seeds. Over the 50 runs, LOO accuracy ranged from 0.75 to 0.86 for BERT3p, from 0.78 to 0.87 for ERNIE3p, and from 0.77 to 0.85 for ERNIE0p. The large variance suggests performance on unseen data is likely to be brittle. Such brittleness is to be expected given the large size of the BERT and ERNIE models and the small size of the training set (108 subjects).

To address this brittleness, we introduced the following ensemble procedure. From the results of LOO cross validation, we calculated the majority vote over 50 runs for each of the $N = 108$ subjects, and used the majority vote to return a single label for each subject. Tables 2-3 and Figure 3 show that this ensemble procedure improves the mean and reduces the standard deviation over estimates based on a single run.

To make sure that the ensemble estimates would generalize to unseen data, we tested the method by selecting $N = 5$, $N = 15$, ..., runs from the 50 runs reported in Figure 3. The results in the first row of Table 2 summarize 100 draws of $N = 5$ runs. The second row is similar, except $N = 15$. All of the rows in Table 2 have better means and less variance than the black line in Figure 3. Table 3 is like Table 2, except the means are even better with ERNIE than BERT. From Table 3 and Figure 3, we can also see that results with pauses are better than results without pauses.

4. Evaluation

Under the rules of the challenge, each team is allowed to submit results of five attempts for evaluation. Predictions on the test set from the following five models were submitted for evaluation: BERT0p, BERT3p, BERT6p, ERNIE0p, and ERNIE3p. To compare with three pauses, 6p represents six bins of pauses, encoded as: "," (under 0.5 sec), "." (.5-1 sec); ".." (1-2 sec), ". . ." (2-3 sec), ". . . ." (3-4 sec), ". . . . ." (over than 4 sec). The

Table 2: *Ensemble improves LOO (leave-one-out) estimates of accuracy; better means with less variance.*

N	BERT with Three Pauses	
	mean ± sd	min - max
5	0.837 ± 0.010	0.815 - 0.861
15	0.840 ± 0.011	0.815 - 0.861
25	0.839 ± 0.011	0.815 - 0.870
35	0.838 ± 0.010	0.824 - 0.861
45	0.839 ± 0.011	0.824 - 0.861

Table 3: *Ensemble also improves LOO for ERNIE (with and without pauses). LOO results are better with pauses than without, and better with ERNIE than BERT.*

N	ERNIE with Three Pauses		ERNIE with No Pauses	
	Mean ± Std	Min - Max	Mean ± Std	Min - Max
5	0.845 ± 0.013	0.806 - 0.880	0.828 ± 0.016	0.796 - 0.870
15	0.851 ± 0.008	0.833 - 0.870	0.831 ± 0.012	0.796 - 0.861
25	0.853 ± 0.007	0.833 - 0.870	0.833 ± 0.010	0.815 - 0.861
35	0.854 ± 0.007	0.824 - 0.861	0.836 ± 0.009	0.815 - 0.852
45	0.854 ± 0.007	0.833 - 0.861	0.834 ± 0.008	0.815 - 0.861

dots are separated from each other, as different tokens.

Following the method proposed in Section 3.2, we made 35 runs of training for each of the five models, with 35 random seeds. The classification of each sample in the test set was based on the majority vote of 35 predictions. Table 4 lists the evaluation scores received from the organizers.

The best accuracy was 89.6%, obtained with ERNIE and three pauses. It is a nearly 15% increase from the baseline of 75.0% [6].

ERNIE outperformed BERT by 4% on input of both three pauses and no pause. Encoding pauses improved the accuracy for both BERT and ERNIE. There was no difference between three pauses and six pauses in terms of improvement in accuracy.

5. Discussion

The group with AD used more *uh* but less *um* than the control group. In speech production, disfluencies such as hesitations and speech errors are correlated with cognitive functions such cognitive load, arousal, and working memory [24, 39]. Hesitations and disfluencies increase with increased cognitive load and arousal as well as impaired working memory. This may explain why the group with AD used more *uh*, as a filled pause and hesitation marker. More interestingly, they used less *um* than the control group. This indicates that unlike *uh*, *um* is more than a hesitation marker. Previous studies have also reported that children with autism spectrum disorder produced *um* less frequently than typically developed children [22, 23], and that *um* was used less frequently during lying compared to truth-telling [24, 40]. All these results seem to suggest that *um* carries a lexical status and is retrieved in speech production. One possibility is that people with AD or autism have difficulty in retrieving the word *um* whereas people who are lying try not to use this word. More research is needed to test this hypothesis.

From our results, encoding pauses in the input was helpful for both BERT and ERINE fine-tuning for the task of AD classification. Pauses are ubiquitous in spoken language. They are distributed differently in fluent, normally disfluent, and abnormally disfluent speech. As we can see from Figure 2, the group with AD used more pauses and especially more long pauses than the control group. With pauses present in the text, the self-attention mechanism in BERT and ERNIE may learn how the pauses are correlated with other words, for example, whether there is a long pause between the determiner *the* and the following noun, which occurs more frequently in AD speech. We think this is part of the reason why encoding pauses improved the accuracy. Both BERT and ERNIE were pre-trained on text corpora, with no pause information. Our study suggests that it may be useful to pre-train a language model using speech transcripts (either solely or combined with text corpora) that include pause information.

Table 4: *Evaluation results: Best accuracy (acc) with ERNIE and three pauses (3p). Pauses are helpful: three pauses (3p) and six pauses (6p) have better accuracy than no pauses (0p).*

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

6. Conclusions

Accuracy of 89.6% was achieved on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, with ERNIE fine-tuning, plus an encoding of pauses. There is a high variance in BERT and ERNIE fine-tuning on a small training set. Our proposed ensemble method improves the accuracy and reduces variance in model performance. Pauses are useful in BERT and ERNIE fine-tuning for AD classification. *um* was used much less frequently in AD, suggesting that it may have a lexical status.

7. Acknowledgements

We thank Julia Li and Hao Tian for their suggestion and help with ERNIE.

8. References

[1] M. P. Mattson, "Pathways towards and away from alzheimer's disease," *Nature*, vol. 430, pp. 631–639, 2004.

[2] K. D. Mueller, R. L. Koscik, B. Hermann, S. C. Johnson, and L. S. Turkstra, "Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer's prevention," *Frontiers in Aging Neuroscience*, vol. 9, 2018.

[3] P. Garrard, L. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early alzheimer's disease on the characteristics of writing by a renowned author." *Brain : a journal of neurology*, vol. 128 Pt 2, pp. 250–60, 2005.

[4] V. Berisha, S. Wang, A. LaCross, and J. M. Liss, "Tracking discourse complexity preceding alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald

reagan and george herbert walker bush." *Journal of Alzheimer's disease : JAD*, vol. 45 3, pp. 959–63, 2015.

[5] C. Laske, H. R. Sohrabi, S. Frost, K. L. de Ipiña, and S. E. O'Bryant, "Innovative diagnostic tools for early detection of alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, pp. 561–578, 2015.

[6] 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

[7] V. O. B. Emery, "Language impairment in dementia of the alzheimer type: a hierarchical decline?" *International journal of psychiatry in medicine*, vol. 30 2, pp. 145–64, 2000.

[8] G. Szatlóczki, I. Hoffmann, V. Vincze, J. Kálmán, and M. Pákáski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 7, 2015.

[9] A. Slegers, R.-P. Filiou, M. Montembeault, and S. M. Brambati, "Connected speech features from picture description in alzheimer's disease: A systematic review," *Journal of Alzheimer's disease*, vol. 65 2, pp. 519–542, 2018.

[10] D. Kempler, "Language changes in dementia of the alzheimer type," in *Dementia and Communication*, R. Lubinski, Ed. Philadelphia: B. C. Decker, 1991, ch. 7, pp. 98–113.

[11] M. M. Kim and C. K. Thompson, "Verb deficits in alzheimer's disease and agrammatism: Implications for lexical organization," *Brain and Language*, vol. 88, pp. 1–20, 2004.

[12] M. Mentis, J. Briggs-Whittaker, and G. D. Gramigna, "Discourse topic management in senile dementia of the alzheimer's type," *Journal of speech and hearing research*, vol. 38 5, pp. 1054–66, 1995.

[13] D. Kempler, S. Curtiss, and C. Jackson, "Syntactic preservation in alzheimer's disease." *Journal of speech and hearing research*, vol. 30 3, pp. 343–50, 1987.

[14] K. Croot, J. R. Hodges, J. H. Xuereb, and K. Patterson, "Phonological and articulatory impairment in alzheimer's disease: A case series," *Brain and Language*, vol. 75, pp. 277–309, 2000.

[15] L. Altmann, D. Kempler, and E. Andersen, "Speech errors in alzheimer's disease : reevaluating morphosyntactic preservation," *Journal of Speech Language and Hearing Research*, vol. 44, pp. 1069–82, 2001.

[16] E. Shriberg, "Preliminaries to a theory of speech disfluencies," phd, University of California, Berkeley, 1994. [Online]. Available: http://www.speech.sri.com/people/ees/publications.html

[17] ——, "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, pp. 153 – 169, 06 2001.

[18] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[19] M. Corley and O. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, pp. 589–602, 07 2008.

[20] G. Tottie, "Uh and um as sociolinguistic markers in british english," *International Journal of Corpus Linguistics*, vol. 16, pp. 173–197, 2011.

[21] M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman, "Variation and change in the use of hesitation markers in germanic languages," *Language Dynamics and Change*, vol. 6, no. 2, pp. 199–234, 2016.

[22] C. A. Irvine, I.-M. Eigsti, and D. Fein, "Uh, um, and autism: Filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 46, pp. 1061–1070, 2016.

[23] K. Gorman, L. Olson, A. Hill, R. Lunsford, P. Heeman, and J. Santen, "Uh and um in children with autism spectrum disorders or language impairment," *Autism research : official journal of the International Society for Autism Research*, vol. 9, pp. 854–865, 2016.

[24] J. Arciuli, D. MALLARD, and G. Villar, ""um, i can tell you're lying": Linguistic markers of deception versus truth-telling in speech," *Applied Psycholinguistics*, vol. 31, pp. 397 – 411, 07 2010.

[25] R.-P. Filiou, N. Bier, A. Slegers, B. Houzé, P. Belchior, and S. M. Brambati, "Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review," *Aphasiology*, pp. 1–33, 2019.

[26] M. L. B. Pulido, J. B. A. Hernández, M. A. F. Ballester, C. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems With Applications*, vol. 150, p. 113213, 2020.

[27] 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 2, pp. 407–22, 2016.

[28] G. Gosztolya, V. Vincze, L. Toth, M. Pakaski, J. Kalman, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's diseasebased on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[29] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of alzheimer's disease using neural network language models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.

[30] F. D. Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *ACL*, 2019.

[31] K. L. de Ipiña, U. M. de Lizarduy, P. M. Calvo, B. Beitia, J. Garcia-Melero, M. Ecay-Torres, A. Estanga, and M. Faúndez-Zanuy, "Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach," *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pp. 1–4, 2017.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," *arXiv preprint arXiv:1907.12412*, 2019.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[35] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," *arXiv preprint arXiv:2002.06305*, 2020.

[36] E. K. H. Goodglass and B. Barresi, *Boston Diagnostic Aphasia Examination – Third Edition*. Philadelphia: Lippincott Williams & Wilkins, 2001.

[37] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[38] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *The Journal of the Acoustical Society of America*, vol. 123, p. 3878, 2008.

[39] M. Daneman, "Working memory as a predictor of verbal fluency," *Journal of Psycholinguistic Research*, vol. 20, pp. 445–464, 1991.

[40] S. Benus, F. Enos, J. Hirschberg, and E. Shriberg, "Pauses in deceptive speech," in *Speech Prosody 2006*, 2006.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection

Aparna Balagopalan[1], *Benjamin Eyre*[1], *Frank Rudzicz*[2,3,4,5], *Jekaterina Novikova*[1]

[1]Winterlight Labs Inc, Toronto, ON
[2]University of Toronto, ON
[3] Vector Institute for Artificial Intelligence, Toronto, ON
[4]Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, ON
[5]Surgical Safety Technologies, Toronto, ON

aparna@winterlightlabs.com, benjamin@winterlightlabs.com, frank@cs.toronto.edu, jekaterina@winterlightlabs.com

Abstract

Research related to automatically detecting Alzheimer's disease (AD) is important, given the high prevalence of AD and the high cost of traditional methods. Since AD significantly affects the content and acoustics of spontaneous speech, natural language processing and machine learning provide promising techniques for reliably detecting AD. We compare and contrast the performance of two such approaches for AD detection on the recent ADReSS challenge dataset [1]: 1) using domain knowledge-based hand-crafted features that capture linguistic and acoustic phenomena, and 2) fine-tuning Bidirectional Encoder Representations from Transformer (BERT)-based sequence classification models. We also compare multiple feature-based regression models for a neuropsychological score task in the challenge. We observe that fine-tuned BERT models, given the relative importance of linguistics in cognitive impairment detection, outperform feature-based approaches on the AD detection task.

Index Terms: Alzheimer's disease, ADReSS, dementia detection, MMSE regression, BERT, feature engineering, transfer learning.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes problems with memory, thinking, and behaviour. AD affects over 40 million people worldwide with high costs of acute and long-term care [2]. Current forms of diagnosis are both time consuming and expensive [3], which might explain why almost half of those living with AD do not receive a timely diagnosis [4].

Studies have shown that valuable clinical information indicative of cognition can be obtained from spontaneous speech elicited using pictures [5]. Several studies have used speech analysis, natural language processing (NLP), and ML to distinguish between healthy and cognitively impaired speech of participants in picture description datasets [6, 7]. These serve as quick, objective, and non-invasive assessments of an individual's cognitive status. However, although ML methods for automatic AD-detection using such speech datasets achieve high classification performance (between 82%-93% accuracy) [6, 8, 9], the field still lacks publicly-available, balanced, and standardised benchmark datasets. The ongoing ADReSS challenge [1] provides an age/sex-matched balanced speech dataset, which consists of speech from AD and non-AD participants describing a picture. The challenge consists of two key tasks: 1) Speech classification task: classifying speech as AD or non-AD. 2) Neuropsychological score regression task: predicting Mini-Mental State Examination (MMSE) [10] scores from speech.

In this work, we develop ML models to detect AD from speech using picture description data of the demographically-matched ADReSS challenge speech dataset [1], and compare the following training regimes and input representations to detect AD:

1. **Using domain knowledge**: with this approach, we extract linguistic features from transcripts of speech, and acoustic features from corresponding audio files for binary AD vs non-AD classification and MMSE score regression. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection [6].
2. **Using transfer learning**: with this approach, we fine-tune pre-trained BERT [11] text classification models at transcript-level. BERT achieved state-of-the-art results on a wide variety of NLP tasks when fine-tuned [11]. Our motivation is to benchmark a similar training procedure on transcripts from a pathological speech dataset, and evaluate the effectiveness of high-level language representations from BERT in detecting AD.

In this paper, we evaluate performance of these two methods on both the ADReSS train dataset, and on the unseen test set. We find that fine-tuned BERT-based text sequence classification models achieve the highest AD detection accuracy with an accuracy of 83.3% on the test set. With the feature-based models, the highest accuracy of 81.3% is achieved by the SVM with RBF kernel model. The lowest root mean squared error obtained for the MMSE prediction task is 4.56, with a feature-based L2 regularized linear regression model.

The main contributions of our paper are as follows:

- We employ a domain knowledge-based approach and compare a number of AD detection and MMSE regression models with an extensive list of pre-defined linguistic and acoustic features as input representations from speech (Section 5 and 6).
- We employ a transfer learning-based approach and benchmark fine-tuned BERT models for the AD vs non-AD classification task (Section 5 and 6).
- We contrast the performance of the two approaches on the classification task, and discuss the reasons for existing differences (Section 7).

http://dx.doi.org/10.21437/Interspeech.2020-2557

Table 1: *Basic characteristics of the patients in each group in the ADReSS challenge dataset are more balanced in comparison to DementiaBank.*

Dataset			Class	
			AD	Non-AD
ADReSS	Train	Male	24	24
		Female	30	30
ADReSS	Test	Male	11	11
		Female	13	13
DementiaBank [17]	-	Male	125	83
		Female	197	146

2. Background

2.1. Domain Knowledge-based Approach

Previous work has focused on automatic AD detection from speech using acoustic features (such as zero-crossing rate, Mel-frequency cepstral coefficients) and linguistic features (such as proportions of various part-of-speech (POS) tags [12, 6, 8]) from speech transcripts. Fraser *et al.* [6] extracted 370 linguistic and acoustic features from picture descriptions in the DementiaBank dataset, and obtained an AD detection accuracy of 82% at transcript-level. More recent studies showed the addition of normative data helped increase accuracy up to 93% [8, 13] .

Yancheva *et al*. [14] showed ML models are capable of predicting the MMSE scores from features of speech elicited via picture descriptions, with mean absolute error of 2.91-3.83.

Detecting AD or predicting MMSE scores with engineered features of speech and thereby infusing domain knowledge into the task has several advantages, such as more interpretable model decisions and potentially lower resource requirement when paired with conventional ML models. However, there are also disadvantages, e.g. a time consuming feature engineering process, and a risk of missing highly relevant features.

2.2. Transfer Learning-based Approach

In the recent years, transfer learning in the form of pre-trained language models has become ubiquitous in NLP [15] and has contributed to the state-of-the-art on a wide range of tasks. One of the most popular transfer learning models is BERT [11], which builds on Transformer networks [16] to pre-train bidirectional representations of text by conditioning on both left and right contexts jointly in all layers.

BERT uses powerful attention mechanisms to encode global dependencies between the input and output. This allows it to achieve state-of-the-art results on a suite of benchmarks [11]. Fine-tuning BERT for a few epochs can potentially attain good performance even on small datasets. However, such models are not directly interpretable, unlike feature-based ones.

3. Dataset

We use the ADReSS Challenge dataset [1], which consists of 156 speech samples and associated transcripts from non-AD (N=78) and AD (N=78) English-speaking participants. Speech is elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia exam [5]. In contrast to other speech datasets for AD detection such as DementiaBank's English Pitt Corpus [17], the ADReSS challenge dataset is matched for age and gender (Table 1). The speech dataset is divided into standard train and test sets. MMSE [10] scores are available for all but one of the participants in the train set.

4. Feature Extraction

The speech transcripts in the dataset are manually transcribed as per the CHAT protocol [18], and include speech segments from both the participant and an investigator. We only use the portion of the transcripts corresponding to the participant. Additionally, we combine all participant speech segments corresponding to a single picture description for extracting acoustic features.

We extract 509 manually-engineered features from transcripts and associated audio files (see Appendix A for a list of all features). These features are identified as indicators of cognitive impairment in previous literature, and hence encode domain knowledge. All of them are divided into 3 categories:

1. **Lexico-syntactic features (297):** Frequencies of various production rules from the constituency parsing tree of the transcripts [19], speech-graph based features [20], lexical norm-based features (e.g. average sentiment valence of all words in a transcript, average imageability of all words in a transcript [21]), features indicative of lexical richness. We also extract syntactic features [22] such as the proportion of various POS-tags, and similarity between consecutive utterances.
2. **Acoustic features (187):** Mel-frequency cepstral coefficients (MFCCs), fundamental frequency, statistics related to zero-crossing rate, as well as proportion of various pauses [23] (for example, filled and unfilled pauses, ratio of a number of pauses to a number of words, etc.)
3. **Semantic features based on picture description content (25):** Proportions of various information content units used in the picture, identified as being relevant to memory impairment in prior literature [24].

5. Experiments

5.1. AD vs non-AD Classification

5.1.1. Training Regimes

We benchmark the following training regimes for classification: classifying features extracted at transcript-level and a BERT model fine-tuned on transcripts.

Domain knowledge-based approach: We classify lexicosyntactic, semantic, and acoustic features extracted at transcript-level with four conventional ML models (SVM, neural network (NN), random forest (RF), naïve Bayes (NB)[1].

Hyperparameter tuning: We optimize each model to the best possible hyper-parameter setting using grid-search 10-fold cross-validation (CV). We perform feature selection by choosing top-k number of features, based on ANOVA F-value between label/features. The number of features is jointly optimized with the classification model parameters (see Appendix C for a full list of parameters).

Transfer learning-based approach: To leverage the language information encoded by BERT [11], we add a linear layer mapping representations from the final layer of a pre-trained 12-layer BERT base, uncased model to binary class labels [25] for the AD vs non-AD classification task. The transcript-level input to the model consists of transcribed utterances with corresponding start and separator special tokens for each utterance, following Liu *et al.* [26]. A pooled embedding summarizing information across all tokens in the transcript using the self-attention mechanism in the BERT base is used as the aggregate transcript

[1] https://scikit-learn.org/stable/

representation, and passed to the classification layer [11, 25]. This model is then fine-tuned on training data for AD detection.

Hyperparameter tuning: We optimize the number of epochs to 10 by varying it from 1 to 12 during CV. Adam optimizer [27] and warmup linear learning rate scheduling [28] are used (details in Appendix B).

5.1.2. Evaluation

Cross-validation on ADReSS train set: We use two CV strategies in our work – leave-one-subject-out CV (LOSO CV) and 10-fold CV at transcript level. We report evaluation metrics with LOSO CV for all models except fine-tuned BERT for direct comparison to challenge baselines. Due to computational constraints of GPU memory, we are unable to perform LOSO CV for the BERT model. Hence, we perform 10-fold CV to compare feature-based classification models with fine-tuned BERT. Values of performance metrics for each model are averaged across three runs with different random seeds in all cases.

Predictions on ADReSS test set: We generate three predictions with different seeds from each hyperparameter-optimized classifier trained on the complete train set, and then produce a majority prediction to avoid overfitting. We report performance on the challenge test set, as obtained from the challenge organizers (see Appendix E for more details).

We evaluate performance primarily using accuracy scores, since all train/test sets are known to be balanced. We also report precision, recall, specificity and F1 with respect to the positive class (AD), and compare to the highest challenge baseline (LDA classifier using language outcome measures [1]).

5.2. MMSE Score Regression

5.2.1. Training Regimes

Domain knowledge-based approach: For this task, we benchmark two kinds of regression models, linear and ridge, using pre-engineered features as input. MMSE scores range from 0 to 30, and so predictions are clipped to range between 0 and 30.

Hyperparameter tuning: Each model's performance is optimized using hyperparameters selected via grid-search LOSO CV. We perform feature selection by choosing top-k features, based on F-Scores computed from the correlation of each feature with MMSE score. The number of features is optimized for all models. For ridge regression, the number of features is jointly optimized with the coefficient for L2 regularization, α.

5.2.2. Evaluation

We report root mean squared error (RMSE) and mean absolute error (MAE) for the predictions produced by each of the models on the training set with LOSO CV. In addition, we include the RMSE for two models' predictions on the ADReSS test set. Hyperparameters for these models were selected using grid-search 10-fold cross validation on the training set. We compare regression performance to the best challenge baseline (decision tree regressor using language outcome measures [1]).

6. Results

6.1. AD vs non-AD Classification

In Table 3, the classification performance with all the models evaluated on the train set via 10-fold CV is displayed. We observe that BERT outperforms all domain knowledge-based ML models with respect to all metrics. SVM is the best-performing domain knowledge-based model. However, accuracy of the fine-tuned BERT model is not significantly higher than that of the SVM classifier based on an Kruskal-Wallis H-test ($H = 0.4838, p > 0.05$).

We also report the performance of all our feature classification models with LOSO CV (Table 4), and compare to the highest challenge baseline [1]. Each of our classification models outperforms the challenge baseline, with a +10% accuracy increase with the SVM classifer. Feature selection results in accuracy increase of about 13% for the SVM classifier.

Results on the unseen, held-out ADReSS test set (Table 5) follow the trend of the cross-validated performance in terms of accuracy, with BERT outperforming the best feature-based classification model, SVM, as well as the challenge baseline.

6.2. MMSE Score Regression

Performance of regression models evaluated on both train and test sets is shown in Table 6. Ridge regression with 25 features selected attains the lowest RMSE on the training set amongst our models, with 4.56 RMSE during LOSO-CV, which is 0.18 higher than the challenge baseline. The results show that feature selection can help achieve a decrease of up to 1.5 RMSE points (and up to 0.86 MAE) for a ridge regressor. Furthermore, a ridge regressor is able to achieve an RMSE of 4.56 on the ADReSS test set, a decrease of 0.64 from the baseline.

7. Discussion

7.1. Feature Differentiation Analysis

We extract a large number of features to capture a wide range of linguistic and acoustic phenomena, based on a survey of prior literature in automatic cognitive impairment detection [6, 14, 30, 31]. In order to identify the most differentiating features between AD and non-AD speech, we perform independent t-tests between feature means for each class in the ADReSS training set. 87 features are significantly different between the two groups at $p < 0.05$. 79 of these are text-based lexicosyntactic and semantic features, while 8 are acoustic. These 8 acoustic features include the number of long pauses, pause duration, and mean/skewness/variance-statistics of various MFCC coefficients. However, after Bonferroni correction for multiple testing, we identify that only 13 features are significantly different between AD and non-AD speech at $p < 9e - 5$, and none of these features are acoustic (Table 2). This implies that linguistic features are particularly differentiating between the AD/non-AD classes here, which explains why models trained on linguistic features only attain performance well above random chance (see Fig. 1 in Appendix for visualization of class separability).

7.2. Analysing AD Detection Performance Differences

Comparing classification performance across model settings, we observe that BERT outperforms the best domain knowledge-based model in terms of accuracy and F1-score on the train set (10-fold CV; though accuracy is not significantly higher) and on the test set (no significance testing possible since only single set of performance scores are available per model; see Appendix E for procedure for submitting challenge predictions). Based on feature differentiation analysis (Section 7.1), we hypothesize that good performance with a text-focused BERT model on this speech classification task is due to the strong utility of linguistic features on this dataset. BERT captures a range of linguistic phenomena due to its training methodology, potentially encap-

Table 2: *Feature differentiation analysis results based on ADReSS train set. μ_{AD} and μ_{non-AD} show the means of the 13 significantly different features at p<9e-5 (after Bonferroni correction) for the AD and non-AD group respectively. We also show Spearman correlation between MMSE score and features, and regression weights of the features associated with the five greatest and five lowest regression weights from a ridge regressor (25 features, $\alpha = 12$).* next to correlation indicates significance at p<9e-5.*

Feature	**Feature type**	μ_{AD}	μ_{non-AD}	**Correlation**	**Weight**
Average cosine distance between utterances	Semantic	0.91	0.94	-	-
Fraction of pairs of utterances below a similarity threshold (0.5)	Semantic	0.03	0.01	-	-
Average cosine distance between 300-dimensional word2vec [29] utterances and picture content units	Semantic (content units)	0.46	0.38	-0.54*	-1.01
Distinct content units mentioned: total content units	Semantic (content units)	0.27	0.45	0.63*	1.78
Distinct action content units mentioned: total content units	Semantic (content units)	0.15	0.30	0.49*	1.04
Distinct object content units mentioned: total content units	Semantic (content units)	0.28	0.47	0.59*	1.72
Average cosine distance between 50-dimensional GloVe utterances and picture content units	Semantic content units)	-	-	-0.42*	-0.03
Average word length (in letters)	Lexico-syntactic	3.57	3.78	0.45*	1.07
Proportion of pronouns	Lexico-syntactic	0.09	0.06	-	-
Ratio (pronouns):(pronouns+nouns)	Lexico-syntactic	0.35	0.23	-	-
Proportion of personal pronouns	Lexico-syntactic	0.09	0.06	-	-
Proportion of RB adverbs	Lexico-syntactic	0.06	0.04	-0.41*	-0.41
Proportion of ADVP–>.RB amongst all rules	Lexico-syntactic	0.02	0.01	-0.37	-0.74
Proportion of non-dictionary words	Lexico-syntactic	0.11	0.08	-	-
Proportion of gerund verbs	Lexico-syntactic	-	-	0.37	1.08
Proportion of words in adverb category	Lexico-syntactic	-	-	-0.4*	-0.49

Table 3: *10-fold CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for BERT is higher, but not significantly so from SVM ($H = 0.4838, p > 0.05$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
SVM	10	0.796	0.81	0.78	0.82	0.79
NN	10	0.762	0.77	0.75	0.77	0.76
RF	50	0.738	0.73	0.76	0.72	0.74
NB	80	0.750	0.76	0.74	0.76	0.75
BERT	-	**0.818**	**0.84**	**0.79**	**0.85**	**0.81**

Table 4: *LOSO-CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for SVM is significantly higher than NN ($H = 4.50, p = 0.034$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	0.768	0.77	0.76	-	0.77
SVM	509	0.741	0.75	0.72	0.76	0.74
SVM	10	**0.870**	**0.90**	**0.83**	**0.91**	**0.87**
NN	10	0.836	0.86	0.81	0.86	0.83
RF	50	0.778	0.79	0.77	0.79	0.78
NB	80	0.787	0.80	0.76	0.82	0.78

sulating many important lexico-syntactic and semantic features. It is thus able to use information present in the lexicon, syntax, and semantics of transcribed speech after fine-tuning [32].

We see a trend of better performance while increasing the number of folds (see SVM in Table 4 and Table 3) in cross-validation. We postulate that this is due to the small size of the dataset, and hence differences in training set size in each fold.

7.3. Regression Weights

To assess the relative importance of individual input features for MMSE prediction, we report features with the 5 highest and 5 lowest regression weights in Table 2. Each value is the average weight assigned to features selected in each LOSO CV fold using ridge regression. We also present the correlation with MMSE score for these features, as well as their significance.We observe that for each of these highly weighted features, a positive or negative correlation is accompanied by a positive or negative regression weight, respectively. This demonstrates that even in the presence of other regressors, the relationship with MMSE score remains the same for these features. We also note that all 10 of these features are linguistic, further demonstrating that linguistic information is particularly distinguishing when it comes to predicting the severity of a patient's AD.

Table 5: *AD detection results on unseen, held-out ADReSS test set presented in same format as the baseline paper [1]. Bold indicates the best result.*

Model	#Features	Class	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	non-AD	0.750	0.70	**0.87**	-	0.78
		AD		**0.83**	0.62	-	0.71
SVM	10	non-AD	0.813	0.83	0.79	0.83	0.81
		AD		0.80	0.83		0.82
NN	10	non-AD	0.771	0.78	0.75	0.78	0.77
		AD		0.76	0.79		0.78
RF	50	non-AD	0.750	0.71	0.83	0.71	0.77
		AD		0.80	0.67		0.73
NB	80	non-AD	0.729	0.69	0.83	0.69	0.75
		AD		0.79	0.63		0.70
BERT	-	non-AD	**0.833**	**0.86**	0.79	**0.86**	**0.83**
		AD		0.81	**0.88**		**0.84**

Table 6: *LOSO-CV MMSE regression results on the ADReSS train and test sets. Bold indicates the best result.*

Model	#Features	α	RMSE Train set	MAE Train set	RMSE Test set
Baseline [1]	-	-	**4.38**		5.20
LR	15	-	5.37	4.18	4.94
LR	20	-	4.94	3.72	-
Ridge	509	12	6.06	4.36	-
Ridge	35	12	4.87	3.79	**4.56**
Ridge	25	10	4.56	**3.50**	-

8. Conclusions

In this paper, we compare two widely used approaches – explicit features engineering based on domain knowledge, and transfer learning using a fine-tuned BERT [11] classification model. Our results show that pre-trained models that are fine-tuned for the AD classification task are capable of performing well, outperforming hand-crafted feature engineering. In the future, we will experiment with different language representation models, and with different tokenization and encoding strategies for transcript representations. A direction for future work is also developing models that combine representations from language representation models like BERT and hand-crafted features [33]. Such feature-fusion approaches could potentially boost performance on the cognitive impairment detection task.

9. Acknowledgements

This paper benefited greatly from feedback and review from multiple people. Most notably, Dr. Jessica Robin (Winterlight Labs), Jordan Ponn (Winterlight Labs), Liam Kaufman (Winterlight Labs) and Maria Yancheva (Winterlight Labs).
Dr. Frank Rudzicz is supported by a CIFAR Chair in AI.

10. References

[1] 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.

[2] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[3] G. Prabhakaran, R. Bakshi *et al.*, "Analysis of structure and cost in a longitudinal study of alzheimer's disease," *Journal of Health Care Finance*, 2018.

[4] E. A. Jammeh, B. C. Camille, W. P. Stephen, J. Escudero, A. Anastasiou, P. Zhao, T. Chenore, J. Zajicek, and E. Ifeachor, "Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study," *BJGP Open*, p. bjgpopen18X101589, 2018.

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

[6] 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.

[7] Z. Zhu, J. Novikova, and F. Rudzicz, "Semi-supervised classification by reaching consensus among modalities," *arXiv preprint arXiv:1805.09366*, 2018.

[8] Z. Noorian, C. Pou-Prom, and F. Rudzicz, "On the importance of normative data in speech-based assessment," *arXiv preprint arXiv:1712.00069*, 2017.

[9] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 2 (Short Papers)*, 2018, pp. 701–707.

[10] J. R. Cockrell and M. F. Folstein, "Mini-mental state examination," *Principles and practice of geriatric psychiatry*, pp. 140–141, 2002.

[11] 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.

[12] S. O. Orimaye, K. Y. Tai, J. S.-M. Wong, and C. P. Wong, "Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams," *arXiv preprint arXiv:1511.02436*, 2015.

[13] A. Balagopalan, J. Novikova, F. Rudzicz, and M. Ghassemi, "The effect of heterogeneous data for alzheimer's disease detection from speech," *arXiv preprint arXiv:1811.12254*, 2018.

[14] 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.

[15] 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.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł . Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "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.

[18] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014.

[19] J. Chae and A. Nenkova, "Predicting the fluency of text with shallow structural features: Case studies of machine tanslation and human-written text," 2009.

[20] N. B. Mota, N. A. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech graphs provide a quantitative measure of thought disorder in psychosis," *PloS one*, vol. 7, no. 4, 2012.

[21] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[22] H. Ai and X. Lu, "A web-based system for automatic measurement of lexical complexity," in *27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June*, 2010, pp. 8–12.

[23] B. H. Davis and M. Maclagan, "Examining pauses in alzheimer's discourse," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 24, no. 2, pp. 141–154, 2009.

[24] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv, abs/1910.03771*, 2019.

[26] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3721–3731.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[30] 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.

[31] 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*, 2019, pp. 1431–1441.

[32] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3651–3657.

[33] M. Yu, M. R. Gormley, and M. Dredze, "Combining word embeddings and feature embeddings for fine-grained relation extraction," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, 2015, pp. 1374–1379.

[34] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International journal of corpus linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS 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

The ADReSS Challenge at INTERSPEECH 2020 defines a shared task through which different approaches to the automated recognition of Alzheimer's dementia based on spontaneous speech can be compared. ADReSS provides researchers with a benchmark speech dataset which has been acoustically pre-processed and balanced in terms of age and gender, defining two cognitive assessment tasks, namely: the Alzheimer's speech classification task and the neuropsychological score regression task. In the Alzheimer's speech classification task, ADReSS challenge participants create models for classifying speech as dementia or healthy control speech. In the the neuropsychological score regression task, participants create models to predict mini-mental state examination scores. This paper describes the ADReSS Challenge in detail and presents a baseline for both tasks, including feature extraction procedures and results for classification and regression models. ADReSS aims to provide the speech and language Alzheimer's research community with a platform for comprehensive methodological comparisons. This will hopefully contribute to addressing the lack of standardisation that currently affects the field and shed light on avenues for future research and clinical applicability.

Index Terms: Cognitive Decline Detection, Affective Computing, computational paralinguistics

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that entails a long-term and usually gradual decrease of cognitive functioning [1]. It is also the most common underlying cause for dementia. The main risk factor for AD is age, and therefore its greatest incidence is amongst the elderly. Given the current demographics in the Western world, where the population aged 65 years or more has been predicted to triple between years 2000 and 2050 [2], institutions are investing considerably on dementia prevention, early detection and disease management. There is a need for cost-effective and scalable methods that are able to identify the most subtle forms of AD, from the preclinical stage of Subjective Cognitive Decline (SCD), to more severe conditions like Mild Cognitive Impairment (MCI) and Alzheimer's Dementia (AD) itself.

Whilst memory is often considered the main symptom of AD, language is also deemed as a valuable source of clinical information. Furthermore, the ubiquity of speech has led to a number of studies investigating speech and language features for the detection of AD, such as [3, 4, 5, 6] to cite some examples. Although these studies propose various signal processing and machine learning methods for this task, the field still lacks balanced and standardised datasets on which these different approaches could be systematically compared.

Consequently, the main objective of the ADReSS Challenge of INTERSPEECH 2020 is to define a shared task through which different approaches to AD detection, based on spontaneous speech, could be compared. This aims to address one of the main problems of this active research field, the lack of standardisation, which hinders its translation into clinical practice. The ADReSS Challenge will therefore: 1) target a difficult automatic prediction problem of societal and medical relevance, namely, the detection of cognitive impairment and Alzheimer's Dementia (AD); 2) to provide a forum for those different research groups to test their existing methods (or develop novel approaches) on a new shared standardized dataset; 3) mitigate 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; and 4) focus on AD recognition using spontaneous speech, rather than speech samples that are collected under laboratory conditions.

To the best of our knowledge, this will be the first such shared-task focused on AD. Unlike some tests performed in clinical settings, where short speech samples are collected under controlled conditions, this task focuses on AD recognition using spontaneous speech. While a number of researchers have proposed speech processing and natural language processing approaches to AD recognition through speech, their studies have used different, often unbalanced and acoustically varied datasets, consequently hindering reproducibility, replicability, and comparability of approaches. The ADReSS Challenge will provide a forum for those different research groups to test their existing methods (or develop novel approaches) on a shared dataset which consists of a statistically balanced, acoustically enhanced set of recordings of spontaneous speech sessions along with segmentation and detailed timestamped transcriptions. The use of spontaneous speech also sets the ADReSS Challenge apart from tests performed in clinical settings where short speech samples are collected under controlled conditions which are arguably less suitable for the development of large-scale monitoring technology than spontaneous speech [7].

As data scarcity and heterogeneity have hindered research into the relationship between speech and AD, the ADReSS Challenge provides researchers with the very first available benchmark, acoustically pre-processed and balanced in terms of age and gender. ADReSS defines two different prediction tasks: (a) the *AD recognition task*, which requires researchers to model participants' speech data to perform a binary classification of speech samples into AD and non-AD classes; and (b) the *MMSE prediction task*, which requires researchers to create regression models of the participants' speech in order to predict their scores in the Mini-Mental State Examination (MMSE).

http://dx.doi.org/10.21437/Interspeech.2020-2571

This paper presents baselines for both tasks, including feature extraction procedures and initial results for a classification and a regression model.

2. ADReSS Challenge Dataset

A dataset has been created for this challenge which is matched for age and gender, as shown in Table 1 and Table 2, so as to minimise risk of bias in the prediction tasks. The data consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [8, 9]. Transcripts were annotated using the CHAT coding system [10]. The recorded speech has been segmented for voice activity using a simple voice activity detection algorithm based on signal energy threshold. We set the log energy threshold parameter to 65 dB with a maximum duration of 10 seconds per speech segment. The segmented dataset contains 1,955 speech segments from 78 non-AD subjects and 2,122 speech segments from 78 AD subjects. The average number of speech segments produced by each participant was 24.86 (standard deviation $sd = 12.84$). 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.

Table 1: *ADReSS Training Set: Basic characteristics of the patients in each group (M=male and F=female).*

	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
$[50, 55)$	1	0	30.0 (n/a)	1	0	29.0 (n/a)
$[55, 60)$	5	4	16.3 (4.9)	5	4	29.0 (1.3)
$[60, 65)$	3	6	18.3 (6.1)	3	6	29.3 (1.3)
$[65, 70)$	6	10	16.9 (5.8)	6	10	29.1 (0.9)
$[70, 75)$	6	8	15.8 (4.5)	6	8	29.1 (0.8)
$[75, 80)$	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

Table 2: *Characteristics of the ADReSS test set.*

	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
$[50, 55)$	1	0	23.0 (n.a)	1	0	28.0 (n.a)
$[55, 60)$	2	2	18.7 (1.0)	2	2	28.5 (1.2)
$[60, 65)$	1	3	14.7 (3.7)	1	3	28.7 (0.9)
$[65, 70)$	3	4	23.2 (4.0)	3	4	29.4 (0.7)
$[70, 75)$	3	3	17.3 (6.9)	3	3	28.0 (2.4)
$[75, 80)$	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

3. Acoustic and Linguistic Features

Acoustic feature extraction was performed on the speech segments using the openSMILE v2.1 toolkit which is an open-source software suite for automatic extraction of features from speech, widely used for emotion and affect recognition in speech [11], and with in-house software [12]. As the purpose of this paper is to describe the prediction tasks and set simple baselines that can be attained without extensive optimisation, we did not perform any feature set reduction procedures. The following is a brief description of the acoustic feature sets used in the experiments described in this paper:

emobase: This feature set contains the mel-frequency cepstral coefficients (MFCC) voice quality, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP) and intensity features with their first and second order derivatives. Several statistical functions are applied to these features, resulting in a total of 988 features for every speech segment [11].

ComParE: The *ComParE 2013* [13] feature set includes energy, spectral, MFCC, and voicing related low-level descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, bringing the total to 6,373 features.

eGeMAPS: The *eGeMAPS* [14] feature set 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 [15]. 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, for a total of 88 features per speech segment.

MRCG functionals: Multi-resolution Cochleagram features (MRCGs) were proposed by Chen et al. [16] and have since been used in speech related applications such as voice activity detection [17], speech separation [16], and more recently for attitude recognition [18]. MRCG features are based on cochleagrams [19]. A cochleagram is generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain so as to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of the audio signal. Four cochleagram features were generated at different levels of resolution. The high resolution level encodes local information while the remaining three lower resolution levels capture spectrotemporal information. A total of 768 features were extracted from each frame: 256 MRCG features (frame length of 20 ms and frame shift of 10 ms), along with 256 Δ MRCG and 256 $\Delta\Delta$ MRCG features. The statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) were applied on the 768 MRCG features for a total of 6,912 features.

Minimal: this feature set consists of basic statistics (mean, standard deviation, median, minimum and maximum) of the duration of vocalisations and pauses and speech rate, and a vocalisation count, similarly to [7].

In sum, we extracted 88 eGeMAPS, 988 emobase, 6,373 ComParE, 6,912 MRCG, and 13 minimal features from 4,077 speech segments. Excepting the minimal feature set, Pearson's correlation test was performed to remove acoustic features that were significantly correlated with duration (when $|R| > 0.2$). Hence, 72 eGeMAPS, 599 emobase, 3,056 ComParE, and 3,253 MRCG features were not correlated with the duration of the speech chunks, and were therefore selected for the machine learning experiments. Examples of features from the ComParE feature set by the above described procedure include L1-norms of segment length functionals smoothed by a moving average filter (including their means, maxima and standard deviations), and the relative spectral transform applied to auditory spectrum (RASTA) functionals (including the percentage of time the signal is above 25%, 50% and 75% of range plus minimum).

In addition, we used the EVAL command in the CLAN program [20] to compute a basic set of 34 language outcome measures (e.g., duration, total utterances, MLU, type-token ratio, open-closed class word ratio, percentages of 9 parts of speech) on the CHAT transcripts.

4. AD classification task

The AD classification task consists of creating a binary classification models to distinguish between AD and non-AD patient speech. These models may use speech data, transcribed speech, or both. Any methodological approach may be taken, but participants will work with the same dataset. The evaluation metric for this task are Accuracy $= \frac{TN+TP}{N}$, precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and $\mathrm{F}_1 = 2\frac{\pi \times \rho}{\pi+\rho}$, where N is the number of patients, TP, FP and FN are the number of true positives, false positives and false negatives, respectively.

4.1. Baseline classification

We performed our baseline classification experiments using five different methods, namely linear discriminant analysis (LDA), decision trees (DT, with leaf size of 20 and the CART algorithm), nearest neighbour (1NN, for KNN with K=1), random forests (RF, with 50 trees and a leaf size of 20) and support vector machines (SVM, with a linear kernel with box constraint of 0.1, and sequential minimal optimisation solver). The classification methods were implemented in MATLAB [21] using the statistics and machine learning toolbox. A leave-one-subject-out (LOSO) cross-validation setting was adopted, where the training data do not contain any information from validation subjects.

Two-step classification experiments were conducted to detect cognitive impairment due to AD (as shown in Figure 1). This consisted of segment-level (SL) classification, where classifiers were trained and tested to predict whether a speech segment was uttered by a non-AD or AD patient, and majority vote (MV) classification, which assigned each subject a class label based on the majority labels of SL classification.

4.2. Results

The classification accuracy is shown in Tables 3 and 4 for LOSO and test settings respectively. These results show that the 1NN (0.574) provides the best accuracy for acoustic features using ComParE set for AD detection, with accuracy above the chance level of 0.50. From the results shown in Table 3, we note that even though 1NN provides the best result (0.574), DT and LDA also exhibit promising performance, being in fact more stable across all feature sets than the other classifiers (the best average accuracy of 0.559 for LDA and 0.570 for DT). We also note that Minimal, ComParE and linguistic also exhibit promising performance, being in fact more stable across all classifiers than the other features (the best average accuracy of 0.552 for Minimal, 0.541 for Compare and 0.713 for linguistic). Based on these findings we have selected the LDA model trained using ComParE as our baseline model for acoustic features.

Table 4 shows that 1NN provides less accurate results on the test set than in LOSO cross validation. However, the results of LDA (0.625) and DT (0.625) improve on the test data for acoustic features. The linguistic features provide an accuracy of 0.75, which is better than automatically extracted acoustic features though it relies on manual transcription. The challenge baseline accuracy for the classification task are therefore 0.625 for acoustic features and 0.75 for linguistic features. The precision, recall and F1 Score are reported in Table 5.

Table 3: *AD classification accuracy on LOSO cross validation.*

Features	LDA	DT	1NN	SVM	RF	mean
emobase	0.500	0.519	0.398	0.491	0.472	0.476
ComParE	**0.565**	0.528	0.574	0.528	0.509	**0.541**
eGeMAPS	0.482	0.500	0.380	0.333	0.482	0.435
MRCG	0.519	0.500	0.482	0.528	0.509	0.507
Minimal	0.519	0.667	0.426	0.565	0.583	0.552
linguistic	**0.768**	0.704	0.740	0.602	0.750	**0.713**
mean	**0.559**	**0.570**	0.500	0.508	0.551	–

Table 4: *AD classification accuracy on test set.*

Features	LDA	DT	1NN	SVM	RF	mean
emobase	0.542	0.688	0.604	0.500	0.729	0.613
ComParE	**0.625**	0.625	0.458	0.500	0.542	0.550
eGeMAPS	0.583	0.542	0.688	0.563	0.604	0.596
MRCG	0.542	0.563	0.417	0.521	0.542	0.517
Minimal	0.604	0.562	0.604	0.667	0.583	0.604
linguistic	**0.750**	0.625	0.667	0.792	0.750	0.717
mean	0.608	0.601	0.573	0.590	0.625	–

Table 5: *Baseline results of AD classification task using the LDA classifier with acoustic and linguistic features.*

	class	Precision	Recall	F1 Score	Accuracy
$LOSO_{Acous}$	non-AD	0.56	0.61	0.58	0.56
	AD	0.57	0.52	0.54	
$TEST_{Acous}$	non-AD	0.67	0.50	0.57	0.62
	AD	0.60	0.75	0.67	
$LOSO_{ling}$	non-AD	0.76	0.78	0.77	0.77
	AD	0.77	0.76	0.77	
$TEST_{ling}$	non-AD	0.70	0.87	0.78	0.75
	AD	0.83	0.62	0.71	

5. MMSE prediction task

The MMSE prediction task consists of generating a regression model for prediction of MMSE scores of individual participants from the AD and non-AD groups. Unlike classification, MMSE prediction is relatively uncommon in the literature, despite MMSE scores often being available. While models may use speech (acoustic) or linguistic data individually or in combination, the baseline described here report results of acoustic and linguistic models built separately.

5.1. Baseline regression

We performed our baseline regression experiments using five different methods, namely decision trees (DT, with leaf size of 20 and CART algorithm), linear regression (LR), gaussian process regression (GPR, with a squared exponential kernel), least-squares boosting (LSBoost, which contains the results of boosting 100 regression trees) and support vector machines (SVM, with a radial basis function kernel with box constraint of 0.1, and sequential minimal optimisation solver). The regression methods are implemented in MATLAB [21] using the statistics and machine learning toolbox. As with classification, the regression experiments were conducted in two steps for acoustic features (Figure 1), with SL regression followed by averaging of predicted MMSE values.

5.2. Results

The regression results are reported as root mean squared error (RMSE) scores in Tables 6 and 7 for LOSOCV and test data. These results show that DT (7.28) provides the best RMSE using MRCG features for MMSE prediction with $r = -0.759$, being more stable across all acoustic feature sets than the other classifiers (the best average RMSE of 6.86 for DT). We also note that Minimal and eGeMaPs also exhibit promising performance, with RMSE of 7.46 and 8.02 respectively across models.

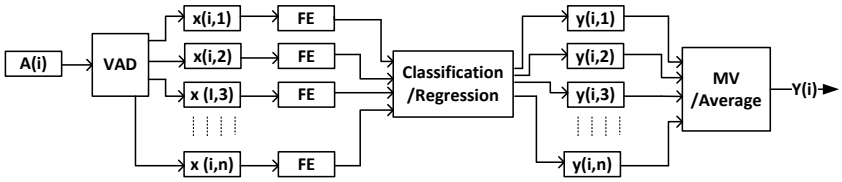


Figure 1: *System Architecture:* $A(i)$*, the recording of is segmented using voice activity detection (VAD) into* n *segments* $x(i,n)$*. Acoustic feature extraction (FE) is performed at segment level. The output of classification or regression for the* n^{th} *segment of the* i^{th} *recording is denoted* $y(i,n)$*. MV outputs the majority voting for classification, and Average the mean regression score.*

Based on this, the DT model trained using the MRCG feature was chosen as the baseline model for the regression task for acoustic features. For linguistic features, we selected the DT model as baseline with RMSE of 4.38 ($r = 0.792$).

Table 7 shows the results of regression methods on test data. The baseline model (DT with MRCG features) provides an RMSE of 6.14 ($r = 0.22$) in the test setting. Hence the challenge baseline accuracy for this task is 6.14 for acoustic features. The linguistic feature model provides an RMSE of 5.20 ($r = 0.57$), which therefore corresponds to the ADReSS challenge baseline accuracy for linguistic features in this task.

Table 6: *MMSE prediction LOSO cross Validation results. the chance level is RMSE of 7.18*

Features	Linear	DT	GP	SVM	LSBoost	mean
emobase	7.44	7.29	7.71	7.71	8.33	7.70
ComParE	15.69	7.29	7.67	7.63	7.84	9.22
eGeMAPS	8.08	7.31	7.72	8.55	8.68	8.07
MRCG	13.46	**7.28,** $r = -0.76$	7.65	7.50	8.02	8.78
Minimal	7.39	7.60	7.18	8.01	7.14	7.46
Linguistic	6.15	**4.38,** $r = 0.79$	7.92	6.34	7.44	6.45
mean	9.70	**6.86**	7.64	7.62	7.91	–

Table 7: *MMSE prediction test results.*

Features	Linear	DT	GP	SVM	LSBoost	mean
emobase	6.80	6.78	6.36	6.18	6.73	6.57
ComParE	6.47	6.52	6.33	6.19	6.72	6.45
eGeMAPS	6.90	5.99	6.28	6.12	6.41	6.34
MRCG	6.70	**6.14,** $r = 0.22$	6.33	6.20	6.31	6.33
Minimal	6.29	6.84	6.58	6.19	7.71	6.72
Linguistic	4.78	**5.20,** $r = 0.57$	5.54	6.24	6.62	5.68
mean	6.32	**6.25**	6.24	6.19	6.75	–

6. Discussion

These results of the classification baseline are comparable to those attained by models based on speech recordings available from spontaneous speech samples in DementiaBank's Pitt corpus [8], which is widely used. Accuracy scores of 81.92%, 80% and 79% and 64% have been reported in the literature [3, 22, 23, 7]. Although these studies report higher accuracy than the baselines presented here, all of those studies (except [7]) combined information from the manual transcripts with acoustic data, and were conducted on an unbalanced dataset (in terms of age, gender and number of subjects in the AD and non-AD classes). It is also worth noting that accuracy for the best performing of these models drops to 58.5% when feature selection is not performed on their original set of 370 linguistic and acoustic features [3]. Models that relied only on acoustic features were reported in [7] (64% accuracy) and [23] (62% accuracy, using an SVM model). It is also noted that previous studies do not evaluate their methods in a complete subject-independent setting (i.e. they consider multiple sessions for a subject and classify a session instead of a subject). This could lead to overfitting, as the model might learn speaker dependent features from a session and then, based on those features, classify the next session of the same speaker.

One strength of our method is its speaker independent nature. Ambrosini et al. reported an accuracy of 80% while using acoustic (pitch, unvoiced duration, shimmer, pause duration, speech rate), age and educational level features for cognitive decline detection using an Italian dataset of an episodic story telling setting [24]. However, this dataset is less easily comparable to ours, as it is elicited differently, and is not age and gender balanced.

Yancheva and colleagues [25] predicted MMSE scores with speech-related features using the full DementiaBank Pitt dataset, which is not balanced and includes longitudinal observations. Their model yielded a mean absolute error (MAE) of 3.83 in predicting MMSE. However, they employed lexicosyntactic and semantic features derived from manual transcription, rather than automatically extracted acoustic features as we used in our analysis. In [25], those linguistic features were the main features selected from a group of 477, with acoustic features typically not being among the most relevant. Therefore no quantitative results were reported for acoustic features.

7. Conclusions

This paper described the ADReSS challenge, and set simple baselines for its tasks, demonstrating the relevance of acoustic and linguistic features of spontaneous speech for cognitive impairment detection in the context of Alzheimer's Disease diagnosis and MMSE prediction. Machine learning methods operating on automatically extracted voice features provide a baseline accuracy of up to 62.5% on the AD classification task, while linguistic features extracted from manually produced transcripts yielded 76.85% accuracy on the same task. These results are well above the chance level of 50%. A baseline RMSE of 6.14 and 5.21 for acoustic and linguistic features respectively on test has been established for the MMSE regression task. It is reasonable to expect that the ADReSS Challenge's participants will attain better accuracy scores by employing further pre-processing, feature set reduction, and more complex models than the ones employed in this paper. By bringing the research community together in order to work on a shared task on the same dataset, ADReSS intends to make comprehensive methodological comparisons. This will hopefully highlight research caveats and shed light on avenues for clinical applicability and future research directions.

8. Acknowledgements

This research is funded by the European Union's Horizon 2020 research programme, under grant agreement 769661, SAAM project. SdlFG is supported by the Medical Research Council.

9. References

[1] American Psychiatric Association, "Delirium, dementia, and amnestic and other cognitive disorders," in *Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR)*, 2000, ch. 2.

[2] World Health Organization, "Mental health action plan 2013-2020," *WHO Library Cataloguing-in-Publication DataLibrary Cataloguing-in-Publication Data*, pp. 1–44, 2013.

[3] 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.

[4] S. Luz, S. D. la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," in *Procs. of LREC'18*, D. Kokkinakis, Ed. Paris, France: ELRA, may 2018.

[5] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations." in *INSTERSPEECH*, 2018, pp. 1893–1897.

[6] 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.

[7] S. Luz, "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *Procs. of the Intl. Symp on Comp. Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[8] 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.

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

[10] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press, 2014.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Procs. of ACM-MM*. ACM, 2010, pp. 1459–1462.

[12] S. Luz, *vocaldia: Create and Manipulate Vocalisation Diagrams*, 2018, R package version 0.8.3. [Online]. Available: https://cran.r-project.org/package=vocaldia

[13] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *ACM-MM*. ACM, 2013, pp. 835–838.

[14] 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.

[15] ——, "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.

[16] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[17] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, 2018.

[18] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *Procs. of ICASSP*, 2019, pp. 3737–3741.

[19] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[20] B. MacWhinney, *Tools for analyzing talk part 2: The CLAN program*, Carnegie Mellon University, Pittsburgh, PA, 2017, retrieved from http://talkbank.org/manuals/CLAN.pdf.

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

[22] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimers disease," in *Procs. of ACL*, 2016, pp. 2337–2346.

[23] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimers disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagn., Asses. & Dis. Mon.*, vol. 10, pp. 260–268, 2018.

[24] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid *et al.*, "Automatic speech analysis to early detect functional cognitive decline in elderly population," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 212–216.

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

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velazquez, Najim Dehak

Center for Language Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{rpappag1,jcho52,laureano,ndehak3}@jhu.edu

Abstract

In this study, we analyze the use of state-of-the-art technologies for speaker recognition and natural language processing to detect Alzheimer's Disease (AD) and to assess its severity predicting Mini-mental status evaluation (MMSE) scores. With these purposes, we study the use of speech signals and transcriptions. Our work focuses on the adaptation of state-of-the-art models for both modalities individually and together to examine its complementarity. We used x-vectors to characterize speech signals and pre-trained BERT models to process human transcriptions with different back-ends in AD diagnosis and assessment. We evaluated features based on silence segments of the audio files as a complement to x-vectors. We trained and evaluated our systems in the Interspeech 2020 ADReSS challenge dataset, containing 78 AD patients and 78 sex and age-matched controls. Our results indicate that the fusion of scores obtained from the acoustic and the transcript-based models provides the best detection and assessment results, suggesting that individual models for two modalities contain complementary information. The addition of the silence-related features improved the fusion system even further. A separate analysis of the models suggests that transcript-based models provide better results than acoustic models in the detection task but similar results in the MMSE prediction task.

1. Introduction

Alzheimers Disease (AD) is the most common cause of dementia and the most prevalent neurodegenerative condition. Its impact on the multiple aspects of society is rising due to the aging of the worldwide population [1]. While two of the most typical signs of AD are memory and cognitive decline, the literature suggests that language impairment is also a common sign that can be employed to support diagnosis and assessment of the severity of the disease, given that speech and language production can provide information about the cognitive status of a person and other aspects related to brain damage. Although the human evaluation of speech and language can be used to diagnose and assess patients in the clinical setting, that type of evaluation does not allow an objective quantitative analysis and reliable repeatability. To this respect, the use of speech recognition and Natural Language Processing (NLP) techniques can deliver new precision medicine tools that will provide objective measures and biomarkers. This will allow faster diagnosis and assessment in a non-invasive and cost-effective manner.

Although the influence of AD in speech and language is diverse and subject-dependent, the literature suggests some common signs such as progressive, logopenic or anomic aphasia [2, 3, 4] (communication and word retrieval impairment, phone substitution) and apraxia of speech [5] (articulatory impairment.) Therefore, several studies indicate that both phonetic-motor signs (related to apraxia) and phonological-linguistic manifestations (related to aphasia and anomia) can be found in cohorts of AD patients [5]. Depending on the patient, the apraxic or aphasic manifestations can be prevalent, suggesting that both acoustic and linguistic analyses are advisable in systems employing speech technologies automatically to detect AD or assess its severity.

In this respect, the combination of acoustic and linguistic features within machine learning based-approaches to automatically detect AD in recordings obtained from the DementiaBank corpus has already been analyzed [6], obtaining 81% cross-validation accuracy. Other studies providing similar results suggest that linguistic features provide higher accuracy than acoustic features in detecting AD [7]. However, the combination of both types of features yields better results than when using these features separately, suggesting that these features are complementary [7]. Additionally, accuracies over 80% have been reported when employing word and silence rates obtained with Voice Activity Detection (VAD) systems and transcripts [8]. Moreover, some linguistic features indicative of lexical diversity such as word frequency, percentage of content words, pronoun ratio or type-token ratio among others have shown a high correlation with Mini-Mental Status Examination (MMSE) in AD patients [9], suggesting that patient's morphosyntactic impairments can be automatically analyzed and employed for severity assessment.

Although the literature includes a fair amount of studies employing acoustic and linguistic features [6, 7, 8, 9, 10, 11, 12] for the automatic detection and assessment of AD, to our knowledge no study analyzes the use of speaker recognition and NLP technologies such as x-vectors [13] and Bidirectional Encoder Representations from Transformers (BERT) [14]. These techniques have become the state-of-the-art in speech technologies, and its acoustic and linguistic characterization properties have been exploited in multiple scenarios such as Parkinson's Disease (PD) detection [15], emotion recognition [16], sentiment analysis [17] or question answering [14], among others.

Consequently, this study aims to analyze the use of these two Deep Neural Networks (DNN)-based techniques, x-vectors and BERT, in AD detection and MMSE prediction scenarios.

2. ADReSS Challenge Dataset

The ADReSS Challenge dataset [18] contains two subsets with speech and transcriptions from speakers with and without AD: the *training* and the *evaluation subsets*. In this study, the *training subset* was used to perform cross-validation and to train models to be evaluated with the *evaluation subset*.

The *training subset* includes two groups of speakers: those diagnosed with AD (AD group) and the age- and sex-matched control speakers (CC group). Each group is composed of 24 male and 30 female participants. Data in both groups contain one audio recording per participant, recorded at 44100 Hz and

http://dx.doi.org/10.21437/Interspeech.2020-2587

with an average length of 72.10 s, demographic information, full transcript, and MMSE score. In our experiments, we down-sampled the recordings according to the models we used, as explained in later sections.

The *evaluation subset* comprises 11 male and 13 female participants in each group, while the age distribution is the same over the two groups. The average session length is 82.51 s. Challenge participants do not have information about AD diagnosis or MMSE assessment for these speakers.

3. Experimental Setup

In this study, we employed two main models to detect AD and predict MMSE from speech. The first model or acoustic model is based on the use of acoustic aspects of speech and employs a speaker characterization technique, i.e., x-vectors and two different back-ends: Probabilistic Linear Discriminant Analysis (PLDA) for detection and Support Vector Regression (SVR) for MMSE prediction. The x-vectors were complemented with heuristic features obtained from the analysis of the silence and pause segments from the speech signal. The second model or transcript-based model is a BERT model that utilizes linguistic contents to detect AD subjects and predict MMSE. We hypothesize that the transcript-based model provides complementary information to the acoustic model. Finally, scores from the two approaches were fused using a Gradient Boosting Regressor (GBR) or averaging, depending on the task.

Moreover, we differentiate two types of results:

- Cross-validation results: obtained training and testing with the *training subset*, using a 10-fold scheme where class and age distributions were consistent over the folds. The cross-validation was done speaker-independently since the dataset has only one session recorded per participant.
- Evaluation results: obtained by testing the models trained with the *training subset* on the *evaluation subset*.

3.1. Acoustic model

3.1.1. x-vectors

To model the speakers' articulatory, prosodic and phonatory characteristics included in the dataset, we employ representation obtained with an x-vector model trained for speaker recognition. An x-vector model is a deep neural network that generates one single vector or embedding per utterance, characterizing the speaker. Although the technique is considered the current state-of-the-art for speaker recognition, several studies suggest that these embeddings also contain information related to emotion, speaking rate, gender [16, 19] and other articulatory, phonatory and prosodic information that can be used to characterize neurological diseases, as Parkinson's Disease [15]. In general terms, an x-vector model contains three main parts: an encoder network to extract frame-level representation from MFCC, a global temporal pooling layer to produce the embedding (x-vector), and a feed-forward classification network to produce speaker class posteriors. Once the model training is done, only the first two parts are used while the last part is discarded. In our case, the three parts consisted of a factorized time delay network encoder (F-TDNN), mean plus standard deviation pooling, and two feed-forward layers, respectively, as detailed in a previous study [15]. Differentiation process between AD and CC speakers followed the same setup as the one explained in the cited study:

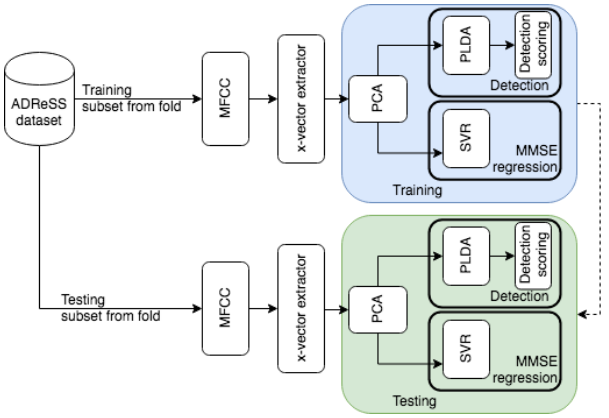


Figure 1: *Diagram of the acoustic model methodology. In cross-validation stage, models obtained with the training folds are used for testing with their respective testing folds. In evaluation stage, the whole training dataset is employed for training while the evaluation dataset is used for testing*

- First, all speech signals were normalized, low-pass filtered and re-sampled to 16 kHz.
- Then, we extracted MFCC features (40 coefficients, frame length of 25 ms with frame shift 10 ms)
- Silence segments were removed employing the standard VAD from Kaldi [20].
- MFCC features were used to extract one x-vector (dimension 512) for each speech recording using an x-vector model trained with VoxCeleb 1 and 2 corpora [21, 22] in Kaldi with sampling frequency 16 kHz.
- At each cross-validation iteration, all the x-vectors from the training folds were employed to train a Principal Component Analysis (PCA) model that was applied to the x-vectors from the training and testing folds in the cross-validation stage.
- For AD detection, x-vector PCA-transformed coefficients from the training folds were used to train a PLDA classifier to differentiate between AD and CC speakers. In the classifier, a likelihood ratio per speech recording is calculated considering two classes (AD and CC) which is employed in scoring to take the decision. The scoring threshold is set to the equal error rate point obtained with the log-likelihoods from the training folds x-vector-PCA coefficients.
- Similarly, for MMSE prediction, we trained and evaluated a linear SVR on the x-vector PCA-transformed coefficients.

Fig. 1 includes a diagram of the described process. To get the best PCA and PLDA models for evaluation on the *evaluation subset*, the whole ADReSS *training subset* was used.

3.1.2. Silence features

To complement the x-vectors characterization, which is data-driven, we also extracted 4-dimensional heuristic features based on the Kaldi energy-based VAD algorithm. Our goal was to characterize the presence of silences in the recordings. The four features are:

- Silence rate (the number of silence regions divided by the recording length)
- Ratio of silence to speech duration

- Mean and standard deviation of the duration of silence regions

We only considered silence regions that were longer than 150 ms. Also, we removed the silences at the start and end of the recordings when these existed. We considered these features since previous studies suggest that silence-related features can help to characterize aphasia and apraxia associated with AD [8]. We used these features in two different manners in this study:

- As single features for PLDA and SVR model training to examine the discrimination capabilities of these features.
- Appended to the x-vector PCA-transformed coefficients, which we denominate *Acoustic model with silence features* scheme. This allows us to observe the complementarity between x-vectors and silence features.

3.2. Transcript-based model

To model the linguistic-phonological manifestations of AD on speech, we employed a BERT model [14] on the spoken transcripts, which has shown state-of-the-art performances in several NLP applications such as question answering, natural language inference, named entity recognition, sentence, and word prediction, among many others. We chose BERT for two reasons: 1) the embeddings obtained from this model act as general text representation and, 2) previous studies reported good results from fine-tuned BERT models for multiple tasks. Two examples are depression detection [23] or sentiment analysis [17].

BERT is a pre-trained language model trained to predict masked words of a sentence and the next sentence. The BERT architecture mainly consists of self-attention layers and feed-forward layers. In general, a pre-trained BERT model is adapted to a down-stream task by fine-tuning the pre-trained parameters with the minimal number of newly introduced parameters for the task [14]. We adapted BERT to our tasks (AD detection and MMSE prediction) in a similar way:

- We replaced the last layer of the BERT model with a task-specific layer: a linear layer having two neurons with a softmax activation function for AD detection or a linear layer having 1 neuron with linear activation function for MMSE prediction.
- We fine-tuned the entire pre-trained model using our data to minimize cross-entropy loss for AD detection or mean square error for MMSE prediction.

The inputs of the model were tokens from the transcript that were tokenized into sub-words using WordPiece tokenizer [24]. These inputs were processed through multiple self-attention and feed-forward layers to obtain embeddings for each sub-word in the penultimate layer. Then, the sequence of sub-word embeddings was pooled to pass through a linear layer to obtain the prediction.

For each iteration of the cross-validation experiments, 9 folds from the *training subset* were employed for BERT fine-tuning and the remaining fold for testing. We used early stopping criterion to stop training the model and trained for 5 epochs.

3.3. Fusion

In this section, we describe our methodology for fusing acoustic and transcript model scores. For the AD detection task, we first obtained the scores from acoustic and transcript-based models for all utterances from the testing folds during the cross-validation stage. Then, we employed these predictions in a cross-validation scheme to train and test the fusion of the scores using a GBR model [25], which provided the cross-validation results. To obtain the *evaluation subset* predictions, we employed the scores from the whole *training subset* to train a final fusion GBR model that was used to perform the fusion of scores coming from the acoustic and transcript-based models for the challenge evaluation. For MMSE prediction, we followed a similar procedure but simply averaged the scores from the different models instead of using a GBR.

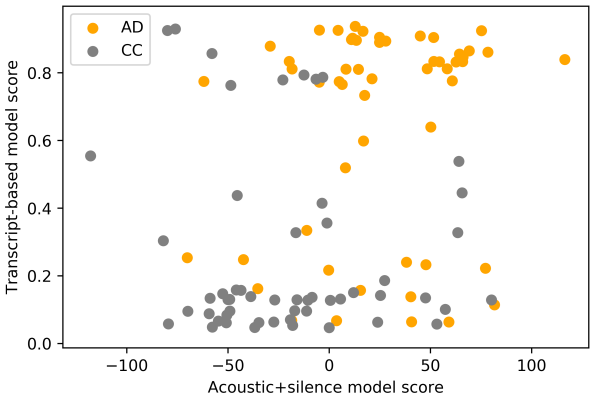


Figure 2: *Score scatterplot for AD and CC speakers in detection task considering the transcript-based model scores (that range between 0 and 1) and the log-likelihood ratio obtained with the PLDA classifier for the acoustic+silence model. Each dot represent one subject.*

4. Results and Discussion

In this section, we present our results on both AD detection and MMSE prediction tasks. For evaluation metrics, we used the same metrics as proposed in [18], namely, accuracy, precision, recall, and F1 score for detection and Root Mean Square Error (RMSE) for MMSE prediction. For simplicity, in cross-validation results (10 folds) we only report accuracy and RMSE.

4.1. Cross-validation results

Table 1 presents the cross-validation results with the proposed models for AD detection and MMSE prediction tasks. From the comparison of acoustic and transcript models, we can observe that the transcript-based model performed better than the acoustic model for AD detection but worse in MMSE prediction. The use of silence features alone did not provide high accuracy to differentiate between AD and CC groups. However, when we concatenated silence features with x-vectors PCA-transformed coefficients, denoted as Acoustic+silence in Table 1, we obtained an absolute 2.4% improvement in AD detection accuracy compared to using acoustic features alone, implying that acoustic and silence features may have complementary information. For MMSE prediction task, we obtained a small improvement in RMSE value after concatenation (0.03, absolute).

We further fused acoustic and transcript-based model scores to exploit their complementary information. The fusion model showed 79.2% accuracy and 5.93 RMSE, which indicates a 0.5% and 0.31 improvement compared to the best individual model, respectively. Thus, results suggest that score fu-

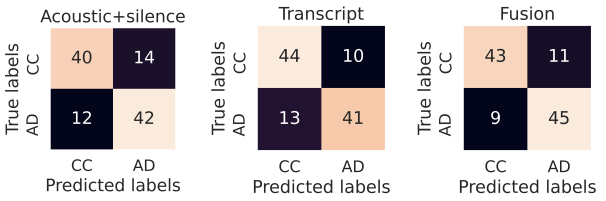
sion provides improvements in both AD detection and MMSE prediction. In the same sense, the fusion of Acoustic+silence and transcript models scores yielded 81.44% AD accuracy and 5.91 RMSE for MMSE prediction, the best cross-validation results.

A scatterplot of detection scores per subject is shown in Figure 2. The figure indicates that in the detection task, the transcript-based analysis is more informative for some speakers, while the acoustic signal analysis is so for some of the others. We can observe that for the majority of subjects, the scores from the two types of models help to cluster the two groups of speakers in the bottom left (CC) and upper right (AD) parts of the score bi-dimensional space, suggesting that both acoustic signal and transcripts contain cues to detect AD. Nevertheless, a few subjects have opposite results in different models, showing a high score from the transcript-based model but a low score from the acoustic model and vice versa. This indicates that different models can provide complementary information.

Figure 3 shows the confusion matrices of the models with the best cross-validation results. We can observe that the models are not biased to any single class, i.e., the recall for each class, AD and CC, is similar. Improvement in the fusion model is reflected with higher diagonal values and lower off-diagonal values in general, compared to the two individual models.

Table 1: *Cross-validation (CV) results for AD detection and MMSE precition tasks. Best results are marked in bold.*

Models	Detection CV accuracy (%)	Prediction RMSE
Acoustic	73.21	6.24
Silence	51.20	8.05
Acoustic+silence	75.93	6.21
Transcript	78.70	6.52
Acoustic & Transcript	79.20	5.93
Acoustic+silence & Transcript	**81.48**	**5.91**



(a) *(b)* *(c)*

Figure 3: *Confusion matrices for the detection tasks using (a) Acoustic model with silence features, (b) Transcript model, (c) Fusion of Acoustic model with silence features and transcript model.*

4.1.1. Evaluation results

Results for the *evaluation subset* were obtained from the submission of our model predictions to the ADReSS challenge organizers. Table 2 shows the evaluation results of our models in AD detection and MMSE prediction tasks. Baseline results are based on the use of the ComParE 2013 feature set [26] and a linear discriminant analysis classifier (for detection) and MRGC features [27] with decision trees (for MMSE prediction.) These baseline results were provided by the ADReSS challenge organizers [18]. We observed that four of our four models outperformed the baseline in the detection task by significant margins, and all of them provided a better RMSE than the baseline. The model comparison showed similar trends in accuracy on the evaluation and cross-validation results, but the overall accuracy was lower in the evaluation than the cross-validation. For MMSE prediction, all RMSE values are lower in the evaluation experiments than in the cross-validation. The model providing the best accuracy was the score-level fusion of acoustic and transcript models with 75% accuracy. When silence features were also used, we obtained the best MMSE prediction results, 5.32 RMSE.

We observed that the acoustic model performance in the *evaluation subset* is much lower than its correspondent cross-validation accuracy, suggesting that the acoustic models might have been overfitted to the *training subset*. We observed the same trends and conclusions from model comparison in cross-validation and evaluation experiments in Tables 1 and 2, as the complementarity between transcript and acoustic models.

Table 2: *ADReSS challenge evaluation results for the detection and prediction tasks. Best results are marked in bold.*

Models	Class	Detection Prec./Rec.	F1	Accuracy (%)	Prediction RMSE
Baseline	CC	0.67/0.50	0.57	62.50	6.14
	AD	0.60/0.75	0.67		
Acoustic	CC	0.61/0.45	0.52	58.00	6.08
	AD	0.57/0.71	0.63		
Acoustic + silence	CC	0.64/**0.75**	0.69	66.70	5.97
	AD	0.70/0.58	0.63		
Transcript	CC	0.79/0.63	0.7	72.92	5.86
	AD	0.69/0.83	0.75		
Acoustic & Transcript	CC	**0.83**/0.63	**0.71**	**75.00**	5.37
	AD	**0.70/0.88**	**0.78**		
Acoustic + silence & Transcript	CC	0.79/0.62	0.70	72.92	**5.32**
	AD	0.69/0.83	0.75		

5. Conclusions and future work

This study presents different approaches to automatically detect AD and predict MMSE from the speech signal and its associated transcript, based on the acoustic characterization of the speech signal and the transcript-based modeling employing DNN. The employed x-vectors and BERT are considered the current state-of-the-art techniques in speaker recognition and NLP, respectively. Our results suggest that transcription-based models provide better results in detection and prediction tasks than acoustic models. At the same time, information about the silences present in the recording improves accuracy for acoustic modeling. The best results in cross-validation and evaluation stages are obtained with the fusion of the scores provided by both the acoustic and transcript-based models.

In future work, we will explore the x-vector adaptation by fine-tuning the extractor [16] for the AD/CC detection task. Also, we will explore the use of automatic speech recognition systems to obtain the speech transcription and compare results with human transcription. Lastly, we will explore the use of BioBERT [28] and other transformer-based architectures for the detection and assessment of AD.

6. References

[1] K. B. Rajan, J. Weuve, L. L. Barnes, R. S. Wilson, and D. A. Evans, "Prevalence and incidence of clinically diagnosed alzheimer's disease dementia from 1994 to 2012 in a population study," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 1–7, 2019.

[2] J. D. Rohrer, M. N. Rossor, and J. D. Warren, "Alzheimer's pathology in primary progressive aphasia," *Neurobiology of aging*, vol. 33, no. 4, pp. 744–752, 2012.

[3] S. Ahmed, C. A. de Jager, A.-M. F. Haigh, and P. Garrard, "Logopenic aphasia in alzheimer's disease: clinical variant or clinical feature?" *J Neurol Neurosurg Psychiatry*, vol. 83, no. 11, pp. 1056–1062, 2012.

[4] S. M. Harnish, "Anomia and anomic aphasia: Implications for lexical processing." *The Oxford Handbook of Aphasia and Language Disorders*, 2018.

[5] E. Rochon, C. Leonard, and M. Goral, "Speech and language production in alzheimer's disease," *Aphasiology*, vol. 32, no. 1, pp. 1–3, 2018.

[6] 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.

[7] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "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.

[8] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[9] G. Kavé and A. Dassa, "Severity of alzheimer's disease and language features in picture descriptions," *Aphasiology*, vol. 32, no. 1, pp. 27–40, 2018.

[10] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 45–46.

[11] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *Proc. of the LREC 2018 Workshop "Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)",*, 2018.

[12] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[16] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *arXiv preprint arXiv:2002.05039*, 2020.

[17] S. Pei, L. Wang, T. Shen, and Z. Ning, "Da-bert: Enhancing part-of-speech tagging of aspect sentiment analysis using bert," in *International Symposium on Advanced Parallel Processing Technologies*. Springer, 2019, pp. 86–95.

[18] 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

[19] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[20] D. Povey, A. Ghoshal, and G. Boulianne, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech*, aug 2017, pp. 2616–2620.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[23] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[24] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[27] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[28] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition

Nicholas Cummins[1,2], Yilin Pan[3], Zhao Ren[1], Julian Fritsch[4,5], Venkata Srikanth Nallanthighal[6], Heidi Christensen[3], Daniel Blackburn[7], Björn W. Schuller[1,8], Mathew Magimai.-Doss[4], Helmer Strik[9], Aki Härmä[6]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Department of Biostatistics and Health Informatics, IoPPN, King's College London, London, UK
[3]Department of Computer Science, University of Sheffield, UK
[4]Idiap Research Institute, Martigny, Switzerland
[5]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[6]Philips Research, Eindhoven,The Netherlands
[7]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, UK
[8]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[9]Centre for Language Studies (CLS), Radboud University Nijmegen

nicholas.cummins@ieee.org

Abstract

In the light of the current COVID-19 pandemic, the need for remote digital health assessment tools is greater than ever. This statement is especially pertinent for elderly and vulnerable populations. In this regard, the INTERSPEECH 2020 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge offers competitors the opportunity to develop speech and language-based systems for the task of Alzheimer's Dementia (AD) recognition. The challenge data consists of speech recordings and their transcripts, the work presented herein is an assessment of different contemporary approaches on these modalities. Specifically, we compared a hierarchical neural network with an attention mechanism trained on linguistic features with three acoustic-based systems: (i) Bag-of-Audio-Words (BoAW) quantising different low-level descriptors, (ii) a Siamese Network trained on log-Mel spectrograms, and (iii) a Convolutional Neural Network (CNN) end-to-end system trained on raw waveforms. Key results indicate the strength of the linguistic approach over the acoustics systems. Our strongest test-set result was achieved using a late fusion combination of BoAW, End-to-End CNN, and hierarchical-attention networks, which outperformed the challenge baseline in both the classification and regression tasks.

Index Terms: Alzheimer's Disease, Bag-of-Audio-Words, Convolutional Neural Network, Siamese Network, Hierarchical Neural Network, Attention Mechanisms

1. Introduction

According to the *World Health Organisation* (WHO), dementia is a major cause of disability in the elderly population worldwide, with at least 10 million new cases reported every year [1]. Alzheimer's Disease (AD) is the most common cause of dementia [1, 2] and is a major public health concern, with considerable associated socio-economic costs [2]. Therefore, there is an urgent need for early diagnosis systems in order to promote timely and optimal management. The current coronavirus disease 2019 (COVID-19) pandemic accelerates this need; people living with dementia are at an increased risk of infection due to an inability to comprehend, recall and follow hygiene and procedures [3].

Declines in speech and language are regarded as key early markers of AD [4]. However, sparse and heterogeneous data sets are limiting the impact of research in this area. The *Alzheimer's Dementia Recognition through Spontaneous Speech* (ADReSS) challenge aims to address this issue by supplying a new AD speech corpus on which competitors perform two different recognition tasks [5]. The database consists of 54 participants with AD and 54 matched controls. The first task is the 2-class classification between the AD and non-AD samples. The second task is a regression task predicting the score of the *Mini-Mental State Examination* (MMSE) [6] of a speaker.

Herein, we present the ***T**raining Network on **A**utomatic **P**rocessing of **PA**thological **S**peech* (TAPAS) – a Horizon 2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network – approach to these two tasks. As both acoustic- and linguistics-based systems have shown promise in the identification of AD, the latter particularly so, we explore the efficacy of combining information gained from these different combinations of state-of-the-art approaches.

Based on previous works that demonstrated their suitability in related tasks, we utilise three different acoustic-based systems. The first, a *Bag-of-Audio Words* (BoAW) system [7] has been successfully applied for other speech-health recognition tasks, e. g., [8, 9]. Based on results achieved in [10], we also test a Siamese network [11]. Finally, building on [12], we include an End-to-End, raw waveform, *Convolutional Neural Network* (CNN). To the best of the authors' knowledge, this is the first time these three systems have been used in AD recognition.

We compare and combine these acoustic systems, with a linguistic system that utilises *Global Vectors* (GLoVe) word embeddings [13] and a hierarchical attention neural network [14]. The strength ofthis approach has been demonstrated across a range of natural language processing (NLP) tasks [15], including AD detection [16, 17]. Given that linguistic features have, in general, shown stronger performances in AD detection tasks [4, 18], we regard this system as our gold standard, and investigate if, (i) a combination of our acoustic systems can match performance with the linguistic systems and, (ii), if the acoustic systems provide complementary information to the linguistic system.

http://dx.doi.org/10.21437/Interspeech.2020-2635

2. Methodology

This section introduces the acoustic and linguistic systems considered in our contribution to the ADReSS challenge.

2.1. Bag-of-Audio-Words

Bag-of-Audio-Words (BoAW) [7] features have been applied for a range of speech-based recognition tasks, including cold and flu detection [8] and level of pain evaluation [9]. BoAW involves the quantisation of acoustic *low-level descriptors* (LLDs), where each frame-level LLD vector is assigned to an audio word from a previously learnt codebook. Typically, the codebook is formed form LLDs extracted from the training partition of a dataset. The subsequent quantisation, undertaken by counting the number of assignments for each audio word, generates a sparse histogram representation of a given speech file. The openXBOW [7] is an open-sourced toolbox for the formation of BoAW features, it has been widely utilised such as in INTERSPEECH Computational Paralinguistics Challenges (COMPARE) [19].

In the formation of BoAW features, LLD vectors are first extracted from the speech files. In this work, three LLDs feature representations are generated using openSMILE [20]: Mel-Frequency Cepstral Coefficient (MFCC), log-Mel, and the COMPARE acoustic feature set [21]. These three feature representations have previously been shown to be suitable for AD recogniton [4, 5, 22] and their bagged representations have performed well in previous studies, especially in health-based tasks [23]. Therefore, the three BoAW representations have promise as effective representations of AD recognition.

2.2. Siamese Network

Inspired by the success of Siamese networks in related tasks [10, 24–26], we investigate this paradigm for the task of AD recognition. A core advantage of Siamese networks is the associated *contrastive loss function* that encourages intra-class compactness and inter-class separability [27]. During training, information from segments of recordings belonging to the same condition (AD speech or healthy speech) is pulled together using contrastive loss, while information relating to segments of recordings from different conditions are pushed away from each other.

Formally we define the contrastive loss L_c as:

$$L_c(\boldsymbol{x}_1, \boldsymbol{x}_2, y) = (1-y) \cdot D(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 + y \cdot \max(0, \alpha_c - D(\boldsymbol{x}_1, \boldsymbol{x}_2))^2, \quad (1)$$

where $y = 1$ if the embeddings $\boldsymbol{x}_1$, and $\boldsymbol{x}_2$ are from different conditions and should thus be distant, and $y = 0$ when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are from the same condition and thus should be close. Additionally, D denotes the Euclidean distance and α_c is the margin which we want to obtain between the two different conditions.

2.3. End-to-End Convolutional Neural Network

We also propose modelling AD's speech in an end-to-end manner, utilising raw waveform based CNNs. This framework was been successfully applied to tasks such as emotion recognition [28], speaker verification [29], gender identification [30], or depression detection [12]. Exploiting this paradigm, we can capture information related to different speech production mechanisms by modifying the initial kernel width (kW) parameter [29, 31]. Setting $kW = 300$ covers a signal length of approximately 20 ms (segmental) allows the first convolution layer to model voice-source-related information. Alternatively, by setting $kW = 30$ covers a signal of approximately 2 ms of length (sub-segmental), encouraging the first convolution layer into tending to capture vocal tract information, such as formants.

In order to verify the importance of changes in fundamental frequency, we also investigated using zero-frequency filtered (ZFF) signals [32]. Taking inspiration from a recent paper showing that voice source related information related to depression can be modelled with CNNs [12], the filtered signals are fed to the same network applied to classification and regression tasks.

2.4. Hierarchical Attention Network

We implement a *bi-directional Hierarchical Attention Network* (bi-HANN) as our linguistic system. This choice was motivated by the success of bi-HANNs in other AD recognition tasks [16, 17]. This approach is a two-stage system which operates at the word- and sentence-level [14]. In our model, w_{it} with $t \in [1, T]$ and $i \in [1, L]$ is used to represent the tth word in ith sentence. Each word w_{ij} is encoded into a fixed dimensional vector x_{ij} by a pre-trained embedding matrix W_e. The choice of word embedding matrix is a trainable parameter in the model.

To extract word-level characteristic patterns from the variable-length sequence, a *bidirectional long short-term memory* (bi-LSTM) is applied on the word vectors, followed by an attention mechanism. After obtaining the sentence representation s_i, a further bi-LSTM layer extracts sentence-level information extraction. Given a sentence vector s_i, this action generates a transcript representation v with a similar structure as for the word level model. Finally, a dense layer with a *sigmoid* function is applied for classification on the transcript representation. See [17] for further information on this paradigm.

3. Experimental Setup

This section introduces the ADReSS AD dataset, as well as the key outlines the key experimental settings associated with our four AD recognition systems.

3.1. Database

The speech data and transcripts used in this work were provided by the ADReSS challenge organisers [5]. The speech data contains both full speech files and segmented speech chunks. The segmented chunks, used to set the challenge baseline [5] were generated by the organisers applying a log-energy threshold-based voice activity detector. The BoAW and End-to-End systems utilised these chunks, while the Siamese network exploits the full recordings. The transcripts contain the linguistic content of an interviewer and a participant, as well as other related annotations. We, therefore, pre-processed all the raw transcripts to remove all content unrelated to the spoken content of the participant and used the remaining information as input to the bi-HANN. For the sake of brevity, the demographics and characteristics of the data set are not given here. The interested reader is referred to [5] for these details.

3.2. Bag-of-Audio-Words

The extraction of the three LLDs representations mentioned in 2.1 is described below. Both MFCC and log-Mel LLDs are extracted with a frame size of 0.025 s and a step size of 0.01 s. The MFCC LLDs consist of MFCC 1-14 and the corresponding delta regression coefficients, leading to 28-dimensional MFCC LLDs. The log-Mel LLD feature set contains 64-band log-Mel frequencies and corresponding 64 delta regression coefficients. The 130 dimensional COMPARE LLDs [21] were obtained by the OPENSMILE configure file *ComParE_2016.conf*.

Next the LLD's were quantised to form the BoAW representations. The input LLDs are split into two subsets, in order to train a separate codebook in each subset. We then quantise 14 LLDs for MFCC, 64 for log-Mel, and 65 for ComParE features for both subsets. The number of word-assignments was set as 10 for all three feature sets. Then, the optimal codebook size was searched in $\{65, 125, 250, 500\}$. Finally, the extracted BoAW features were then fed into a linear Support Vector Machine (SVM) for classification or regression. The combination of BoAW and SVM has worked well in similar tasks [8, 9]. The complexity hyperparameter in the SVM is optimised from the setting of $\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1\}$.

3.3. Siamese Networks

This model generates embeddings using a deep Siamese neural network consisting of convolutional layers. The network was trained using a contrastive loss between the two different conditions (Section 2.2). Note, as the Siamese network and contrastive loss function are not suited to regression analysis, we only present use of this system in the classification task. As an input, the model used either 8-second or 16-second segments, with a 2 second stride size, extracted from the full, rather than the chunked audio recordings (Section 3.1). Log-Mel spectrograms were then extracted from these segments using a frame size of 25 ms and stride of 10 ms. Our deep Siamese network consists of two CNNs to extract embeddings, one for each class. The encoded embeddings are then concatenated and fed into a fully connected network to estimate their similarity. Specifically, each CNN has 4 convolutional layers, each of which is followed by rectified linear unit (ReLU) activations, and batch normalisation. After the embeddings from the two CNNs are extracted, they are concatenated and fed into a 2-layer Fully Connected Network with each layer followed by ReLU activation. The final layer uses a sigmoid activation function to squash the output value between 0 and 1, which is regarded as the similarity value.

3.4. End-to-End Convolutional Neural Network

Raw waveform CNN networks typically consist of an initial filter stage followed by a classification stage. Our proposed network has four convolution layers, kernel widths *30-10-4-3* for subsegmental modelling, and *300-5-4-3* for segmental modelling (Section 2.3). Convolution layers are followed by maximum pooling and ReLU activations. The final stage of the network is a multilayer perceptron. At the output, the classification network predicts a probability for AD using a sigmoid function, while the output is a linear value for the regression model. In both cases, this is a per frame action. These frame-level values are then averaged to get per-utterance posterior probabilities.

The input to the CNNs, w_{seq}, is a 250 ms length speech segment, shifted by 10 ms. We randomly-initialised CNNs with a batch-size of 256 and employ a cross-entropy cost function or mean squared error for the two tasks, respectively. We opted for a decaying learning schedule which halves the learning rate between 10^{-3} and 10^{-7} whenever the validation loss stops reducing. In initial testing, we observed that a combination of *ZFF* with subsegmental modelling was better suited to the classification task. In contrast, the combination of *ZFF* with segmental modelling was better suited to the regression task. Herein, *ZFF* denotes this combination.

3.5. Hierarchical Attention Network

Only the transcripts that corresponded to participants are used for the bi-HANN model (Section 3.1). GloVe 100-dimensional word vectors trained on Wikipedia 2014 and Gigaword-5 data were taken as our pre-trained embedding matrix [13]. The bi-HANN was trained on a fixed number of epochs (20) and evaluated on the development set at each epoch. Batch size was set to 20 and the best model selected via the F_1-score on the training set. The number of LSTM units was set to 100, and the dense layer dimension in word-level was set as 50. For the attention layer's dimension, both the sentence and word level is set to 30. The sentence length was set to 30, and we zero-padded the shorter sentences. The sentence numbers in a transcript were set to 30, with zero-padding used on the shorter transcript. We opted for *Adam* optimisation with a learning rate of $1e^{-5}$. Dropout was applied after all the functional layers with 0.3 dropout rate.

We compare the *bi-HANN* with two simplified linguistics systems, a *bi-LSTM* and *bi-LSTM with attention* (bi-LSTM-Att). These models follow the same parameters setting as in the *biHANN*. The maximum word number for each transcript is 200, with zero-padding being applied if the word number is less than this amount. Dropout layers are adopted after the LSTM layer and attention, and dense layers.

Table 1: *A comparison of the proposed approaches on the ADReSS Challenge training set. Results are the average performance across a nine-fold cross-validation step up.*

Approach			Acc.	F1	RMSE
BoAW	MFCC	65	.611	.604	7.03
		125	**.630**	**.623**	7.05
		250	.602	.593	7.00
		500	.620	.610	7.17
	LogMel	65	.565	.540	7.18
		125	.556	.526	**6.97**
		250	.537	.522	7.15
		500	.556	.544	7.03
	ComParE	65	.620	.601	7.04
		125	.593	.582	7.04
		250	.574	.556	7.17
		500	.574	.567	7.13
	Fusion	–	**.639**	.639	6.99
SiameseNet	LogMel	8 s	.586	.693	–
		16 s	**.628**	**.731**	–
End-to-End	Raw seg		.713	.762	8.89
	ZFF		**.741**	**.780**	7.58
Linguistic	bi-LSTM		.694	.736	5.99
	bi-LSTM-Att		**.842**	**.842**	5.49
	bi-HANN		.827	.826	**4.86**
Fusion	Maj. / Wt.		.850	.855	4.91
Fusion	bi-LSTM-Att		.887	.887	7.73
Fusion	bi-HANN		.831	.829	7.64

3.6. Evaluation Metrics

As per the challenge organisers [5], we report our results in terms of accuracy and F_1-score for the classification score, and *root mean squared error* (RMSE) for the MMSE prediction task. We divided the 108 speaekers in the training set into 9 folds of 12 speakers[1] and report the average of each score across each fold in the results section.

[1]Partitioning of folds available on request

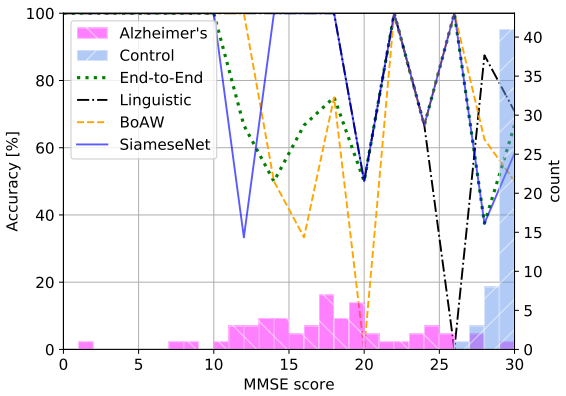


Figure 1: *Accuracy per MMSE score of our for best systems on the development set, together with a histogram of MMSE scores.*

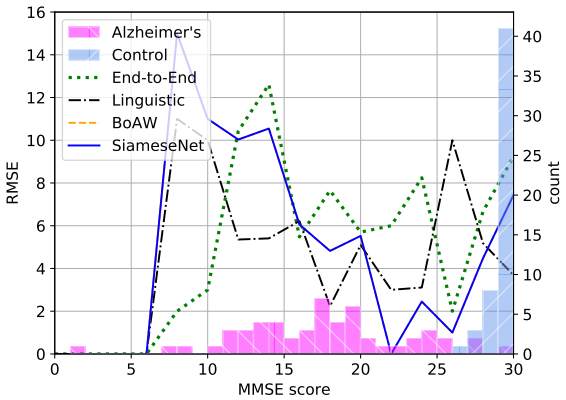


Figure 2: *RMSE per MMSE score of our for best systems on the development set, together with a histogram of MMSE scores.*

4. Results and Discussion

4.1. Training Set Results

As expected, the linguistic systems outperforms the acoustic systems (Table 1). The *bi-HANN* system achieves the strongest result on the regression task; however, the simpler *bi-LSTM-Att* system achieves the strongest performances on the classification task. This result does not match with similar systems in the literature [17]. We speculate the more even performances between the *bi-HANN* and *bi-LSTM-Att* systems are due to the smaller size of the ADReSS database. The end-to-end CNNs produce the strongest performance of the acoustic systems on the classification task, highlighting the benefits of self-learning features. The inclusion of the *ZFF* signals improves the performance of this set-up, indicating the importance of fundamental frequency in AD recognition tasks. Finally, the *BoAW-logMel-C125* gains the lowest RMSE of our acoustic systems; verifying the strength of this feature representation in paralinguistic tasks [23].

Figure 1 and Figure 2 show accuracy and RMSE per MMSE, respectively, for the best performing systems from each group on the training set. In terms of accuracy, none of the systems in Figure 1 show any consistency. Whereas, in terms of RMSE, we observe high errors at low MMSE values and another peak around 26, where control and AD histograms start overlapping.

The late fusion between the best-performing systems from each grouping did not improve system performance beyond the linguistic only approaches (Table 1). This approach adopted a majority voting for the classification task or a weighted average for the regression task. However, in the classification task, when fusing the *bi-LSTM-Att* and *ZFF* systems, we were able to improve on the performance of the bi-LSTM-Att system. This approach exploited the learnt representations from the second to last layer of the *ZFF* CNN. These features were concatenated with the attention output of the bi-LSTM attention layer and the network trained as per (Section 2.4). However, this feature fusion approach was not as well suited to the regression task.

Table 2: *A comparison of the best performing approaches from Table 1 on the ADReSS Challenge test set*

Approach	**Acc.**	**F1**	**RMSE**
Baseline [5]	.625	.620	6.14
BoAW	.563	.561	6.88
BoAW fusion (3-best)	.625	.625	6.45
SiameseNet	**.708**	**.708**	–
End-to-End	.667	.664	6.75
bi-LSTM-Att	.813	.812	4.66
bi-HANN	.729	.726	4.74
Fusion Feat. (bi-LSTM-Att)	.771	.766	5.62
Fusion Feat. (bi-HANN)	.813	.810	6.65
Fusion Maj./ W-avg (3-best)	**.852**	**.854**	**4.65**

4.2. Test Set Results

The SiameseNet performs the strongest out of the acoustic systems in the classification task (Table 2). Interestingly, despite their stronger performance in the classification task, none of the acoustic systems trailed on the test set out-performs the regression baseline. The *bi-LSTM-Att* system was our strongest standalone system, highlighting the strength of considering linguistics in AD recognition tasks. The benefits of fusion are more apparent in the test set, with our best result being achieved through a majority vote (classification) / weighted average (fusion) of the *BoAW-MFCC-C125* (classification) / *BoAW-logMel-C125* (regression), *ZFF*, and *bi-LSTM-Att* systems. This set-up achieves an accuracy of .852 and an RMSE of 4.65.

5. Conclusions

This paper described the TAPAS Training Network approach to the INTERSPEECH 2020 ADReSS challenge. We compared and combined information from four different speech-based Alzheimer's recognition approaches; three acoustic and one linguistic. The linguistic systems outperformed our acoustics approaches; such a result is unsurprising given a human observer generated the transcripts. Thus, they contain considerably fewer sources of noise than the audio recordings. Small gains were found when fusing acoustics and linguistics approaches. In future work, we will explore the effect of performing similar analysis when combining acoustic information with linguistics systems based on transcripts generated from an automatic speech recognition system.

6. Acknowledgements

This work was supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS). The four early stage researchers from the project: Yilin Pan, Zhao Ren, Julian Fritsch, and Srikanth Nallanthighal, all contributed equally to this manuscript.

7. References

[1] World Health Organization, "Towards a Dementia Plan: A WHO Guide," WHO, 2018, 82 pages. [Online]. Available: https://www.who.int/mental_health/neurology/dementia/guidelines_risk_reduction/en/

[2] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 13, no. 4, pp. 325–373, 2017.

[3] H. Wang, T. Li, P. Barbarino, S. Gauthier, H. Brodaty, J. L. Molinuevo, H. Xie, Y. Sun, E. Yu, Y. Tang, and X. Yu, "Dementia care during COVID-19," *The Lancet*, vol. 395, no. 10231, pp. 1190–1191, 2020.

[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] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *Proc. INTERSPEECH 2020*, Shanghai, China, 2020, 5 pages. [Online]. Available: https://arxiv.org/abs/2004.06833

[6] T. N. Tombaugh and N. J. McIntyre, "The Mini-Mental State Examination: A Comprehensive Review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[7] M. Schmitt and B. Schuller, "openXBOW — Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[8] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, and B. Schuller, "You sound ill, take the day off: Classification of speech affected by Upper Respiratory Tract Infection," in *Proc. EMBC 2017*. Jeju Island, South Korea: IEEE, 2017, pp. 3806–809.

[9] Z. Ren, N. Cummins, J. Han, S. Schnieder, J. Krajewski, and B. Schuller, "Evaluation of the pain level from speech: Introducing a novel pain database and benchmarks," in *Proc. ITG 2018*. Oldenburg, Germany: IEEE/VDE, 2018, pp. 56–60.

[10] S. Boelders, V. S. Nallanthighal, V. Menkovski, and A. Härmä, "Detection of Mild Dyspnea from Pairs of Speech Recordings," in *Proc. ICASSP 2020*. Barcelona, Spain: IEEE, 2020, pp. 4102–4106.

[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification using a "Siamese" Time Delay Neural Network," in *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan-Kaufmann, 1994, pp. 737–744.

[12] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning Voice Source Related Information for Depression Detection," in *Proc. ICASSP 2019*. Brighton, United Kingdom: IEEE, 2019, pp. 6525–6529.

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP 2014*. Doha, Qatar: ACL, 2014, pp. 1532–1543.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proc. NAACL-HLT 2016*, San Diego, California, USA, 2016, pp. 1480–1489.

[15] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 221–225.

[16] J. Chen, J. Zhu, and J. Ye, "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 4085–4089.

[17] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic Hierarchical Attention Neural Network for Detecting AD," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 4105–4109.

[18] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An Avatar-Based System for Identifying Individuals Likely to Develop Dementia," in *Proc. INTERSPEECH 2017*. Stockholm, Sweden: ISCA, 2017, pp. 3147–3151.

[19] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge:: Atypical & Self-Assessed Affect, Crying & Heart Beats," in *Proc. INTERSPEECH 2018*. Hyderbad, India: ISCA, 2018, pp. 122–126.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM MM '10*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[21] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM '13*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[22] 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.

[23] N. Cummins, A. Baird, and B. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and Opportunities," *Methods*, vol. 151, pp. 41–54, 2018.

[24] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech Emotion Recognition via Contrastive Loss under Siamese Networks," in *Proc. ASMMC-MMAC '18*. Seoul, Republic of Korea: ACM, 2018, pp. 21–26.

[25] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 3885–3889.

[26] M. Harandi, S. R. Kumar, and R. Nock, "Siamese Networks: A Thing or Two to Know," Data61, CSIRO, 2017. [Online]. Available: https://pdfs.semanticscholar.org/c03c/e09b419b6a15c1228e344a900d8c54bddc78.pdf

[27] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *ProcĊVPR '05*. San Diego, CA, USA: IEEE, 2005, pp. 539–546.

[28] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network," in *Proc. ICASSP 2016*. Shanghai, China: IEEE, 2016, pp. 5200–5204.

[29] H. Muckenhirn, M. Magimai-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proc. ICASSP 2018*. Calgary, AB, Canada: IEEE, 2018, pp. 4884–4888.

[30] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On Learning to Identify Genders from Raw Speech Signal Using CNNs," in *Proc. INTERSPEECH 2018*. Hyderbad, India: ISCA, 2018, pp. 287–291.

[31] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," *Speech Communication*, vol. 108, pp. 15–32, Apr. 2019.

[32] K. S. R. Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's Dementia recognition from spontaneous speech

Morteza Rohanian[1], *Julian Hough*[1], *Matthew Purver*[1,2]

[1]Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
[2]Department of Knowledge Technologies, Jožef Stefan Institute
{m.rohanian, j.hough, m.purver}@qmul.ac.uk

Abstract

This paper is a submission to the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, which aims to develop methods that can assist in the automated prediction of severity of Alzheimer's Disease from speech data. We focus on acoustic and natural language features for cognitive impairment detection in spontaneous speech in the context of Alzheimer's Disease Diagnosis and the mini-mental state examination (MMSE) score prediction. We proposed a model that obtains unimodal decisions from different LSTMs, one for each modality of text and audio, and then combines them using a gating mechanism for the final prediction. We focused on sequential modelling of text and audio and investigated whether the disfluencies present in individuals' speech relate to the extent of their cognitive impairment. Our results show that the proposed classification and regression schemes obtain very promising results on both development and test sets. This suggests Alzheimer's Disease can be detected successfully with sequence modeling of the speech data of medical sessions.

Index Terms: Cognitive Decline Detection, Affective Computing, Computational Paralinguistics

1. Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative condition and the most common form of dementia. AD gradually affects the memory, language and cognitive skills and ultimately the ability to perform basic tasks in the everyday lives of patients. Early diagnosis of AD has become essential in disease management as it has not been possible to reverse the degenerative process, even with significant efforts focused on therapies [1].

Discrepancies in speech comprehension, speech production and memory functions are closely tied in with AD as suggested by a decrease in global vocabulary and a loss in evocative memory [2]. Patients with AD have difficulty performing tasks that leverage semantic information; they exhibit problems with verbal fluency and identification of objects [3]. The semantics and pragmatics of their language appear affected throughout the entire span of the disease more than syntax [4]. AD Patients talk more gradually with longer pauses and invest extra time seeking the right word, which contributes to disfluency of speech [3].

AD diagnosis demands the existence of cognitive dysfunction to be validated by neuropsychological assessments like the mini mental state examination (MMSE) performed in medical clinics [5]. Diagnosis is typically based on the clinical analysis of patients' history and the presence of typical neurological and neuropsychological features. It is costly and not accessible to all patients who have concerns about their memory functions.

Recent experimental research has looked at AD's automated analysis from multimodal data as alternative, less invasive tools for diagnostics. Studying behaviours of individuals could also help detect AD earlier. There has been research on building systems which use a broad range of multimodal features to identify AD severity. A meaningful association between MMSE scores and language measures such as articulation and disfluency has been found [6].

Much of the work to date has looked separately at the properties of the language of an individual: acoustic and lexical characteristics of speech, or syntax, fluency, and content of information. Usually these are studied within language tasks in specific domains or in conversational dialogue [7]. Several studies have suggested various forms of speech analysis to identify AD. Researchers found that the number of pauses, pause proportion, phonation time, phonation–to-time ratio, speach rate, articulation rate, and noise-to-harmonic ratio correlate with the severity of AD [8]. Weiner et al. [9] developed a Linear Discriminant Analysis (LDA) classifier with a set of acoustic features such as the mean of silent segments, speech and silence durations and silence to speech ratio to distinguish subjects with AD from the control group and achieved a classification accuracy of 85.7 percent. Ambrosini et al. [10] showed an accuracy of 73 percent when using selected acoustic features (pitch, voice breaks, shimmer, speech rate, syllable duration) to detect mild cognitive impairment from a spontaneous speech task.

In terms of the features which aid AD detection, lexical features from spontaneous speech are shown to be informative. Jarrold et al. [11] extracted the frequency occurrence of 14 different part of speech features and combined them with acoustic features. Abel et al. [12] modeled patient speech errors (naming and repetition disorders) to the problem of AD diagnosis.

There has also been work on modelling multimodal input for AD detection. Gosztolya et al. [13] examined the fusion of two SVM models with separate feature sets. The first model used a set of acoustic features, and the second model was developed using linguistic features extracted from manually annotated transcripts. Their work showed the complementary information that audio and lexical features may contain about a subject with AD.

Among other similar tasks, using multimodal fusion to predict a cognitive state, research has been done on integrating temporal information from two or more modalities in a recurrent approaches to classify emotions or detecting different mental states, such as depression [14]. One key challenge these mod-

http://dx.doi.org/10.21437/Interspeech.2020-2721

els have is addressing the various predictive capacity of each modality and their different levels of noise. The application of a gating mechanism in various multimodal tasks has been shown to be successful in controlling the level of contribution of each modality to the eventual prediction.

This paper addresses AD classification and MMSE score regression tasks, which are part of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [15]. In ADReSS, participants are required to assess the AD severity of different subjects, where the target severity is based on their MMSE scores.

We performed a binary classification of samples of speech into AD and non-AD classes and create regression models to predict MMSE scores. Using the ADReSS Challenge data which consists of speech recordings and transcripts of spoken picture descriptions, we explored various features as diagnostically relevant tools. We focused in particular on sequential modelling of sessions and whether the disfluencies and self-repairs present in individuals' speech can help predict the level of cognitive impairment.

Our approach is motivated by [14] that developed the ability to learn difficult decision boundaries which other models with different methods of fusion have trouble managing, and maximise the use and combination of each modality. We employed data of individuals under controlled conditions, and modeled the sessions with audio and text features in a Long-Short Term Memory (LSTM) neural network to detect AD. Our findings indicate that AD can be detected with minimal information available on the structure of the description tasks by pure sequential modelling of a session. We also found that disfluency markers have predictive power for AD recognition.

2. Proposed Approach

Our approach is to model the speech of individuals giving picture descriptions as a sequence to predict whether they have AD or not, and if so, to what degree. To predict AD, we performed three sets of experiments using features from the audio and text data:

1 LSTM models utilising unimodal audio and text features.

2 LSTM model with gating to test the effect of using multimodality.

3 A multimodal LSTM model using acoustic and lexical information, including disfluency tagging.

The details of the three experiments are outlined below in the following sub-sections. In line with the standard assumption in deep learning, we take the approach that for a model to be genuinely data-driven, minimal feature engineering is required. The model's power is in its capacity to represent information through non-linear transforms, at varying spatial and temporal units, and from different modalities. Since we were interested in modelling temporal session changes, we used a bi-directional Long Short-Term Memory (LSTM) neural network as it has the added benefit of sequential data modelling. For each of the audio and text modalities we trained an LSTM model separately, using the audio and text features.

2.1. Multimodal Features

Lexical Features from Text A pre-trained GloVe model [16] was used to extract the lexical feature representations from the picture description transcript and convert the utterance sequences into word vectors. We selected the hyperparameter values, which optimised the output of the model on the training set. The optimal dimension of the embedding was found to be 100.

Audio Features A set of 79 audio features were extracted using the COVAREP acoustic analysis framework software, a package used for automatic extraction of features from speech [17]. We sampled the audio features at 100Hz and used the higher-order statistics (mean, maximum, minimum, median, standard deviation, skew, and kurtosis) of COVAREP features. The features include prosodic features (fundamental frequency and voicing), voice quality features (normalized amplitude quotient, quasi open quotient, the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum, maxima dispersion quotient, parabolic spectral parameter, spectral tilt/slope of wavelet responses, and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics) and spectral features (Mel cepstral coefficients 0-24, Harmonic Model and Phase Distortion mean 0-24 and deviations 0-12). Segments without audio data were set to zero. A standard zero-mean and variance normalization was applied to features. We omitted all features with no statistically significant univariate correlation with the results of training set.

2.2. Sequence Modeling

The potential of neural networks lies in the power to derive representations of features by non-linear input data transformations, providing greater power than traditional models. As we were interested in modelling temporal nature of speech recordings and transcripts, we used a bi-directional LSTM. For each of the audio and text modalities we trained a separate unimodal LSTM model, using different sets of features. For the input data we explored different timesteps and strides. After exploring different hyper-parameters, the model using audio data has a timestep of 20 and stride of 1 with 4 bi-directional LSTM layers with 256 hidden nodes. The model using text input has an input with a timestep of 10 and stride of 2 and has 2 LSTM layers with 16 hidden nodes. The code used in the experiments are publicly available in an online repository.[1]

2.3. Multimodal Fusion with Gating

Audio and text features can include not only discriminative and temporarily changing information about the current state of a subject, but supporting information as well.

The model consists of two branches of the LSTM, one for each of the modalities, with their outputs combined into final feed-forward highway layers. The branches are made up of different hyperparameters and configured with respect to each modality's properties. Their outputs are concatenated and passed through N highway layers (where the best value N was determined from optimizing on heldout data). We pad the size of the training examples in the text set (which was the smaller set) to meet the audio set by mapping together instances that occurred in the same session, as the audio and text inputs for each branch of the LSTM had different timesteps and strides.

Gating Mechanism Data from two modalities affect the final output differently, and it is important to consider the amount of noise when aggregating them into a single representation. Since learned representation for the text can be undermined by corresponding audio representation, during multimodal fusion we need to minimise the effects of noise and overlaps. We use feed-forward highway layers [18], with gating units that learn by weighing text and audio inputs at each time step to regulate

[1] https://github.com/mortezaro/ad-recognition-from-speech

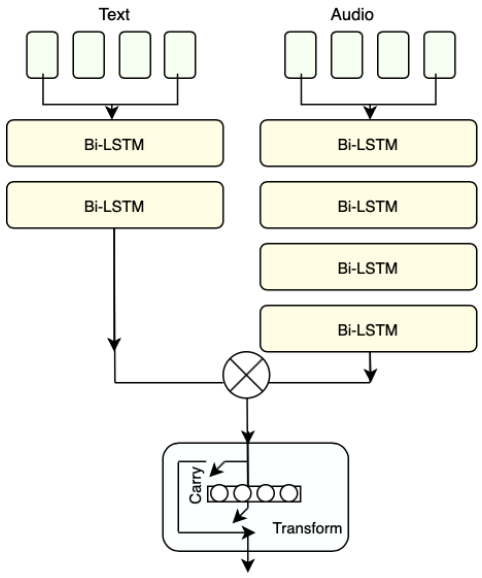


Figure 1: *Multimodal fusion with gating.*

information flow through network work.

Each highway layer consists of two non-linear transformations: a Carry (Cr) and a Transform (Tr) gate which determine the degree to which the output is generated by transforming and carrying the input. Each layer uses the gates and feed-forward layer H to regulate its input vector at timestep t, D_t, to generate output y:

$$y = Tr \cdot H + Cr \cdot D_t \tag{1}$$

where Cr is simply defined as $1 - Tr$, giving:

$$y = Tr \cdot H + (1 - Tr) \cdot D_t \tag{2}$$

The transform gate Tr is defined as $\sigma(W_{Tr}D_t + b_{Tr})$, where W_{Tr} is the weight matrix and b_{Tr} the bias vector for the gates. Based on the transform gates outputs, highway layers adjusts their performance from multiple-unit layers to layers that only pass through their inputs. As inspired by [18] and to help resolve long-term learning dependencies faster we initialise b_{Tr} with a negative value (biased towards the Carry gate). We use a block of 3 stacked highway layers. The overall architecture of the LSTM with Gating model is shown in Figure 1.

2.4. Multi-modal Model with Disfluency Markers

Disfluencies like self-repairs, pauses and fillers are widespread in everyday speech [19]. Disfluencies are usually seen as indicative of communication problems, caused by production or self-monitoring issues [20]. Individuals with AD are likely to deal with troubles in language and cognitive skills. Patients with AD speak more slowly and with longer breaks, and invest extra time seeking the right word, which in effect contributes to disfluency [3]. The present research explores the disfluencies present in the speech of AD patients as they contribute to severity of symptoms.

Self-repair disfluencies are typically assumed to have a reparandum-interregnum-repair structure, in their fullest form as speech repairs [21]. A reparandum is a speech error subsequently fixed by the speaker; the corrected expression is a repair. An interregnum word is a filler or a reference expression between the words of repair and reparandum, often a halting step as the speaker produces the repair, giving the structure as in (3)

$$\texttt{John} \ \underbrace{\texttt{[likes}}_{\texttt{reparandum}} + \underbrace{\texttt{\{ uh \}}}_{\texttt{interregnum}} \ \underbrace{\texttt{loves]}}_{\texttt{repair}} \ \texttt{Mary} \tag{3}$$

In the absence of reparandum and repair, the disfluency reduces to an isolated *edit term*. A marked, lexicalized edit term such as a filled pause ("uh" or "um") or more phrasal terms like "I mean" and "you know" can occur. Recognizing these elements and their structure is then the task of disfluency detection.

We automatically annotated self-repairs using a deep-learning-driven model of incremental detection of disfluency developed by Hough and Schlangen [22, 23]. It consists of deep learning sequence models that use word embeddings of incoming words, part-of-speech annotations, and other features in a left-to-right, word-by-word manner to predict disfluency tags. Here each word is either tagged as a repair onset tag (marking first word of the repair phase) edit term, or fluent word by the disfluency detector- we concatenate the disfluency tags with the word vectors to create the input for text-based LSTM.

3. Experiments

3.1. Data

The ADReSS challenge's data consists of speech recordings and transcripts of spoken picture descriptions gathered from participants via the Boston Diagnostic Aphasia Exam's Cookie Theft picture [15]. The training set includes 108 subjects, and the state of the subjects is assessed on the basis of the MMSE score. MMSE is a commonly used cognitive function test for older people. It involves orientation, memory, language, and visual-spatial skills tests. Scores of 25-30 out of 30 are considered as normal, 21-24 as mild, 10-20 as moderate and $<$10 as severe impairment.

The total number of speech segments each participant had generated was 24.86 on average. The annotations for the test set were not included in the public release of the ADReSS Challenge, so all models were tested on both the development and test set. The data is pre-processed acoustically and is balanced in terms of age and gender.

3.2. Implementation and Metrics

We set up our model to learn the most useful information from modalities for predicting AD. All experiments are carried out without being conditioned on the identity of the speaker. The sizes of layers and the learning rates are calculated by grid search on validation test. The LSTM models were trained using ADAM [24] with a learning rate of 0.0001. For the loss function we used Binary Cross-Entropy to model binary outcomes, and Mean Square Error (MSE) to model regression outcomes. For binary classification of AD and non-AD, we report accuracy, precision, recall, and F1 scores and for the MMSE prediction task we report the Root Mean Square Error (RMSE).

3.3. Baseline Models

We compare the performance of our models to the ADReSS Challenge baseline [15] with an ensemble of audio features which was provided with the dataset. The baseline classification experiments were different methods of linear discriminant

analysis (LDA), decision trees (DT), and support vector machines (SVM). The baseline regression experiments were different methods of DT, gaussian process regression (GPR), and SVM.

Table 1: *Result of the AD classification and regression experiments with our models in cross validation*

Models	Features	Accuracy	RMSE
LSTM	Acoustic	0.64	6.01
LSTM	Lexical	0.69	5.42
LSTM	Lexical+ Dis	0.73	5.08
LSTM with Gating	Acoustic + Lexical	0.76	5.01
LSTM with Gating	Acoustic + Lexical + Dis	**0.77**	**4.98**

Table 2: *Result of the AD classification and regression experiments with our models against baseline models on test set*

Models	Features	Accuracy	RMSE
Baseline ([15])			
LDA	Acoustic	0.625	-
DT	Acoustic	0.625	6.14
SVM	Acoustic	0.563	6.12
GPR	Acoustic	-	6.33
Our Models			
LSTM	Acoustic	0.666	5.93
LSTM	Lexical	0.708	5.45
LSTM	Lexical + Dis	0.729	4.88
LSTM with Gating	Acoustic + Lexical	0.771	4.57
LSTM with Gating	Acoustic + Lexical + Dis	**0.792**	**4.54**

4. Results

In Table 1, we present our proposed model's performance in a cross-validation setting and in Table 2 against that of baselines models on AD detection on the provided test set. For AD detection, our proposed LSTM model with gating and disfluency features achieves an accuracy of **0.792** and RMSE of **4.54**, outperforming all the baselines. The overall findings confirm our assumption that a model with a gating structure can more efficiently minimise the errors and noise of the individual modalities.

Effect of disfluency features We found that disfluency tags help as features in AD detection. Adding disfluency features to the lexical features lead to improvement in both unimodal (ACC 0.70 vs. 0.72; RMSE 5.45 vs. 4.88) and multimodal models (ACC 0.77 vs. 0.79; RMSE 4.57 vs. 4.54).

Effect of multimodality The multimodal LSTM with gating model outperforms the single modality AD detection models in both the classification and regression tasks. A performance increase is obtained by combining textual and audio modalities with gating over single modality models (ACC 0.72 vs. 0.79; RMSE 4.88 vs. 4.54). Adding audio features improves performance despite having different steps and timesteps inputs for each LSTM branch. In terms of our competitor baselines (without the information from the manual transcripts), multimodal classifiers performed better than all the baseline models, indicating the potential benefits of multimodal fusion in AD detection. We found that while the baseline audio-based models have some discriminative capacity, sequence modelling is more accurate (ACC scores 0.67 vs. 0.63) and has lower (better) RMSE (5.93 vs. 6.12) for predicting AD.

For AD classification, the text features alone are more informative than the audio features, as using only the text modality gives a better AD prediction than utilizing unimodal audio modality sequentially (Acc scores 0.73 vs. 0.67; RMSE 4.88 vs. 5.93).

We can see that all models provide more accurate results on the test set than in cross validation. LSTM with gating models accuracy improved more than other models on the test set (RMSE 4.54 and 4.57 vs. 4.98 and 5.01).

Error analysis The results in Table 3 show that the LSTM model with gating and disfluency features obtains the highest precision and recall for both AD and non-AD classes. The model achieves F1 scores of 0.7826 for AD and 0.8000 for non-AD. The addition of gating particularly improves the recall of AD class: the LSTM model with lexical and disfluency features without gating has a recall 0.6667 for the AD class compared to the 0.7500 achieved with gating, while its 0.7910 recall for the non-AD class is not as far beneath the 0.8333 achieved by the full gating model. Depending on the application the model is used for, false negatives or false positives for AD detection will be more or less desirable, but as it stands our full gating model considerably reduces the false negatives of diagnosis whilst still marginally reducing the false positives.

Table 3: *Results of AD classification task on test set*

Models	Class	Precision	Recall	F1 Score	Accuracy
LSTM	AD	0.7619	0.6667	0.7111	0.7292
(Lexical+ Dis)	non-AD	0.7037	0.7910	0.7451	
LSTM with Gating	AD	0.7826	0.7500	0.7660	0.7708
(Acoustic + Lexical)	non-AD	0.7600	0.7917	0.7755	
LSTM with Gating	AD	0.8182	0.7500	0.7826	0.7917
(Acoustic + Lexical+ Dis)	non-AD	0.7692	0.8333	0.8000	

5. Conclusions

We have presented a deep multi-modal fusion model that learns the AD indicators from audio and text modalities as well as disfluency features. We trained and tested the model on audio and transcript data from individuals doing a description task under controlled conditions, and modeled the sessions with an LSTM and feed-forward highway layers as gating mechanism for AD detection. Our findings indicate that AD can be identified by pure sequential modelling of a session, with limited information available on the structure of the description tasks. We also found that markers of disfluency hold predictive power for identification of AD.

In future work we intend to study a series of language markers associated with AD severity, as well as interactions between them. In particular, we want to undertake a more principled approach to lexical markers, disfluency markers in terms of a study of self-repair and structural markers with a look at grammatical fluency. Furthermore, we want to find acoustic features that contribute more to the prediction of AD and have higher correlation with linguistic information.

6. Acknowledgments

Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

7. References

[1] A. Burns and S. Iliffe, "Alzheimer's disease," *B M J*, vol. 338, no. 7692, pp. 467–471, 2 2009.

[2] D. Kempler, *Neurocognitive disorders in aging*. Sage, 2005.

[3] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[4] K. A. Bayles and D. R. Boone, "The potential of language tasks for identifying senile dementia," *Journal of Speech and Hearing Disorders*, vol. 47, no. 2, pp. 210–217, 1982.

[5] 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.

[6] M. F. Weiner, K. E. Neubecker, M. E. Bret, and L. S. Hynan, "Language in alzheimer's disease," *The Journal of clinical psychiatry*, vol. 69, no. 8, p. 1223, 2008.

[7] S. Nasreen, M. Purver, and J. Hough, "Interaction patterns in conversations with alzheimer's patients."

[8] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[9] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[10] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid *et al.*, "Automatic speech analysis to early detect functional cognitive decline in elderly population," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 212–216.

[11] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[12] S. Abel, W. Huber, and G. S. Dell, "Connectionist diagnosis of lexical disorders in aphasia," *Aphasiology*, vol. 23, no. 11, pp. 1353–1378, 2009.

[13] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "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] M. Rohanian, J. Hough, M. Purver *et al.*, "Detecting depression with word-level multimodal fusion," *Proc. Interspeech 2019*, pp. 1443–1447, 2019.

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

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[17] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[19] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language*, vol. 53, no. 2, pp. 361–382, 1977.

[20] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[21] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, Citeseer, 1994.

[22] J. Hough and D. Schlangen, "Recurrent neural networks for incremental disfluency detection," ser. Interspeech 2015, 2015, pp. 849–853.

[23] ——, "Joint, incremental disfluency detection and utterance segmentation from speech," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 326–336.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Comparing Natural Language Processing Techniques for Alzheimer's Dementia Prediction in Spontaneous Speech

Thomas Searle[1], *Zina Ibrahim*[1], *Richard Dobson*[1,2]

[1]Department of Biostatistics and Health Informatics,
Institute of Psychiatry, Psychology and Neuroscience,
King's College London, London, U.K.
[2]Institute of Health Informatics, University College London,
London, London, U.K.
{firstname}.{lastname}@kcl.ac.uk

Abstract

Alzheimer's Dementia (AD) is an incurable, debilitating, and progressive neurodegenerative condition that affects cognitive function. Early diagnosis is important as therapeutics can delay progression and give those diagnosed vital time. Developing models that analyse spontaneous speech could eventually provide an efficient diagnostic modality for earlier diagnosis of AD. The Alzheimer's Dementia Recognition through Spontaneous Speech task offers acoustically pre-processed and balanced datasets for the classification and prediction of AD and associated phenotypes through the modelling of spontaneous speech. We exclusively analyse the supplied textual transcripts of the spontaneous speech dataset, building and comparing performance across numerous models for the classification of AD vs controls and the prediction of Mental Mini State Exam scores. We rigorously train and evaluate Support Vector Machines (SVMs), Gradient Boosting Decision Trees (GBDT), and Conditional Random Fields (CRFs) alongside deep learning Transformer based models. We find our top performing models to be a simple Term Frequency-Inverse Document Frequency (TF-IDF) vectoriser as input into a SVM model and a pre-trained Transformer based model 'DistilBERT' when used as an embedding layer into simple linear models. We demonstrate test set scores of 0.81-0.82 across classification metrics and a RMSE of 4.58.

Index Terms: adress shared task, spontaneous speech classification, alzheimers dementia classification

1. Introduction

Alzheimer's Dementia (AD) is a progressive neurodegenerative condition that largely affects cognitive function. With our globally aging population, conditions such as AD are likely to become more prevalent[1]. Despite there being no cure currently, early diagnosis can offer interventions to slow or delay progression of symptoms[2]. Prior work has used machine learning methods for the prediction of cognitive impairment (CI) conditions, including AD, using patient structured data[3] and medical imaging data[4]. Linguistic phenomenon have also been identified in those already diagnosed with AD[5, 6].

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge presents two tasks in the modelling of spontaneous speech[7]. Firstly, to classify presence of AD vs controls and secondly, to predict the 'Mental Mini State Exam' score, a common set of questions designed to assess cognitive function[8]. The challenge provides 108 training, 54 AD vs 54 Control samples, and 48 unseen test samples. Spontaneous speech audio and associated transcripts of participants describing the 'Cookie Theft' picture from the Boston Diagnostic Aphasia Exam[9] are provided. Samples are demographically and acoustically balanced and longer in duration than previous clinical studies[7]. The challenge provides an environment for researchers to test competing methods with recommendations for future work. Using machine learning techniques to predict AD from spontaneous speech could potentially offer an efficient early diagnostic modality. For example, audio samples could be collected via a mobile device with results directing individuals to seek more formal medical evaluation.

2. Data Prepossessing

In this work we exclusively focus on the textual transcriptions that are provided alongside the audio samples. Transcripts are supplied using the CHAT transcription format[10]. The transcription schema provides the linguistic content alongside some prosodic content such as: pauses, laughter, discourse markers such as 'um' and 'ah', and abbreviations such as '(be)cause'. We preprocess each transcript before feeding into our model pipelines. All code to re-create the data prepossessing, experiments and analysis is available open-source[1].

The preprocessing parses participant metadata such as age, sex, AD diagnosis and MMSE score. Each transcription line is parsed to remove time duration suffixes, specific speech artifacts such as '[','[' or '>', '<' and excess white-space such as tabs and newlines. We purposely leave discourse markers such as 'um' 'ah' and other speech artifacts such as '+...', '&=laughs' and '(...)' that indicate various pause types, or laughter in the audio.

2.1. Data Splits and Granularity

We split the transcripts into multiple competing datasets providing the candidate models with greatest opportunity to find adequate signal for the prediction tasks.

2.1.1. Transcript Level Data

Within these datasets each transcript is a single data point with their corresponding AD label and assigned MMSE score. This includes:

1. A dataset with only participant utterances concatenated together into a single paragraph as they appear in the transcript. Denoted **PAR**.

[1]https://github.com/tomolopolis/ADReSS_Challenge

http://dx.doi.org/10.21437/Interspeech.2020-2729

2. A dataset with both participant and interviewer speech concatenated into a single paragraph as they appear in the transcript. Denoted **PAR+INV**.

2.1.2. Utterance Level Data

For these datasets we define each utterance as an individual data point. This provides N=1,476, AD(N=740), controls (N=736). The target labels and MMSE scores are replicated to each utterance. Segments maintain a reference to their source transcript so random shuffling does not produce data leakage between the train and test phases. We only consider participant spoken utterances here as initial experiments indicated including interviewer speech lead to a reduction in performance. This includes:

1. A dataset with only participant utterance as individual classification & regression data points. Denoted **PAR_SPLT**.
2. Further datasets that extend the text based features with the inclusion of temporal and participant demographic features such as: time duration per sentence, time between sentences, average/max/min sentence time denoted **PAR_SPLT+T**, and participant age and sex denoted **PAR_SPLT+T+D**.

3. Methods

The baseline paper accompanying the challenge[7] only creates a single baseline result using only the text transcripts. Therefore, we present a range of models both as a new baseline result for the linguistic features, Section 3.1, alongside our more advanced approaches in Section 3.2.

3.1. Baseline Methods

We make extensive use of Scikit-Learn[11], a python based machine learning framework that provides APIs for common machine learning models, feature extraction, cross validation, hyper parameter optimisation and performance metric calculation. We use the integrated Term-Frequency-Inverse Document Frequency (TF-IDF)[12] 'bag-of-words' vectoriser. With this method text inputs forgo their sequence order and words are counted within and across documents. TF-IDF down-weights the counts of common cross-document terms, and increases weights of rare cross-document but frequent intra-document terms. This embedding method is a common first stage in any textual modelling exercise due to its efficiency and ease of use.

Scikit-learn provides APIs for optimised implementations of common machine learning algorithms such as libsvm[13] for Support Vector Machines(SVM)[14] and XGBoost[15] for Gradient Boosted Decision Trees(GBDT)[16] allowing for fast model fitting. We use both algorithms in the development of our baseline models for the transcript level and utterance level datasets presented in Section 2.1

SVMs and GBDTs are effective techniques to learn non-linear relationships between input features and the decision boundaries for both classification and regression tasks.

3.1.1. Utterance Level Methods

For the segmented speech datasets, PAR_SPLT, PAR_SPLT+T, PAR_SPLT+T+D presented in Section 2.1.2, our modelling approach does not support MMSE prediction so we only report results for AD classification. We train and cross validate TF-IDF/SVM and TF-IDF/GBDT models on each utterance, and feed output prediction probability sequences to a Conditional Random Field (CRF)[17]. CRFs are effective in the modelling of sequential data as input feature representations can depend on previous and future states of the sequence. For the overall classification of the transcript we take the final classification state of the CRF.

3.1.2. Hyper Parameter Optimisation

Table 1 lists the model configuration and associated hyper-parameter spaces we search across during an exhaustive 5-fold cross validation grid-search. As our dataset is fairly small, performing this only took a couple of minutes for each model configuration and each dataset despite the many individual model fits.

Table 1: *Baseline methods hyper-parameter searched and found optimal parameters.* * *values are* $\times 10^3$. † *the parameter spaces are sampled from an exponential probability distributions 15 times with specified* λ

Model	Hyper Parameter	Param Space	Optimal
TF-IDF/GBDT	Max Features	0.1, 0.5, 1, 2, 10$*$	1$*$
TF-IDF/GBDT	Stop Words	english, None	english
TF-IDF/GBDT	Analyser	word, char	word
TF-IDF/GBDT	sublinear TF	True, False	True
TF-IDF/GBDT	N-Estimators	100, 200, 500	100
TF-IDF/GBDT	Max Depth	3, 5, 10	5
TF-IDF/SVM	Max Features	0.1, 0.5, 1, 2, 10$*$	0.1$*$
TF-IDF/SVM	Stop Words	english, None	None
TF-IDF/SVM	Analyser	word, char	word
TF-IDF/SVM	sublinear TF	True, False	True
TF-IDF/SVM	Kernel	rbf, sigmoid	sigmoid
TF-IDF/SVM	C	0.1, 0.5, 1	1
SVM+CRF	c1	$\lambda = 0.5^\dagger$	0.0036
SVM+CRF	c2	$\lambda = 0.05^\dagger$	0.018
GBDT+CRF	c1	$\lambda = 0.5^\dagger$	0.314
GBDT+CRF	c2	$\lambda = 0.05^\dagger$	0.009

3.2. Deep Learning Methods

To converge successfully deep learning (DL) models often require more training data than methods such as SVMs and GBDTs. Training set sizes are often 50 or 100 times larger than available here. Transfer learning presents a compelling option to enable re-use of deep learning models for smaller domain specific data sets. Recently, transfer learning approaches have been successfully applied to a variety of NLP problems[18].

Large pre-trained language models are an example of transfer learning, and can be used to provide semantically rich embedding layers, allowing researchers to re-use knowledge acquired by the model from a prior training process. The language modelling task can be defined as predicting the next word given the sequence of previous words, or formally in Equation 1, modelling the probability distribution of all words w in a vocabulary V conditioned on previous words w_{i-1} to w_1.

$$P(w_i|w_{i-1}, w_{i-2} \cdots w_1) \forall w \in V \quad (1)$$

The task enables the usage of large corpora of existing texts without any explicit manual annotation, often referred to as self-supervised learning[19]. Each model we use is based upon the Transformer architecture first presented for sequence to sequence problems such as machine translation[20]. The Transformer consists of layers of encoder and decoder blocks of multi-headed self-attention followed by fully connected layers.

Each successive layer learns sophisticated latent representations of the input texts.

We use the 'transformers'[21] library to load, and re-use the BERT[22], RoBERTa[23] and DistilBERT/DistilRoBERTa[24] models as embedding layers for the PAR and PAR+INV datasets.

Running the input transcripts through the pre-trained models produces a fixed size embedding representation for each provided transcript. This is an embedding matrix of size $N \times H$, where N is the number of transcripts and H is the hidden dimension of the pre-trained model. As recommended in prior work we fit simple linear models, Logistic Regression model for AD classification and LASSO Regression for MMSE prediction, to produce our final predictions.

4. Results

Table 2 provides results for average 10 fold cross-validation for hyper parameter selection and best train/development set performance. This attempts to compare model robustness with the available training data, especially for our transcript level datasets where dataset size is limited. Metrics follow the standard definitions as outlined in the baseline work[7] and are averaged between the classes Non-AD / AD for precision, recall and F1. We then pick our 5 best performing models / dataset configurations and run on the unlabelled test dataset, containing 48 samples, sending our AD and MMSE predictions to challenge organisers. Organisers subsequently responded with aggregate results as reported in Section 4.1 for AD classification and Section 4.2 for MMSE predictions.

Table 2: *Average 10-fold CV AD Classification and MMSE prediction results. Results are highlighted if within 0.02 of the highest score.* * *indicates best score for given metric.*

Dataset	Model	Acc	Prec	Recall	F1	RMSE
PAR	GBDT	.82	.84	.82	.81	5.93
PAR	SVM	.86	**.90**	.83	**.86**	6.57
PAR	DistilBERT	**.87**	**.90**	.87	**.87**	**4.49***
PAR	DistilRoBERTa	.84	.86	.85	.82	5.12
PAR	BERT(base)	.84	.86	.85	.82	5.12
PAR	RoBERTa(base)	.75	.79	.72	.74	7.11
PAR	BERT(large)	.77	.80	.77	.76	6.64
PAR	RoBERTa(large)	.77	.81	.73	.76	7.13
PAR+INV	GBDT	.79	.80	.82	.79	5.60
PAR+INV	SVM	**.88**	**.92***	.87	**.87**	6.74
PAR+INV	DistilBERT	**.87**	.89	**.89**	**.88***	4.85
PAR+INV	DistilRoBERTa	.80	.87	.79	.78	7.11
PAR+INV	BERT(base)	.75	.76	.78	.74	7.13
PAR+INV	RoBERTa(base)	.72	.71	.71	.69	5.45
PAR+INV	BERT(large)	.75	.78	.73	.74	7.13
PAR+INV	RoBERTa(large)	.81	.88	.76	.79	6.64
PAR_SPLT	SVM+CRF	.88	.88	**.88**	**.87**	-
PAR_SPLT	GBDT+CRF	.80	.84	.74	.78	-
PAR_SPLT+T	SVM+CRF	**.89***	.87	**.90***	**.88***	-
PAR_SPLT+T	GBDT+CRF	.82	.84	.79	.81	-
PAR_SPLT+T+D	SVM+CRF	.86	.85	.87	**.86**	-
PAR_SPLT+T+D	GBDT+CRF	.83	.86	.79	.81	-

4.1. AD Classification

Table 3 shows our test set results for each metric. We show results for metrics both labels (AD vs No AD) for precision, recall and F1 metrics as defined in baseline work[7].

Table 3: *Test set results for AD classification*

Dataset / Model	Class	Prec	Recall	F1	Acc
PAR / DistilBERT	Non-AD	0.76	0.79	0.78	0.77
	AD	0.783	0.75	0.77	
PAR+INV / DistilBERT	Non-AD	**0.83**	0.79	0.81	**0.81**
	AD	0.80	**0.83**	**0.82**	
PAR / TF-IDF/SVM	Non-AD	0.70	0.83	0.75	0.73
	AD	0.79	0.63	0.70	
PAR_SPLT / SVM+CRF	Non-AD	0.78	0.88	**0.82**	**0.81**
	AD	**0.86**	0.75	0.80	
PAR_SPLT+T / SVM+CRF	Non-AD	0.75	**0.88**	0.81	0.79
	AD	0.85	0.71	0.77	

4.2. MMSE Prediction

Table 4 provides our MMSE prediction results on the provided test set. We observe that the deep learning embedding methods perform best, and in particular the DistilBERT model using only participant sections of the transcript performs best RMSE. Interestingly, the deep learning methods perform well despite having not been trained with regression tasks in mind. Our CRF models do not support regression so we cannot report MMSE prediction scores for those model configurations.

Table 4: *Test set results for MMSE score prediction, 'DBL' indicates our DistilBERT embedding with LASSO linear model.*

Dataset/Model	PAR/DBL	PAR+INV/DBL	PAR/SVM
RMSE Score	5.37	4.58	5.22

4.3. Alternative Configurations

The supplied transcripts included temporal metadata for each sentence for participants and interviewers. We experimented with these time time based features alongside the text features at the transcript level, i.e. a PAR+TIME dataset. This included features: participant average / minimum / maximum and median sentence times, and time between sentences. Intuitively, we assumed that AD subjects would exhibit distinctly different time based features due to their impaired cognitive function. However, this dataset (PAR+TIME) performed poorly across the modelling approaches so we do not include in our results. We suggest this is due to the aggregation in the transcript level dataset removing any signal to that could be detected by the modelling approaches. PAR_SPLT+T does include temporal level features but does not perform as as well as linguistic features only.

5. Discussion

We discuss our results in context of model complexity, model generalisability and potential utility as a diagnostic modality. Our most effective models are DistilBERT with PAR+INV and SVM+CRF with PAR_SPLT. Both models perform similarly for

the AD classification task, but the deep learning approach can also output MMSE score predictions. The DL methods will likely generalise better as the majority of the modelling is accomplished by the embedding layer. Both models could be deployed to mobile devices for a potentially ubiquitous early diagnostic tool.

In a potential diagnostic scenario, models would seek to balance recall and precision. A true-positive label of AD would prompt the user to seek a formal evaluation, under medical supervision, potentially leading to an earlier diagnosis allowing for slower progression of the disorder. However, ensuring the false-positive rate is low would minimise unnecessary anxiety during the following formal clinical evaluation.

5.1. Baseline Approaches

The SVM models report higher performance than the GBDT models across all metrics and tasks. They are also computationally faster to fit and cross validate. It is unclear if these models will generalise to alternative or larger datasets. The models have captured correlations in frequency of appearances of 'key words' as identified by TF-IDF vectoriser. Further datasets may result in variations in performance as the frequencies of 'key words' change providing insufficient signal for accurate modelling of the decision boundaries necessary for prediction.

Despite offering the worst performance across all configurations and datasets, GBDTs do provide good model interpretability. For example, we find the top 20 most informative word level features from the TF-IDF vectoriser contain discourse markers such as 'oh', 'uh' and 'um' for both tasks. This indicates the model has found occurrences of these words are useful in making predictions for both tasks. However, prior work has suggested more informative features are more complex[25].

5.2. Deep Learning Approaches

Our results are inline with previous work that has empirically shown that large, pre-trained, DL Transformer based, language models are an effective embedding layer that capture a variety of linguistic phenomenon[26]. As models are pre-trained we incur no model fit expense to use them. Training from scratch requires days or weeks with specialised hardware and large data sets. The simple LR or LASSO models that are fit on top of the fixed size output embeddings are as efficient to fit as the baseline SVM models.

BERT and RoBERTa models are available in their 'base' and 'large' varieties. In prior work, 'large' often performs better due to the increased parameter space and longer training time[22]. However, we observe in our experiments 'large' models are often equivalent or worse performing. We also observe this trend with DistilBERT / DistilRoBERTa that have further reduced parameters compared to 'base' varieties that broadly produce better results in our experiments although prior work would suggest the contrary.

6. Future Work

6.1. Further NLP Modelling

For future work we would look to replicate findings with larger datasets to demonstrate model robustness. We also currently use the models 'out-of-the-box' so they have only been trained with large corpora of prepared speech (Wikipedia and the Toronto Book Corpus[22]). For future work we would also look to fine-tune the deep learning embedding models specifically to spontaneous speech as fine-tuning to domain specific data often boosts performance[27]. Spontaneous speech corpora would likely show a difference in lexicon and grammar as well subtle prosodic differences such as rhythm, tempo that are often captured within spontaneous speech transcripts. An example of such a corpus is the 'The British National Corpus[28]'. A large corpora of informal spontaneous speech containing 1251 recordings and $\sim$11 million words from 668 speakers. We have cleaned and prepared the corpus using a sliding sentence window producing a dataset of $\sim$767k 'documents'. We successfully begun the fine-tuning process observing a reduction in training loss. However, due to extenuating circumstances our GPU resource became unavailable and we were unable to complete the fine-tuning. We make the data pre-processing, and language model fine-tuning scripts available open-source[2].

6.2. Feature Combinations and Model Ensembling

Combining features or model ensembling that incorporated the acoustic data (i.e. prosodic/articulatory features) may provide further gains in performance. Audible phenomenon such as changes in pitch, intonation, stress and subtle changes in tempo would only be available in the audio dataset and have shown to be useful during prior work[29, 25]. We leave the investigation and ensembling of these features to future work.

7. Conclusions

We have presented a range of NLP techniques applied to the ADReSS challenge dataset, a shared task for the prediction of AD and MMSE scores of AD patients and controls. Each dataset and model configuration is rigorously optimised and tested. We observe promising results, above published baselines, for machine learning techniques such as SVMs and Deep Learning approaches. We highlight that the Deep Learning approaches are particularly effective when used as embedding layers for both the AD classification and MMSE score prediction tasks even despite the lack of domain and task specific fine-tuning.

8. Acknowledgements

RD's work is supported by 1.National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. 2. Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. 3. The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, MRC, NIHR or the Department of Health and Social Care.

[2] https://github.com/tomolopolis/ADReSS_Challenge/blob/master/Fine-Tune-LanguageModel.ipynb

9. References

[1] R. Mayeux and Y. Stern, "Epidemiology of alzheimer disease," Cold Spring Harb. Perspect. Med., vol. 2, no. 8, Aug. 2012.

[2] J. Rasmussen and H. Langerman, "Alzheimer's disease - why we need early diagnosis," Degener. Neurol. Neuromuscul. Dis., vol. 9, pp. 123–130, Dec. 2019.

[3] M. J. Kang, S. Y. Kim, D. L. Na, B. C. Kim, D. W. Yang, E.-J. Kim, H. R. Na, H. J. Han, J.-H. Lee, J. H. Kim, K. H. Park, K. W. Park, S.-H. Han, S. Y. Kim, S. J. Yoon, B. Yoon, S. W. Seo, S. Y. Moon, Y. Yang, Y. S. Shim, M. J. Baek, J. H. Jeong, S. H. Choi, and Y. C. Youn, "Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, p. 231, Nov. 2019.

[4] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, Jr, J. Ashburner, R. S. J. Frackowiak, and Alzheimer Disease Neuroimaging Initiative, "Predicting clinical scores from magnetic resonance scans in alzheimer's disease," Neuroimage, vol. 51, no. 4, pp. 1405–1413, Jul. 2010.

[5] Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from continuous speech using language models," J. Alzheimers. Dis., vol. 70, no. 4, pp. 1163–1174, 2019.

[6] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in alzheimer's disease and in its assessment," in Interspeech, 2016, pp. 1948–1952.

[7] 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.

[8] C. de Boer, F. Mattace-Raso, J. van der Steen, and J. J. M. Pel, "Mini-Mental state examination subscores indicate visuomotor deficits in alzheimer's disease patients: A cross-sectional study in a dutch population," Geriatr. Gerontol. Int., vol. 14, no. 4, pp. 880–885, Oct. 2014.

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

[10] B. MacWhinney, "Tools for analyzing talk part 1: The chat transcription format," 2017.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, 2011.

[12] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," IBM J. Res. Dev., vol. 1, no. 4, pp. 309–317, Oct. 1957.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," May 2011.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.

[17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jun. 2001, pp. 282–289.

[18] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 2019, pp. 15–18.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and Others, "Huggingface's transformers: State-of-the-art natural language processing," ArXiv, abs/1910. 03771, 2019.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019.

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019.

[25] 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.

[26] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3651–3657.

[27] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," Jan. 2018.

[28] B. N. C. Consortium and Others, "The british national corpus, version 3 (BNC XML edition)," Distributed by Oxford University Computing Services on behalf of the BNC Consortium, vol. 5, no. 65, p. 6, 2007.

[29] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," Alzheimers. Dement., vol. 1, no. 1, pp. 112–124, Mar. 2015.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multiscale System for Alzheimer's Dementia Recognition through Spontaneous Speech

Erik Edwards, Charles Dognin, Bajibabu Bollepalli, Maneesh Singh

Verisk Analytics

erik.edwards4@gmail.com, charles.dognin@verisk.com, bajibabu.bollepalli@aalto.fi, msingh@verisk.com

Abstract

This paper describes the Verisk submission to The ADReSS Challenge [1]. We analyze the text data at both the word level and phoneme level, which leads to our best-performing system in combination with audio features. Thus, the system is both multi-modal (audio and text) and multi-scale (word and phoneme levels). Experiments with larger neural language models did not result in improvement, given the small amount of text data available. By contrast, the phoneme representation has a vocabulary size of only 66 tokens and could be trained from scratch on the present data. Therefore, we believe this method to be useful in cases of limited text data, as in many medical settings.

Index Terms: Dementia detection, voice classification, computational paralinguistics

1. Introduction and Related Work

Alzheimer's disease (AD) is the most common cause of dementia, a group of symptoms affecting memory, thinking and social abilities. Detecting and treating the disease early is important to avoid irreversible brain damage. Several machine-learning (ML) approaches to identify probable AD and MCI (Mild Cognitive Impairment) have been developed in an effort to automate and scale diagnosis. A comprehensive review of medical-imaging-based approaches was provided by [2], but methods that are less invasive and expensive still require exploration.

Acoustic Approaches: Detection of AD using only audio data could provide a lightweight and non-invasive screening tool that does not require expensive infrastructure, and can be used in peoples' homes. Speech production with AD differs qualitatively from normal aging or other pathologies, and such differences can be used for early diagnosis of AD [3]. Several studies have been proposed to detect AD using speech signals. [4] showed that spectrographic analysis of temporal and acoustic features from speech can characterise AD with high accuracy. [5] used only acoustic features extracted from the recordings of DementiaBank for AD detection, and reported accuracy results of up to 97%.

Linguistic Approaches: There has also been recent work in text-based diagnostic classification approaches; these techniques use either engineered features or deep features.

Engineered Features: [6] showed that classifiers trained on automatic semantic and syntactic features from speech transcripts were able to discriminate between semantic dementia, progressive nonfluent aphasia, and healthy controls. This work was later extended to AD vs healthy control classification [7] using lexical and n-gram linguistic biomarkers.

Deep Features: Deep learning models to automatically detect AD have also recently been proposed. Orimaye et al. [8] introduced a combination of deep language models and deep neural networks to predict MCI and AD. One limitation of a deep-learning-based approach is the paucity of training data typical in medical settings. [9] attempted to interpret what the neural models learned about the linguistic characteristics of AD patients. Text embeddings of transcribed text have also been recently explored for this task. For instance, Word2Vec and GloVe have been successfully used to discriminate between healthy and probable AD subjects [10], while, more recently, multi-lingual FastText embedding combined with a linear SVM classifier has been applied to classification of MCI versus healthy controls [11].

Multimodal Approaches using representations from images have been recently used to detect AD [12, 13]. [14] used lexicosyntactic, acoustic and semantic features extracted from spontaneous speech samples to predict clinical MMSE scores (indicator of the severity of cognitive decline associated with dementia). The work of [15] extended this approach to classification, and obtained state-of-the-art results on DemantiaBank-fused linguistic and acoustic features extracted into a logistic regression classifier.

Multimodal and Multiscale Deep Learning Approaches to AD detection have been applied using medical imaging data [16]. Inspired by this, we propose an Acoustic-Linguistic approach with late fusion to classify AD vs healthy controls. Our contributions are as follows:

1. We introduce a multiscale approach for linguistic features by learning phoneme-level representation from scratch using FastText [17] and Sent2Vec [18]. We show that this phoneme-level embedding can be learned with a very small amount of data, which is a considerable advantage over existing work and ideally suited for clinical settings.
2. We combine speech and text domains to obtain a novel multiscale and multimodal approach to AD recognition. We find that subword (phoneme) information helps the classifier discriminate between healthy and ill participants.

2. Dataset

The dataset was provided by the ADReSS Challenge [1]. The participants were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [19]. Both the speech and corresponding text transcripts were provided. It was released in two parts: train and test sets. The train data had 108 subjects (48 male, 60 female) and the test data had 48 subjects (22 male, 26 female). For the train data, 54 subjects were labeled with AD and 54 with non-AD. The speech transcriptions were provided in CHAT format [20], with 2169 utterances in the train data (1115 AD, 1054 non-AD), and 934 in the test data.

http://dx.doi.org/10.21437/Interspeech.2020-2781

Table 1: *Acoustic features and their dimensions. CFS denotes correlation feature selection and RFECV denotes recursive feature elimination using cross-validation.*

Feature	Dim. (All)	Dim. (CFS)	Dim. (RFECV)
GEMAPS	64	53	3
eGEMAPS	90	76	4
emobase	979	626	6
emobase2010	1583	995	19
emolarge	6511	1810	21
ComParE2016	6375	3592	54
MRCG	6914	367	5

3. Acoustic Systems

All audio started as 16-bit WAV files at 44.1 kHz sample rate. These were provided as 'chunks', which were sub-segments of the above speech dialog segments that had been cropped to 10 seconds or shorter duration (2834 chunks: 1476 AD, 1358 non-AD). In general, the audio data was found to be very noisy and some of the chunks were unintelligible to the human ear. For example, a basic audio classification into 'speech' vs. 'other' using pyAudioAnalysis [21] found only 49.8% of audio chunks were clearly 'speech'.

We applied a basic speech-enhancement technique using VOICEBOX [22], which slightly improved the audio results, but is not essential to our method. We also tried rejecting noisy chunks, or using a 3-category classification scheme to separately identify the noisiest chunks. These attempts did not significantly improve the results, however, and so were not pursued further. We also attempted using voice activity detection, using OpenSMILE [23] or rVAD [24], and weighting audio results accordingly. This led to small improvements for some analyses, but was also not included in the final results, as it was apparent that more radical changes in methodology would be required to deal with these noise levels (e.g., a windowing into fixed-length frames). We decided therefore to use the noisy audio 'chunks' as given, with only the basic speech enhancement applied, and to defer additional improvements to future work.

3.1. Acoustic Features

Acoustic features were extracted on the enhanced speech segments downsampled to 16-kHz sample rate. We used the same feature sets as in the baseline Challenge paper [1], along with a few additional sets, but also added a stage of feature selection. Features are computed every 10-ms to give "low-level descriptors" (LLDs) and then statistical functionals of the LLDs (such as mean, standard deviation, kurtosis, etc.) are computed over each audio chunk of 0.5-10 sec duration (chunks shorter than 0.5 s were rejected). Using OpenSMILE [23], we extracted the following sets of functionals: emobase [25], emobase2010, GeMAPS [26], extended GeMAPS (eGeMAPS), and ComParE2016 (a minor update of numerical fixes to the ComParE2013 set [27]). Using code from the Cacophony Project (https://github.com/TheCacophonyProject), we extracted multi-resolution cochleagram (MRCG) LLDs [28], and then several statistical functionals of these. The dimensions of each functionals set are given in Table 1, and details can be found in the cited references.

3.2. Acoustic Feature Selection

As the dimensionality of each functionals set was large (Table 1), we explored feature selection techniques to improve subsequent classification. First, we used correlation feature selection (CFS), which discards highly-correlated features. Second, we used recursive feature elimination with cross validation (RFECV), where a classifier is employed to evaluate the importance of the each feature dimension. In each recursion, the feature that least improves or most degrades classifier importance is discarded, leading to a supervised ranking of features.

Table 2: *Accuracy scores of feature selection. These numbers calculated by taking majority vote on segments.*

Feature	All	CFS	RFECV
GEMAPS	0.490	0.472	0.629
eGEMAPS	0.453	0.462	0.620
emobase	0.555	0.555	0.657
emobase2010	0.555	0.574	0.601
emolarge	0.595	0.629	0.666
ComParE2016	**0.601**	**0.629**	**0.694**
MRCG	0.546	0.509	0.611

Table 3: *Accuracy scores of the ComParE2016 acoustic feature set with different classifiers. LR: Logistic regression, SVM: support vector machine, and LDA: linear discriminant analysis.*

Feature	LR	SVM	LDA
ComPareE2016	0.694	**0.740**	**0.740**

Table 1 shows the raw ("All") feature dimensions and after each feature selection method. We further appended age and gender to each acoustic feature set. With CFS, we discarded features with correlation coefficient ≥ 0.85. For RFECV, we used logistic regression (LR) as the base classifier with leave-one-subject-out (LOSO) cross validation. CFS reduced the dimensionality by 15-95%, and the RFECV method further brought the dimensionality down to 3-54 for all sets.

Table 2 shows the performance of feature selection methods employed in this study, assessed with LOSO cross-validation on the train set. There is considerable improvement in accuracy after the CFS and RFECV methods. Since the performance of the ComParE2016 features is best among the acoustic feature sets, we used only the ComParE2016 features for further experiments. However, it is noted that equivalent performance could be obtained with emobase2010 using other feature selection methodology (not included here).

Table 3 presents the accuracy scores achieved by the ComParE2016 features using different ML classification models. SVM (support vector machine) and LDA (linear discriminant analysis) models gave better performance than LR. The best accuracy obtained using acoustic features alone is 0.74. For our ensemble models, we used the posterior probabilities from the LDA model averaged over all chunks for each subject.

4. Linguistic Systems

The linguistic system contains two parts: the natural language representation and the phoneme representation.

4.1. Natural Language Representation

We applied a basic text normalisation to the transcriptions by removing punctuation and CHAT symbols and lower casing. Table 4 shows the accuracy and F_1 score results on a 6-fold cross validation of the training data-set (segment level). For each model used, hyper-parameter optimisation was performed to allow for fair comparisons.

4.1.1. Engineered Features

Following [7] and [9], we extract seven features from text segments: richness of vocabulary (measured by unique word count), word count, number of stop words, number of coordinating conjunction, number of subordinated conjunction, average word length, and number of interjections. Using CHAT symbols, we extract four more features: number of repetitions (using [/]), number of repetitions with reformulations (using [//]), number of errors (using [*]), and number of filler words (using [&]).

4.1.2. Deep Learning Features

We experimented with three different settings: Random Forest with deep pre-trained Features (DRF), fine-tuning of pre-trained models (FT) and training from scratch (FS).

Deep Random Forest Setting: We extract features using three pre-trained embeddings: Word2Vec (CBOW) with subword information [29] (pre-trained on Common Crawl), GloVe [30] pre-trained on Common Crawl and Sent2Vec [31] (with uni-grams) pre-trained on English Wikipedia. The procedure is the same for each model: each text segment is represented by the average of the normalised word embeddings. The segment embeddings are then fed to a Random Forest Classifier. In this setting the best performing model is Sent2Vec with unigram representation. Sent2Vec is built on top of Word2Vec, but allows the embedding to incorporate more contextual information (entire sentences) during pre-training.

Training from Scratch Setting: In this setting, models are trained from scratch on the given dataset. The only model fast enough to allow us to find the best hyper-parameters while being a good baseline is FastText. With an embedding dimension as low as 5 and with as low as 16 words in its vocabulary, FastText performs competitively compared to most of the Deep Random Forest Settings. Subword information determined by character n-grams are keys to this result as explained below.

Fine-Tuning Setting: For this final setting, pre-trained embeddings (Word2Vec, GloVe, Sent2Vec) or models (Electra [32], Roberta [33]) are fine-tuned on the data. Electra uses a Generator/Discriminator pre-training technique more efficient than the Masked Language Modeling approach used by Roberta. Though the results of the two models are approximately the same at the segment level, Electra strongly outperforms Roberta at the participant level. The best models still remain the ones using subword information: GloVe (FT) and Word2Vec (FT). Both of those pre-trained embeddings are fine-tuned with the FastText classifier. The later turn sentences into character-n-gram augmented sentences (we found that a maximum character n-grams of 6 was optimal). Though FastText from scratch also have the sub-word information, it does not have the pre-trained representation of those sub-words learnt using GloVe or CBOW (Word2Vec).

4.1.3. Interpretation and Discussion

Subword Information appears to be a key discriminative feature for effective classification. As Figure 1 shows, not using subword information is detrimental to the discriminative power of the model. As a result, we can make the hypothesis that in low resource settings like in this case of medical data, taking into account subword information might be the key to good performance. We explore even further this hypothesis by transforming sentences into phoneme level sentences.

Table 4: *Best performance after hyper-parameters optimisation for each model, metrics are the average of accuracy and f1 scores across 6-fold cross-validation, participant level (softmax average).*

Model	Dim.	Accuracy	F1-score
Random (DRF)	11	0.463	0.482
Engineered Feat (DRF)	11	0.704	0.68
Sent2Vec (FT)	600	0.787	0.758
GloVe (FT)	300	0.861	0.865
Word2Vec (FT)	300	**0.926**	**0.923**
Word2Vec (DRF)	300	0.787	0.785
GloVe + EF (DRF)	311	0.796	0.792
Sent2Vec (DRF)	600	0.833	0.83
GloVe (DRF)	300	0.824	0.822
FastText (FS)	5	0.796	0.776
Roberta-Base (FT)	768	0.787	0.753
Electra-Base (FT)	768	0.861	0.845

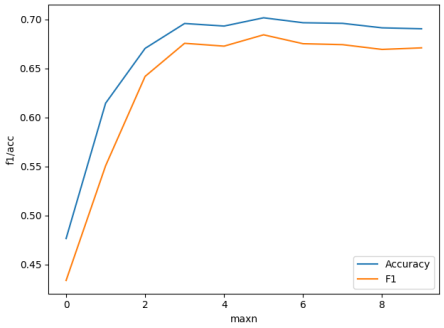


Figure 1: *F1 and Accuracy on 6-fold cross validation as a function of the maximum size of character n-grams (maxn) using FastText supervised classifier*

Word Order: When word order is important, FastText tends to not perform well as it averages the word embeddings of the input sentences without accounting for their original position. We confirmed this hypothesis by measuring the impact of adding word n-grams as additional features to the classifiers. Figure 2 shows that adding word n-grams, thus introducing temporality, does not impact the performance or even degrade it.

Performance of Transformers Though Transformers have subword information through the use of Byte Pair Encoding tokenizer for Roberta and WordPiece tokenizer for Electra, there are too few data points for their large number of parameters.

Experiment Details For the Random Forest (RF), we found that the best results on the 6-fold cross validation were obtained using 200 estimators, entropy criterion, square root for the maximum number of features. A StandardScaler (subtracting the mean and scaling to unit variance) was also applied to the features. FastText From Scratch (FS) hyper-parameters are: wordNgrams=1, 100 epochs, max number of character n-grams=6, minimum number of word occurences=100, learning rate of 0.05 and embedding dimension of 5. We kept the same hyper-parameters for FastText fine-tuned except for the dimension that we set to 300 for Word2Vec and GloVe and 600 for Sent2Vec. Roberta-Base and Electra-Base performance was measured on the best hyper-parameters found. The hyper-parameters that were found to work best are: a batch size of 16, 5 epochs, a maximum token length of 128 and a learning rate of 2e-05.

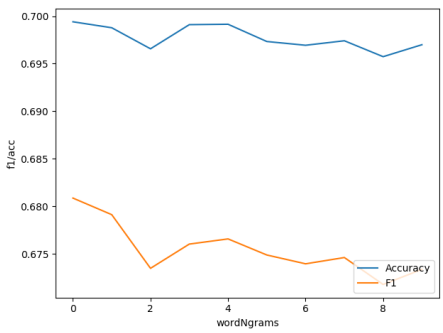


Figure 2: *F1 and Accuracy on 6-fold cross validation as a function of the word n-grams (wordNgrams) features using FastText supervised classifier*

Table 5: *Results of 9-fold CV on the Train set for several combined systems, using simple LR on posterior probability outputs. Audio represents the LDA posterior probabilities of ComParE2016. Word2Vec and GloVe were text (word-based) systems (Section 4.1) and Phonemes are as in Section 4.2. Age and speaking rate were added to each system.*

Model	Accuracy
GloVe + Phonemes	0.8981
GloVe + Phonemes + Audio	0.9074
Word2Vec + Phonemes	**0.9352**
Word2Vec + Phonemes + Audio	**0.9352**

4.2. Phonetic Representation

The discriminative importance of the subword information was confirmed by our phoneme transcription experiments. We transcribed the segment text into phoneme written pronunciation using CMUDict [34]. The most likely pronunciation is used for words with multiple pronunciations. Thus, "*also taking cookies*" becomes "*ao1 l s ow0 t ey1 k ih0 ng k uh1 k iy0 z*". In several experiments, it always helped to include vowel stress in the pronunciation (0 is no stress, 1 is full stress, 2 is part stress). With stress, there were 66 phones total.

Several text classifiers were trained on the phoneme representation (FastText, Sent2Vec, StarSpace), and FastText was again found to perform best (and fastest). Our best performance on the Test set (Table 6) used only the phoneme representation and FastText classification, along with the audio. However, in 9-fold CV tests with the Train set, the best result was a combination of natural language and phonetic representation (Table 5).

The numbers appended to vowel phonemes are stress indicators according to the convention of CMUdict. Our experiments showed that removing stress always caused a decrease in performance. The discriminative importance of phonetic and articulatory representation in AD patient is in accord with previous medical research (e.g., [35]), and deserves future experimentation for ML purposes.

Experiment Details For the phonetic experiments, we used FastText supervised classifier with the following hyperparameters: 4 wordNgrams, an embedding dimension of 20, a learning rate of 0.05, 300 epochs, and a bucket size of 50000. The other hyperparameters were at default. We did not use character n-grams (many phones are already characters).

Table 6: *Challenge Test Set Results*

Model	Class	Precision	Recall	F1 Score	Accuracy
System 1	non-AD	0.6316	0.5	0.5581	0.6042
	AD	0.5862	0.7083	0.6415	
System 2	non-AD	0.7407	0.8333	0.7843	0.7708
	AD	0.8095	0.7083	0.7556	
System 3	non-AD	0.7692	0.8333	0.8	**0.7917**
	AD	0.8182	0.7500	0.7826	
System 4	non-AD	0.7308	0.7917	0.76	0.75
	AD	0.7727	0.7083	0.7391	
System 5	non-AD	0.75	0.75	0.75	0.75
	AD	0.75	0.75	0.75	

5. Discussion

- **System 1**: Audio (LDA posterior probabilities of ComParE2016 features)
- **System 2**: Phonemes (as in Section 4.2)
- **System 3**: Phonemes and Audio
- **System 4**: Phonemes and Word2vec (as in Section 4.1)
- **System 5**: Phonemes and Audio and Word2Vec

For each combined system (Tables 5 and 6), we appended the age and speaking rate as auxiliary features. Those two variables are well studied for identifying AD (see [36] for the positive correlation with age and [37] for the negative correlation with speech rate).

Acoustic Features alone are not as discriminative as text features alone. There is indeed a 15 points difference in accuracy between System 1 which mainly use acoustic features and System 4 which mainly uses text features. However, the audio was very noisy in this set; new feature sets and robustness measures should be explored.

Deep learning text systems easily overfit for small data. RoBERTa and Electra models performed worse than Word2Vec on this small dataset (Table 4), and systems 4 and 5 perform worse on the final Test set than just Phonemes alone (Table 6). However, 9-fold CV on the Train set (Table 5) found that the best performing system was multiscale (Word2Vec and Phonemes) as well as multimodal (text and audio) (Table 5). We believe this would also give the best result for the Test set if the amount of data were larger.

Using Phoneme/Subword is key. The effectiveness of using subword features to discriminate between AD and non-AD people can be understood as analogous to data augmentation. Splitting tokens into subwords or mapping them to phonemes reduces the size of the vocabulary and at the same time expands the number of tokens in the training set. Also, several studies like [35] have found that AD patients show articulatory difficulties and patterns which would show on the phonetic transcription. Phoneme representations also capture many simple aspects of word-based text models, noting that phoneme 4-grams as used here already include many basic words.

6. Conclusions

We propose a multiscale approach to the problem of automatic Alzheimer's Disease (AD) detection. We find that subword information, and in particular phoneme representation, helps the classifier discriminate between healthy and ill participants. This finding could prove useful in many medical or other settings where lack of data is the norm.

7. References

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

[2] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation," *Medical image analysis*, p. 101694, 2020.

[3] 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*, p. 113213, 2020.

[4] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia?" *Dementia and geriatric cognitive disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[5] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for Alzheimer's disease," in *Proc SLPAT Workshop*, 2016, pp. 32–36.

[6] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, pp. 43–60, 2014.

[7] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC bioinformatics*, vol. 18, no. 1, p. 34, 2017.

[8] S. O. Orimaye, J. S.-M. Wong, and C. P. Wong, "Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia," *PloS one*, vol. 13, no. 11, 2018.

[9] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," *arXiv:1804.06440*, 2018.

[10] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations." in *Proc INTERSPEECH*. ISCA, 2018, pp. 1893–1897.

[11] K. C. Fraser, K. L. Fors, and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Computer speech & language*, vol. 53, pp. 121–139, 2019.

[12] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen *et al.*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.

[13] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann *et al.*, "Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection," *The Lancet neurology*, vol. 14, no. 10, pp. 1037–1053, 2015.

[14] M. Yancheva, K. C. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *Proc SLPAT Workshop*, 2015, pp. 134–139.

[15] 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.

[16] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv:1607.01759*, 2016.

[18] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *arXiv:1703.02507*, 2017.

[19] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "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.

[20] B. MacWhinney, *The CHILDES project: tools for analyzing talk*. Psychology Press, 2014, vol. I.

[21] T. Giannakopoulos, "pyAudioAnalysis: an open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[22] M. Brookes, "VOICEBOX: speech processing toolbox for matlab," Imperial College London, UK, 2010. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[23] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc ACM Conf Multimedia*, 2010, pp. 1459–1462.

[24] Z.-H. Tan, N. Dehak *et al.*, "rVAD: an unsupervised segment-based robust voice activity detection method," *Computer speech & language*, vol. 59, pp. 1–21, 2020.

[25] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc ICASSP*, vol. 1. IEEE, 2004, pp. 577–581.

[26] 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 transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[27] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[28] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[30] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for word representation," in *Proc Conf EMNLP*, 2014, pp. 1532–1543.

[31] P. Gupta, M. Pagliardini, and M. Jaggi, "Better word embeddings by disentangling contextual n-gram information," in *Proc NAACL-HLT Conf*, vol. I. ACL, 2019, pp. 933–939.

[32] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: pre-training text encoders as discriminators rather than generators," *arXiv:2003.10555*, 2020.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: a Robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.

[34] R. Weide, "The Carnegie Mellon pronouncing dictionary [CMUdict. 0.6]," *Pittsburgh, PA: Carnegie Mellon Univ.*, 2005.

[35] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, "Phonological and articulatory impairment in Alzheimer's disease: a case series," *Brain and language*, vol. 75, no. 2, pp. 277–309, 2000.

[36] R. Guerreiro and J. Bras, "The age factor in Azheimer's disease," *Genome medicine*, vol. 7, no. 1, p. 106, 2015.

[37] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in Alzheimer's disease, is that an early sign? importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge

Anna Pompili[1], *Thomas Rolland*[1,2], *Alberto Abad*[1,2]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal

anna.pompili@inesc-id.pt, thomas.rolland@hlt.inesc-id.pt, alberto.abad@inesc-id.pt

Abstract

This paper describes a multi-modal approach for the automatic detection of Alzheimer's disease proposed in the context of the INESC-ID Human Language Technology Laboratory participation in the ADReSS 2020 challenge. Our classification framework takes advantage of both acoustic and textual feature embeddings, which are extracted independently and later combined. Speech signals are encoded into acoustic features using DNN speaker embeddings extracted from pre-trained models. For textual input, contextual embedding vectors are first extracted using an English Bert model and then used either to directly compute sentence embeddings or to feed a bidirectional LSTM-RNNs with attention. Finally, an SVM classifier with linear kernel is used for the individual evaluation of the three systems. Our best system, based on the combination of linguistic and acoustic information, attained a classification accuracy of 81.25%. Results have shown the importance of linguistic features in the classification of Alzheimer's Disease, which outperforms the acoustic ones in terms of accuracy. Early stage features fusion did not provide additional improvements, confirming that the discriminant ability conveyed by speech in this case is smooth out by linguistic data.

Index Terms: Alzheimer's Disease, automatic multi-modal diagnosis, acoustic and textual feature embeddings

1. Introduction

Alzheimer's Disease (AD), the most common cause of Dementia [1], is a neurodegenerative disorder characterized by loss of neurons and synapses in the cerebral cortex. Its prevalence increases with age, a study on the U.S. census reported that 3% of people aged 65-74, 17% of people aged 75-84, and 32% of people aged 85 and older have AD [2]. As most countries are experiencing a general increase in average lifespan, it is expected a rapidly escalation of AD cases worldwide in the next thirty years [3]. Pharmacological treatments may temporarily improve the symptoms of the disease, but they can not stop or reverse its progression. For these reasons, there is an increasing need for additional, noninvasive, and cost-effective tools allowing a preliminary identification of AD in its early clinical stages. Currently, AD is diagnosed through an analysis of patient clinical history and disability, neuropsychological tests, brain imaging and cerebrospinal fluid exams. Although the prominent symptoms of the disease are alterations of memory and of spatial-temporal orientation, language impairments are also an important factor confirmed by current literature [4, 5]. Some of the most well known language impairments found in AD speech include naming [4], word-finding difficulties [6], repetitions [7], an overuse of indefinite and vague terms [8], and inappropriate use of pronouns [9].

Over the last years, there has been an increased interest from the research community in the automatic identification of AD through the analysis of speech and language abilities. Some studies have focused on syntactic or semantic features [10, 11], some targeted plain acoustic approaches [12, 13], while other works have investigated a combination of temporal speech parameters and lexical measures [14, 15]. Most of these approaches use handcrafted features and traditional classification algorithms. Very recent works investigated the use of automatically learned representations from deep neural networks [16–19]. Regardless of the approach used, the studies existing in the literature are difficult to analyze and compare due to the different datasets used. In this scenario, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge has been proposed, with the aim of providing researchers with a common, statistically balanced and acoustically enhanced dataset to test their approaches [20].

In this work, we present the multi-modal system proposed by the Human Language Technology Laboratory of INESC-ID for the ADReSS 2020 challenge. Our framework is designed to solve the task of automatically distinguishing AD patients from healthy individuals. In our previous approaches to this topic [21, 22] we exploited lexical, syntactic, and semantic features with measures of local, global, and topic coherence, in order to provide a more comprehensive characterization of language abilities in AD and thus a more reliable identification. In this work, we take the challenge of using automatically learned representations instead of traditional and consolidated handcrafted features, which already proven to achieve good classification results. Inspired by recent studies, we push the limit of deep neural models to work with extreme conditions, such the ones in the health domain, in which data scarcity is ordinary. Additionally, we combine both acoustic and linguistic information to have a complete picture of patient's disabilities, in a similar way to the type of information that clinicians receive during their interactions with patients.

The rest of this work is organized as follows: Section 2 introduces the relevant state on the art on the automatic identification of AD. Then, in Section 3 and 4, we present the dataset used in this study and a description of our methodology. Finally, classification results are reported in Section 5, while conclusions are summarized in Section 6.

2. Related work

The computational analysis of speech and language impairments in AD has gained growing attention in recent years. Initially, existing studies explored engineered temporal and acoustic parameters of speech, linguistic features, or a combination

This work has been partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and by European Union funds through Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287.

http://dx.doi.org/10.21437/Interspeech.2020-2833

of both. König *et al.* [12] computed several temporal speech features on a dataset composed of 26 AD and 15 healthy subjects, while performing different tasks of isolated and continuous speech. By considering different features according to the task, the authors achieved an accuracy of 87% in the automatic identification of AD. Fraser *et al.* [11] used more than 350 features to capture lexical, syntactic, grammatical, and semantic phenomena from the transcriptions of a picture description task. With a selection of 35 features, the authors achieved a classification accuracy of 81.92% in distinguishing individuals with AD from healthy controls. Pompili *et al.* [21] exploited lexical, syntactic, semantic and pragmatic features from the descriptions of the Cookie Theft picture [23] attaining an accuracy of 85.5% in the task of classifying AD patients. Gosztolya *et al.* [14] collected a dataset composed of 75 Hungarian speakers (25 AD, 25 MCI, and 25 healthy subjects) performing two tasks eliciting continuous speech. The set of features used considered demographic attributes, acoustic and linguistic features. Using only acoustic or linguistic information the authors achieved an accuracy of 82% in distinguishing AD patients from healthy subjects. When the two types of features were combined, the accuracy increases to 86%.

More recently, researchers are shifting their focus towards more complex architectures capable of overcoming the limitations of traditional approaches. Warnita *et al.* [18] proposed an approach relying only on acoustic data computed from continuous speech and gated Convolutional Neural Network (GCNN). Using majority voting on speaker and the Paralinguistic Challenge (IS2010) feature set, the authors achieved an accuracy of 73.6%. Karlekar *et al.* [19], on the other hand, investigated linguistic impairments using CNN, LSTM-RNNs, and a combination of both. In this way, they obtained an accuracy of 91.1% in classifying AD patients. Chen *et al.* [16] went further, proposing a network based on attention mechanism and composed of a CNN and GRU module. In this way, the architecture should be able to analyze both local speech patterns and global macro-linguistic functions. The accuracy achieved in distinguishing AD patients was of 97.42%. Finally, Zargarbashi *et al.* [17] designed a multi-modal feature embedding approach based on N-gram, *i-vectors*, and *x-vectors*. Classification accuracy results achieved with each of these models were, respectively, of 78.2%, 75.9%, and 75.1%. The joint fusion of the three models reached an accuracy of 83.6%.

Our work differs from previous studies for several reasons. First, to process the text data, we use contextual embeddings vectors as input to two different systems. One based on the training of a Global Maximum pooling and a bidirectional LSTM-RNNs architectures, and one based on the statistical computation of sentence embeddings. The latter presents the advantage of being a simple approach, which does not require the training of deep, data-demanding architectures. Second, for the audio recordings, we use DNN speaker embeddings extracted from pre-trained models. These learned, speaker representative vectors have recently shown their potential in the discrimination of neurodegenerative disorders [24]. To the best of our knowledge, this is the first work that jointly uses automatically learned representations from neural models, instead of engineered features, for both audio signals and textual data. In fact, although existing studies have shown that linguistic impairments in AD appear to be more important than acoustic ones, traditional literature provide convincing evidence that using both source of information will definitively improve the accuracy of automatic diagnosis methods.

Table 1: *Statistical information on the ADReSS dataset*

	Train		**Test**
	Control	AD	–
Audio Full	00:55:46	01:14:00	01:06:00
Audio chunks	00:30:11	00:26:31	00:26:32
# words (unique)	6097 (567)	5494 (552)	5536 (602)

3. Corpus

The ADReSS dataset contains the speech recordings and corresponding annotated transcriptions of 156 subjects, 78 AD patients, and 78 healthy control matched for age and gender. Data were divided into two partitions, training and test sets composed of 108 and 48 subjects, respectively. Recorded participants were required to provide the descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [23]. Speech recordings were segmented using Voice Activity Detection (VAD) and later normalised [20]. The dataset contained both full enhanced audio, and normalised audio chunks.

In our approach, we have used both the full enhanced audio and the transcriptions. The latter were annotated with disfluencies, filled pauses, repetitions, and other more complex events. However, to build an automated system requiring a minimal annotation effort, we removed all the annotations not corresponding to the plain textual representation of words, thus, better resembling the output that can be generated by an Automatic Speech Recognition (ASR) system. Overall, the whole set of transcriptions contained 17127 words, of which 1009 were unique. More detailed information about the duration and size of the ADReSS dataset are reported in Table 1.

4. Proposed methods

As shown in Figure 1, our multi-modal framework is based on the independent generation of acoustic and textual feature embeddings. Then, we perform an early fusion of the output of the two systems to obtain a single feature vector containing a compact representation of both speech and language characteristics. Finally, classification is performed with an SVM classifier with linear kernel. The two systems are described in the remainder of this section.

4.1. Acoustic system

The acoustic system is strongly based on two models borrowed from the speaker verification field, *i-vectors* [25] and *x-vectors* [26]. *i-vectors* are statistical speaker representation vectors that have been recently used for the classification of Parkinson's Disease and for the automatic prediction of dysarthric speech metrics [27, 28]. *X-vectors* are discriminative deep neural network-based speaker embeddings that have outperformed *i-vectors* in speaker and language recognition tasks [26, 29, 30] and have been successfully applied to AD, obstructive sleep apnea and pathological speech detection [24, 31]. Both models allow to extract a fixed sized feature vector from variable length audio signal.

Taking into consideration the small size of the ADReSS dataset, we preferred to exploit already existing pre-trained models to produce our acoustic feature embeddings, rather than training them using in-domain challenge data. To this end, for the *x-vectors* framework we use both the SRE and the Voxceleb models. The first was trained mainly on telephone and microphone speech using data from the Switchboard corpus, Mixer 6,

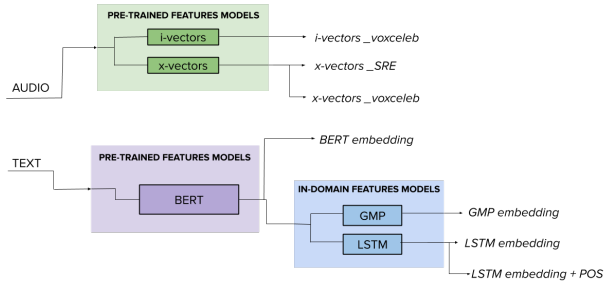


Figure 1: *Summary of embedding-based approaches*

and NIST SREs [29]. The latter was trained on augmented VoxCeleb 1 and VoxCeleb 2 datasets, which contains speech from speakers spanning a wide range of different ethnicities, accents, professions and ages. [29, 32]. This dataset was used also to build the *i-vectors* pre-trained model used in this work.

The inputs to the pre-trained SRE and Voxceleb models consisted of 23 and 30-dimensional MFCC vectors, extracted with Kaldi [33] from the full recordings, using default values for window size and shift. Non-speech frames were removed using energy-based VAD. For the *x-vectors* model, the last layers of the pre-trained model, before the softmax output layer, can be used to compute the embeddings. In this work, we extracted a 512-dimensional *x-vectors* at layer *segment6* of the network.

The *i-vectors* models, is based on GMM-UBM. The universal background model (UBM) is used to capture statistics about intra-domain and inter-domain variabilities and a projection matrix is used to compute *i-vectors*. We extracted a 400-dimensional *i-vectors*.

4.2. Linguistic system

We followed two different approaches to obtain textual feature embeddings. First, we investigated the feasibility of training deep architectures with a corpus of reduced dimension like the one used in this challenge. Then, this method is compared with a less data-demanding one, based on the statistical computation of sentence embeddings using a pre-trained model. Both strategies rely on contextual word embeddings as input, but they provide different types of learned representations as output. In fact, to combine the information from the linguistic and the acoustic systems, the trained architectures are used only to extract linguistic features from the last layer of the models, before the final classification. In this way, we obtain a single 768-dimensional feature vector for an entire description. The sentence embedding approach, on the other hand, provide a single 768-dimensional vector for each sentence of a description. These features are then used to classify between AD patients and healthy subjects. For both approaches, the first step of the pipeline deals with the normalization of the data provided in the ADReSS dataset. In fact, we recall that besides the plain transcription of the descriptions these also contain additional annotations and information that were removed. Then, we encode each word of the clean transcriptions into a 768-dimensional context embedding vector using a frozen English Bert model pre-trained with 12-layers, 768-hidden. This representation is fed to our two linguistic systems, described hereafter.

The first system is derived from the ComParE2020 Elderly Challenge baseline [34], and was obtained by adapting the original code to deal with the classification of AD. With this approach, three different neural models are trained on top of contextual word embeddings: (i) a Global Maximum pooling, (ii) a bidirectional LSTM-RNNs provided with an attention module, and (iii) the second model augmented with part-of-speech (POS) embeddings. During training, the loss is evaluated on the development set.

The second system provides the advantage of not requiring an additional phase of model training. Similarly to the approach followed with the acoustic system, we use automatically learned representations extracted from a pre-trained model to directly characterize linguistic deficits in AD. The contextual word embeddings obtained for each word of the clean transcriptions are now used to compute an embedding vector of fixed size for each sentence of a description. Sentence embeddings were successfully employed in tasks of humor detection and more generally sentiments analysis [35, 36] and information retrieval [36]. In our approach, sentence embeddings are computed by averaging the second to twelfth hidden layers of each word.

Table 2: *Results of different acoustic approaches on the development set*

	Accuracy	Precision	Recall	F1 Score
x-vectors_Vox	0.6818	0.6834	0.6919	0.6812
x-vectors_SRE	**0.7273**	**0.7273**	**0.7273**	**0.7273**
i-vectors_Vox	0.6818	0.7292	0.6818	0.6645
***i-vectors*_Vox_*x-vectors*_Vox**	0.7273	0.7273	0.7273	0.7273
***i-vectors*_Vox_*x-vectors*_SRE**	0.7273	0.7351	0.7273	0.7250

5. Results and discussion

The ADReSS dataset contains only training and test partitions and for the latter the ground truth is not provided. Thus, in order to test our approaches, we retain the 20% of the data from the training set and use it as development set. In this way, we are left with 86 subjects for training, 22 for development, and 48 for testing. While creating the additional partition, we kept the dataset gender balanced.

As briefly mentioned, our evaluation method relies on SVM [37] with linear kernel, based on a liblinear implementation. The complexity parameter C was optimised during the development phase. The results reported in Tables 2 and 3 are obtained using the best complexity configuration. Features were normalized to have zero mean and unit variance. In the remainder of this section we first describe our results on the development set for each system independently and then for their final fusion. Finally, for the best systems, we report the results obtained on the test set.

5.1. Results on the development set

5.1.1. Acoustic system

Results using different automatically learned acoustic features embeddings are summarized in Table 2. Also in this case, we explored different independent models and then we do an early fusion of the best acoustic results attained. From Table 2 is possible to observe that the *x-vectors* Voxceleb model usually achieve a lower classification accuracy. However, when we combine both *i-vector*s and *x-vectors* extracted from this model, the accuracy resulting from their fusion is comparable to that of *x-vectors* using the SRE model, which is currently our best result on the development set. These outcomes are slightly lower than those found in the literature review for similar works. In fact, we recall that Warnita *et al.* [18] and Zargarbashi *et al.* [17]

Table 3: *Results of different linguistic approaches on the development set*

	Accuracy	Precision	Recall	F1 Score
Global Max Pool.	0.7727	0.7947	0.7728	0.7684
LSTM-RNNs	0.8182	0.8182	0.8182	0.8182
LSTM-RNNs Pos	0.8636	0.8667	0.8637	0.8634
GMax/LSTM-RNNs/LSTM-RNNs-Pos	**0.9091**	**0.9091**	**0.9091**	**0.9091**
Sentence emb. - maj. vote	0.7727	0.7947	0.7728	0.7684

Table 4: *Results of different acoustic and linguistic approaches on the test set*

	Class	Accuracy	Precision	Recall	F1 Score
Fusion of system	AD	**0.8125**	0.9412	0.6667	0.7805
	non-AD		0.7419	0.9583	0.8364
Sentence embedding	AD	**0.7292**	0.8235	0.5833	0.6829
	non-AD		0.6774	0.8750	0.7636
x-vectors_SRE	AD	**0.5417**	0.5417	0.5417	0.5417
	non-AD		0.5417	0.5417	0.5417

obtained an accuracy of 73.6%, 75.9%, and 75.1%, using, respectively a gated CNN with the IS10 acoustic feature set and the *i-vectors*/*x-vectors* paradigms. Our approach, however, is different from the ones of these authors since we are using a smaller dataset and do not rely on DNN training. Nevertheless, since we are interested in corroborating these results on the test set, we select the acoustic feature embeddings extracted from the pre-trained *x-vectors* SRE model for the evaluation.

The use of pre-trained acoustic embedding extractors has been motivated by the reduced size of the ADReSS dataset, that we considered to be insufficient for data hungry deep learning approaches. To confirm this, we also trained an end-to-end LSTM model for AD classification. The architecture consisted of one dense and two LSTM layers with a softmax activation function. The network took as input chunks of 500 voiced frames using 23-dimensional MFCC with delta and delta-delta. Majority voting was performed over all the chunks from the same speaker to generate a single prediction per speaker. This end-to-end approach performed very poorly, with an accuracy around chance result in the development set, confirming our expectations that the ADReSS dataset is not suited for training a deep learning end-to-end system.

5.1.2. Linguistic system

Results obtained with our different linguistic systems are summarized in Table 3. This table reports the performance for the features trained with the three neural models, their fusion, and finally for the sentence embeddings approach. For the latter, we present only results achieved using a majority voting over the entire description. Our best classification result attained an accuracy of 90.91% on the development set using the fusion of the linguistic features sets generated by the three neural models. Comparing this result with the one obtained by sentence embeddings, we acknowledge that neural models outperform simpler strategies even with constrained training data. This was somehow surprising and in contradiction with similar experiments performed with the acoustic system. We hypothesize that the large amount of contextual information provided by the Bert model is helpful in overcoming the limited size of the ADReSS dataset. Nevertheless, we suspect that the high accuracy attained with neural models may be too optimistic, due to the fact of having used the development set both for testing and evaluating the model's loss. Thus, in spite of their lower outcome, the sentence embeddings approach is selected as one of the systems to be evaluated on the test set. In fact, on the one hand, we think that they may represent a more reliable system, since do not require additional training. On the other hand, we also observe that they achieve higher classification scores, when compared with a similar approach based on GloVe embeddings [38], thus corroborating our decision.

5.1.3. Fusion of systems

To provide a comprehensive evaluation of speech and language impairments in AD, the best results obtained with both the acoustic and the linguistic systems where combined together in an early fusion fashion. We merged the *x-vectors* features set obtained with the SRE model with the combination of linguistic feature sets (GMax/LSTM-RNNs/LSTM-RNNs-Pos) generated by the three neural models. Unfortunately, results on the development set using this extended set of features did not provide any further improvements with respect to using the linguistic system alone. We believe that, in this case, the predictive ability of linguistic features completely override acoustic ones. Nevertheless, we select the combination of these two systems as our main system for the evaluation.

5.2. Results on the test set

Overall, we submitted three systems for the evaluation: (i) the fusion of the best results achieved by the linguistic and acoustic systems, (ii) sentence embeddings, (iii) the best acoustic system. A summary of these results is reported in Table 4. In general, we found a consistent impoverishment of the performance of our methods when evaluated on the test set, even for those systems based on features that do not required a training phase. The first system submitted achieved the best result, with an accuracy of 81.25%, showing that the use of deep architectures with contextual word embeddings are actually able of overcoming the limitation of a constrained dataset. The worse result is achieved by the acoustic system alone, with an average accuracy of 54.17%. This outcome is lower than the one found in the ADReSS baseline (62.50%) [20], indicating that there is still room for improving our acoustic approach. We relied on pre-trained models to overcome the lack of data, but we ended up with a similar problem. It is likely the case that an adaptation of these models to the characteristics of elderly speech would allow for better performance.

6. Conclusions

In this work we presented a multi-modal approach to the classification of AD based on automatically learned feature representations. Both for the acoustic and linguistic systems, we investigated feature embedding vectors extracted from pre-trained models, as well as the feasibility of training deep neural architectures. Using a combination of both approaches, we attained an accuracy of 90.91% and 81.25% on the development and test sets, respectively. Our results showed that acoustic systems, in comparison to linguistic ones, require more data in order to improve the predictive ability of neural models and obtain fine-tuned features representations. Nonetheless, it is worth noting that linguistic systems used manually generated transcriptions. In the presence of potential ASR errors –which are commonly exacerbated in the case of atypical speech, such as AD speech–, acoustic systems may play a more relevant role. The impact of these errors could be an interesting analysis for future work, as well as the investigation of robust acoustic methods and models specially tailored to the elderly and AD speech characteristics.

7. References

[1] "World health organization. dementia: Fact sheet no. 362," September 2017, 2 (2017).

[2] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans, "Alzheimer disease in the United States (2010–2050) estimated using the 2010 census," *Neurology*, vol. 80, no. 19, pp. 1778–1783, 2013.

[3] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and P. Matthew, "World Alzheimer Report 2015 - The Global Impact of Dementia. An AnalysIs of Prevalence, Incidence, Cost and Trends," Alzheimer's Disease International, Tech. Rep., 2015.

[4] J. Reilly, J. Troche, and M. Grossman, "Language processing in dementia," *The handbook of Alzheimer's disease and other dementias*, pp. 336–368, 2011.

[5] D. Kempler, "Language changes in dementia of the Alzheimer type," *Dementia and communication*, pp. 98–114, 1995.

[6] D. Kempler and E. Zelinski, "Language in dementia and normal aging," *Dementia and normal aging*, pp. 331–365, 1994.

[7] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with Alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[8] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[9] D. N. Ripich and B. Y. Terrell, "Patterns of discourse cohesion and coherence in Alzheimer's disease," *Journal of Speech and Hearing Disorders*, vol. 53, no. 1, pp. 8–15, 1988.

[10] L. Hernández-Domínguez, S. Ratté, G. S. Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimers Dement (Amst)*, vol. 10, pp. 260–268, 2018.

[11] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech." *J Alzheimers Dis*, vol. 49, no. 2, pp. 407–422, 2016.

[12] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *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.

[13] 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*, 2019.

[14] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "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.

[15] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Computer Speech & Language*, vol. 53, pp. 65–79, 2019.

[16] J. Chen, J. Zhu, and J. Ye, "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech," *Proc. Interspeech 2019*, pp. 4085–4089, 2019.

[17] S. Zargarbashi and B. Babaali, "A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language," *arXiv preprint arXiv:1910.00330*, 2019.

[18] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data," *arXiv preprint arXiv:1803.11344*, 2018.

[19] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.

[20] 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.

[21] A. Pompili, A. Abad, D. M. de Matos, and I. P. Martins, "Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer's Disease," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 261–271, 2020.

[22] ——, "Topic coherence analysis for the classification of Alzheimer's disease." in *IberSPEECH*, 2018, pp. 281–285.

[23] H. Goodglass, E. Kaplan, and B. Barresi, *The Boston Diagnostic Aphasia Examination*, Baltimore: Lippincott, Williams & Wilkins, 2001.

[24] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[25] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[26] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification." in *Interspeech*, 2017, pp. 999–1003.

[27] 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," *Proc. Interspeech 2019*, pp. 2498–2502, 2019.

[28] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. Interspeech 2017*, 2017, pp. 1834–1838. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1363

[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[30] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken Language Recognition using X-vectors." in *Odyssey*, 2018, pp. 105–111.

[31] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Anton-Martin, M. A. Barbero-Alvarez, and L. A. Hernandez, "Modeling Obstructive Sleep Apnea voices using Deep Neural Network Embeddings and Domain-Adversarial Training," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[34] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[35] I. Annamoradnejad, "ColBERT: Using BERT Sentence Embedding for Humor Detection," *arXiv preeprint arXiv:2004.12765*, 2020.

[36] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[38] 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.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues

Shahla Farzana, Natalie Parde

Department of Computer Science
University of Illinois at Chicago
851 S. Morgan St., Chicago, IL 60607
{sfarza3, parde}@uic.edu

Abstract

The Mini Mental State Examination (MMSE) is a standardized cognitive health screening test. It is generally administered by trained clinicians, which may be time-consuming and costly. An intriguing and scalable alternative is to detect changes in cognitive function by automatically monitoring individuals' memory and language abilities from their conversational narratives. We work towards doing so by predicting clinical MMSE scores using verbal and non-verbal features extracted from the transcripts of 108 speech samples from the ADReSS Challenge dataset. We achieve a Root Mean Squared Error (RMSE) of 4.34, a percentage decrease of 29.3% over the existing performance benchmark. We also explore the performance impacts of acoustic versus linguistic, text-based features and find that linguistic features achieve lower RMSE scores, providing strong positive support for their inclusion in future MMSE score prediction models. Our best-performing model leverages a selection of verbal and non-verbal cues, demonstrating that MMSE score prediction is a rich problem that is best addressed using input from multiple perspectives.

Index Terms: spoken language processing, spoken language analysis, healthcare applications, dementia detection

1. Introduction

Scientific progress and improved healthcare standards in many areas of the world have resulted in older populations than ever before [1]. Although this is in many ways cause for celebration, it also introduces new challenges to administering effective clinical care. A growing elderly population creates an increased demand for a wide range of healthcare services, including cognitive assessment. Managing clinician burden and allowing medical professionals to allocate their time effectively is key to maximizing health outcomes and minimizing patient distress. One way to do this is by automating lower-risk tasks, such as routine cognitive assessment.

Cognitive assessment is often performed using straightforward, clinically validated tests such as the Mini Mental State Examination (MMSE) [2]. Clinicians administering the MMSE ask patients a series of questions in five different areas (orientation, registration, attention, memory, and language); their responses to these questions ultimately result in a score ranging from 0 (greatest cognitive decline) to 30 (no cognitive decline). Although simple to administer, the assessment can be burdensome, requiring the patient to travel to a clinical setting for in-person assessment. It may also be subject to biases from various demographic factors [3]. As an alternative to the structured, in-person MMSE, preliminary evidence suggests that automated methods can be used to predict MMSE scores from open-ended narrative descriptions [4]. The availability of easily-accessible, automated mechanisms could also enable assessment of individuals at more frequent, regular intervals, potentially facilitating quicker diagnosis of early-stage dementia [5].

We work toward this goal of simple, efficient dementia diagnosis by investigating a wide range of spoken language features for automated MMSE score prediction. It is well-known that dementia can influence spontaneous speech production, with declines in verbal fluency often manifesting with longer hesitations, lower speech rates, more frequent repetition, and other aphasic conditions [6, 7]. We design features that account for these discourse characteristics, in addition to incorporating promising linguistic features from prior work. Our findings suggest that a combination of verbal and non-verbal features results in strong predictive ability. Our contributions are as follows:

1. We propose a suite of features for MMSE score prediction, and run experiments to assess their utility for the task. We find that a blend of features drawn from multiple linguistic and discourse perspectives exhibits the strongest performance.
2. We extract features designed to encode properties of hesitation and verbal fluency, which are important biomarkers of Alzheimer's disease. Since identifying these subtle characteristics directly from audio files remains a challenging task [8, 9], we leverage the extensive set of annotations for non-verbal cues already present in the transcripts. To the best of our knowledge, the use of these features for MMSE score prediction is novel.
3. We compare the performance of acoustic and textual features for the task, finding that models trained only on text features outperform those trained only on acoustic features. This provides strong support for the inclusion of linguistic features in future models.
4. We analyze patterns in the features found to be most beneficial, finding that function words and discourse connectives offer high predictive value.

Our best-performing model outperforms the existing task benchmark by a wide margin (RMSE=4.34, a 29.3% decrease from the acoustic benchmark (RMSE=6.14) at the time of submission, and a 16.5% decrease from the linguistic benchmark (RMSE=5.20) added before the camera-ready deadline [4]).

2. Related Work

There is growing interest in automated dementia detection, although most work to date has focused on the binary task of dementia classification (wherein an individual is predicted to either have or not have dementia) [10, 11, 12, 13, 4] rather than the more nuanced problem of assigning continuous MMSE scores [14, 4]. Unlike most recent natural language processing tasks, which have migrated almost exclusively to using neural models with implicitly learned features, small dataset sizes and a

http://dx.doi.org/10.21437/Interspeech.2020-3085

strong interest in maintaining model interpretability have kept the problem space of automated dementia detection refreshingly diverse. Recent high-performing models have relied on a wide range of engineered features [10, 14, 11, 12, 15, 4], at the same time that others have explored neural solutions [16, 13].

Although we examine one neural solution for comparative purposes, our focus in this work is on identifying high-performing interpretable feature sets. Previously, others have explored both acoustic [11, 15, 4] and linguistic [10, 11, 12, 13] engineered features, primarily for dementia classification [14, 15, 4] rather than regression [4]. Acoustic features that have proved successful for the task include fundamental frequency [4], measures of vocal quality [4], Mel Frequency Cepstral Coefficients [11, 4], and pause- and duration-based features [15], among others. High-performing linguistic features have included verbal markers (e.g., indicators of repetition or backtracking) [10], syntax patterns [10, 11], lexical characteristics [10, 12], part-of-speech tags [11, 13], syntactic complexity [11], psycholinguistic traits [11, 13], vocabulary richness [11, 12], information content [11], repetitiveness [11], n-grams [12], and sentiment [13]. We draw inspiration from many of these prior approaches in selecting and designing features for our MMSE prediction models. Specifically, we make use of an expanded n-gram set, non-verbal speech and discourse markers via CHAT transcript [17] annotations, and measures of word familiarity, imageability, concreteness, sentiment, and typical age of word acquisition, as well as MFCC acoustic features.

3. Methods

We employ a set of automatically-extracted lexicosyntactic, psycholinguistic, discourse-based, and acoustic features for estimating continuous MMSE scores on a scale of 0 to 30. Although MMSE scores are often present in dementia detection datasets, the task is generally approached as a binary classification problem; its framing as a regression task is under-explored. We experiment with several machine learning techniques for representing relationships between our observed features and the underlying clinical scores. We explored this task in the context of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge.

3.1. Data

The ADReSS Challenge dataset is a subset of DementiaBank's Pitt Corpus [18]. The Pitt Corpus consists of anonymized recordings and transcripts of spoken picture descriptions elicited from participants who were shown the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [19]. In the recordings and transcripts, the interviewer asks the participant to describe what is in the picture, with no time constraints and relatively little structure (on occasion, the interviewer prods the participant for clarification or additional details). The audio from these conversations was manually transcribed, with discourse markers added for false starts, pauses, word repetition, phrase tracing, incomplete sentences, and other nonverbal cues, using the CHAT coding system [17]. For the ADReSS Challenge, the original speech recordings were also segmented into volume-normalized clips of at most ten seconds in length.

The dataset was divided by the task organizers into training and test sets. The training set contained 108 transcripts with an average conversation length (in terms of number of words uttered by the participant) of 98.5 (SD=55.37), and the test set contained 48 transcripts with an average conversation length of 93.38 (SD=56.20). The dataset (unlike the Pitt Corpus as a whole) was gender- and age-balanced across participants with and without dementia. Individual participant demographic information, cognitive status (Dementia or Control), and MMSE score were provided for all training samples; cognitive status and MMSE score were not provided for test samples. We preprocessed the transcripts to remove interviewer utterances, as well as numbers, punctuation, and unwanted symbols.

Table 1: *Token-level psycholinguistic and sentiment features.*

Feature	Description
Age of Acquisition	The age at which a particular word is usually learned.
Concreteness	A measure of a word's tangibility.
Familiarity	A measure of how often one might expect to encounter a word.
Imageability	A measure of how easily a word can be visualized.
Sentiment	A measure of a word's valence.

3.2. Features

We automatically extracted a variety of features from each transcript, described in more detail below.

3.2.1. Lexicosyntactic Features

We extracted n-grams for $n \in \{1, 2, 3\}$ from all training set samples, retaining only n-grams that appeared at least five times and at most 50 times across the training data and including coded non-verbal cues (e.g., *laugh*, *cough*, *breath intake*, or *sigh*). This resulted in a sparse feature vector for each utterance containing one dimension for each n-gram. Feature values were filled using TFIDF counts for a given transcript, computed as follows where TF is the term frequency within the transcript and DF is the number of documents containing the term:

$$TFIDF = TF \times \frac{1}{DF} \tag{1}$$

Each vector was L2-normalized with unit modulus. The final vocabulary size across all n-grams was 613.

3.2.2. Psycholinguistic Features

Psycholinguistic characteristics play a key role in verbal processing [20], and thus we suspected that they may have high utility for predicting MMSE scores. We considered four classic psycholinguistic properties (age of acquisition, concreteness, familiarity, and imageability), as well as sentiment scores. These features (five total, described further in Table 1) were all extracted from third-party lexical resources as token-level scores, which we then averaged across all tokens in a given transcript. Sentiment scores were obtained using NLTK's SentimentAnalyzer library,[1] and psycholinguistic scores were obtained from an open source repository[2] containing scores from multiple aspects of the MRC Psycholinguistic Database [21].

[1] `https://www.nltk.org/api/nltk.sentiment.html`

[2] `https://github.com/vmasrani/dementia\_classifier`

3.2.3. Discourse-Based Features

To model global discourse patterns across the entire transcript, we extracted an array of count-based features for discourse tags. These features include CHAT transcript [17] markers for different pause types (including filled pauses containing, e.g., *uh* or *umm*), word repetition, retracing (restarting the same phrase or segment), and incomplete utterances. Our full list of discourse-based features included: short_pause_count, long_pause_count, very_long_pause_count, word_repetition_count, retracing_count, filled_pause_count, and incomplete_utterance_count. We normalized these counts by the number of words uttered in the conversation. We also examined both word count and utterance count as features, ultimately dropping utterance count due to its high correlation ($r > 0.5$) with the former, but retaining word count, resulting in a total set of eight discourse features.

3.2.4. Acoustic Features

Finally, we extracted acoustic features due to their success in prior work on dementia detection [11, 22, 15, 4] and MMSE score prediction [4]. Specifically, we computed Mel Frequency Cepstral Coefficients (MFCCs) and extracted the first 14 MFCCs for each speech segment. We identified mean values, variance, skewness, and kurtosis for these features, and then computed the same for velocity and acceleration. This resulted in a total of 171 audio features for each segment.

3.3. Model

We designed separate models for our textual (lexicosyntactic, psycholinguistic, and discourse-based) and acoustic features due to underlying differences in how the data was handled. Since we extracted our acoustic features from local audio segments (maximum duration 10 seconds), we employed a segment-based model similar to that seen in the existing performance benchmark [4]. The model predicted individual MMSE scores for each discrete segment, and these scores were then averaged across an entire transcript to produce a transcript-level MMSE score. We employed a transcript-level model for our textual features since they were extracted from the transcript as a whole. We experimented with two high-performing statistical regression algorithms: Support Vector Regression (SVR) with a polynomial kernel, regularization parameter $C = 100$, and kernel coefficient γ="auto"; and Gaussian Process Regression (GP) with a squared exponential kernel, $\alpha = 0.1$, and optimizer restarts set to 10. All other parameters for the respective algorithms were kept at their default values.

To empirically validate the utility of our engineered features relative to neural alternatives, we also experimented with a fine-tuned DistilBERT sequence classification model [23] for the task. We illustrate the architecture for this model in Figure 1. The pre-trained DistilBERT tokenizer processes unseen tokens (e.g., discourse tags in our transcripts) as subword units, allowing it to make use of vocabulary not present in its original corpus. Input is thus tokenized and then encoded, and the resulting hidden representation is subsequently passed to a final fully-connected network, which applies linear transformations to the data to ultimately predict a single output neuron representing the predicted MMSE value for the specific patient.

4. Evaluation

We selected a diverse set of five models for entry to the ADReSS Challenge:

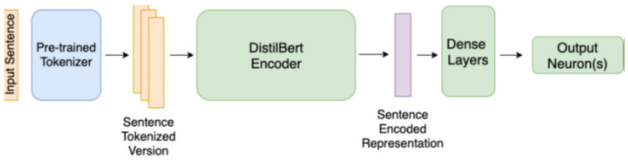


Figure 1: *Model Architecture for DistilBERT.*

- **ALL:** All textual features described in Section 3.2.
- **N-GRAM:** All lexicosyntactic features.
- **SELECTED-FEATURE:** A selection of the 90 highest-performing features from the training corpus. To obtain this feature subset, we employed a Random Forest regression model with 100 trees and selected features based on their mean decrease impurity (MDI), where impurity was measured as variance. We retained only features having MDI values exceeding a predefined threshold (10^{-3}). We show the top ten most important features measured using this process in Table 4.
- **DISTILBERT:** The DistilBERT model described in Section 3.3.
- **ACOUSTIC-ALL:** All acoustic features.

Although not entered into the ADReSS Challenge, we also experimented with a selection of the highest-performing acoustic features (ACOUSTIC-SELECTED), using the same feature selection technique as applied to SELECTED-FEATURE. We additionally ran some experiments using a late fusion neural network model to map acoustic and textual features to the same hidden space,[3] but the model performance was significantly lower than alternatives in the leave-one-out (LOO) experiment (RMSE>10). We report both our LOO cross-validation results on the training corpus, and our ADReSS Challenge results on the test data. We report both root mean squared error (RMSE) and R-squared values for the LOO setting, and RMSE for the results on the test data.

4.1. Results

We present the results from our LOO cross-validation experiment in Table 2, and our ADReSS Challenge results on the test set in Table 3. Our LOO experiment included both SVR and GP versions of each model; since SVR outperformed GP in more cases and we were limited to a batch of five results submissions, we submitted only SVR models (along with our DistilBERT alternative) to the ADReSS Challenge. Our best-performing model in the LOO experiment was ALL using an SVR classifier, achieving an RMSE of 4.97. Interestingly, ALL and ACOUSTIC-ALL exceeded the performance of SELECTED-FEATURE and ACOUSTIC-SELECTED, respectively, in the LOO experiments. Although ACOUSTIC-SELECTED was not entered in the ADReSS Challenge, this advantage did not persist for ALL vs. SELECTED-FEATURE on the test data. The R-squared values in Table 2 provide insight into the variance from the regression line; $R^2 = 0.52$ is considered moderate [25]. Our highest-performing model on the test set (Table

[3]Specifically, we encoded words using 300-dimensional English GloVe embeddings [24] and passed them to a bidirectional LSTM (Bi-LSTM) layer. We fed the acoustic features for each segment to a separate LSTM layer, and then we concatenated the resulting hidden representations of the Bi-LSTM and LSTM layers. We merged this concatenated vector with the vector of discourse-based features, and fed the merged vector into a feedforward layer with an output linear activation.

Table 2: *LOO results, formatted as RMSE (R^2).*

Features	SVR	GP
ALL	**4.97 (0.52)**	6.43 (-0.001)
N-GRAM	5.00 (0.514)	5.60 (0.225)
SELECTED-FEATURE	5.49 (0.415)	5.31 (0.451)
ACOUSTIC-ALL	6.59 (-0.093)	6.71 (-0.135)
ACOUSTIC-SELECTED	7.67 (-0.481)	7.31 (-0.271)

Table 3: *Test set results.*

Features	RMSE
ALL	4.87
NGRAM	4.61
SELECTED-FEATURE	**4.34**
DISTILBERT	4.63
ACOUSTIC-ALL	6.42

3) employed the SELECTED-FEATURE subset with SVR. This model (RMSE=4.34) outperformed the best-performing baseline model on the test set (RMSE=5.20 [4]) by 16.5%.

4.2. Analysis

We analyzed trends in RMSE scores across binned MMSE score groups to identify weaknesses in our best model and areas for potential improvement, and present our findings in Figure 2. We found that in general our model's predictive power was best for high MMSE scores, which is likely an artifact of the training set distribution—although samples in the ADReSS Challenge dataset are balanced across age and gender, they are not evenly distributed across the MMSE score continuum.

We also sorted the features in SELECTED-FEATURE in descending order based on their MDI importance score to analyze the strongest identified patterns, and present the top ten features in Table 4. Interestingly, we found many non-content function words and discourse connectives in this list, along with some discourse-based count features. In general, individuals with higher MMSE scores created longer descriptions of the picture and used more content words and complex phrases (e.g. *fall, cookie jar and*), whereas those with lower MMSE scores used shorter descriptions and more pauses and filler words. This provides evidence that verbal disfluency markers are important indicators of cognitive status, and also supports our hypothesis that a wide range of features can be productively leveraged in tandem for this task.

5. Discussion and Conclusion

Overall, we found text-based features to be more informative than acoustic features for the MMSE score prediction task. We speculate that this may be an important distinction between this and the dementia classification task, for which acoustic features have achieved considerable success [11, 15]. Our source code

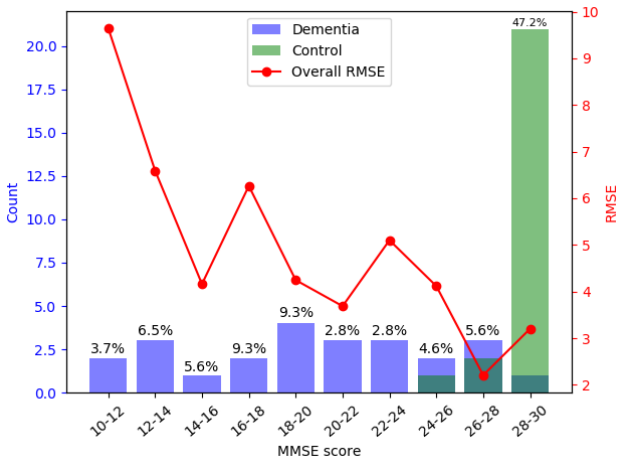


Figure 2: *Binned MMSE scores and frequency counts, with corresponding average RMSE per bin. Frequency counts (left y-axis and associated histogram bars) and RMSE (right y-axis and associated line graph) are for test instances, whereas percentages above histogram bars indicate the corresponding training set frequency for the same MMSE bins.*

Table 4: *Top 10 features based on MDI importance.*

Features	Importance
this	0.284
here	0.050
word_count	0.044
fall	0.037
well	0.034
laughs (non-verbal)	0.034
short_pause_count	0.021
in the	0.015
cookie jar and	0.014
it uh	0.013

is publicly available.[4] Further investigation into more informative features (e.g., acoustic disfluency markers) from the normalized speech signal could potentially transfer insights from our text-based features to high-performing acoustic analogues. Likewise, we are interested in leveraging the segment-based model with the text transcripts (casting utterances as segments). Finally, while automated MMSE score prediction may make testing more accessible, reliable, and resource-effective, future work could additionally explore more precise measures such as the Montreal Cognitive Assessment (MoCA) or the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [26, 27], which have higher sensitivity than the MMSE to subtle changes in cognitive decline.

6. Acknowledgements

We thank Flavio Di Palo for contributing the DISTILBERT model, and the anonymous reviewers for their helpful feedback.

[4] https://github.com/treena908/MMSE-Prediction

7. References

[1] V. Fuster, "Changing demographics," *J. Am. Coll. Cardiol*, vol. 69, no. 24, pp. 3002–3005, 2017. [Online]. Available: http://www.onlinejacc.org/content/69/24/3002

[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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022395675900266

[3] R. N. Jones and J. J. Gallo, "Education and Sex Differences in the Mini-Mental State Examination: Effects of Differential Item Functioning," *The Journals of Gerontology: Series B*, vol. 57, no. 6, pp. P548–P558, 11 2002. [Online]. Available: https://doi.org/10.1093/geronb/57.6.P548

[4] 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

[5] S. Farzana, M. Valizadeh, and N. Parde, "Modeling dialogue in conversational cognitive health screening interviews," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. Marseilles, France: European Language Resources Association, May 11-16, 2020 2020.

[6] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010, pMID: 20380247. [Online]. Available: https://doi.org/10.3109/17549500903137256

[7] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] K. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proc. of the 4th Workshop on Speech and Language Processing for Assistive Technologies*. Grenoble, France: Assoc. for Computational Linguistics, 2013, pp. 47–54. [Online]. Available: https://www.aclweb.org/anthology/W13-3909

[10] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 78–87. [Online]. Available: https://www.aclweb.org/anthology/W14-3210

[11] 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.

[12] D. Weissenbacher, T. A. Johnson, L. Wojtulewicz, A. Dueck, D. Locke, R. Caselli, and G. Gonzalez, "Automatic prediction of linguistic decline in writings of subjects with degenerative dementia," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1198–1207. [Online]. Available: https://www.aclweb.org/anthology/N16-1143

[13] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 302–308. [Online]. Available: https://www.aclweb.org/anthology/P19-2042

[14] M. Yancheva, K. 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*. Dresden, Germany: Association for Computational Linguistics, Sep. 2015, pp. 134–139. [Online]. Available: https://www.aclweb.org/anthology/W15-5123

[15] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features," *Proceedings of Interspeech 2018*, pp. 2324–2328, 2018.

[16] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 701–707. [Online]. Available: https://www.aclweb.org/anthology/N18-2110

[17] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Psychology Press, 2000, vol. 1.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "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, 06 1994. [Online]. Available: https://doi.org/10.1001/archneur.1994.00540180063015

[19] C. Roth, *Boston Diagnostic Aphasia Examination*. New York, NY: Springer New York, 2011, pp. 428–430. [Online]. Available: https://doi.org/10.1007/978-0-387-79948-3_868

[20] T. Salsbury, S. A. Crossley, and D. S. McNamara, "Psycholinguistic word information in second language oral discourse," *Second Language Research*, vol. 27, no. 3, pp. 343–360, 2011. [Online]. Available: https://doi.org/10.1177/0267658310395851

[21] M. Coltheart, "The mrc psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, no. 4, pp. 497–505, 1981.

[22] 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*, ser. ICBRA 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 57–61. [Online]. Available: https://doi.org/10.1145/3175587.3175589

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[25] H. Jörg, R. C. M., and S. R. R., *The use of partial least squares path modeling in international marketing*, ser. Advances in International Marketing. Emerald Group Publishing Limited, Jan 2009, vol. 20, pp. 277–319. [Online]. Available: https://doi.org/10.1108/S1474-7979(2009)0000020014

[26] 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," *Jour. of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53221.x

[27] A. R. Loughan, S. E. Braun, and A. Lanoye, "Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary utility in adult neuro-oncology," *Neuro-Oncology Practice*, vol. 6, no. 4, pp. 289–296, 12 2018. [Online]. Available: https://doi.org/10.1093/nop/npy050

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity

Utkarsh Sarawgi[*], *Wazeer Zulfikar*[*], *Nouran Soliman, Pattie Maes*

Massachusetts Institute of Technology
{utkarshs, wazeer, nouran, pattie} @mit.edu

Abstract

Alzheimer's disease is estimated to affect around 50 million people worldwide and is rising rapidly, with a global economic burden of nearly a trillion dollars. This calls for scalable, cost-effective, and robust methods for detection of Alzheimer's dementia (AD). We present a novel architecture that leverages acoustic, cognitive, and linguistic features to form a multimodal ensemble system. It uses specialized artificial neural networks with temporal characteristics to detect AD and its severity, which is reflected through Mini-Mental State Exam (MMSE) scores. We first evaluate it on the ADReSS challenge dataset, which is a subject-independent and balanced dataset matched for age and gender to mitigate biases, and is available through DementiaBank. Our system achieves state-of-the-art test accuracy, precision, recall, and F1-score of 83.3% each for AD classification, and state-of-the-art test root mean squared error (RMSE) of 4.60 for MMSE score regression. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy of 88.0% when evaluated on the full benchmark DementiaBank Pitt database. Our work highlights the applicability and transferability of spontaneous speech to produce a robust inductive transfer learning model, and demonstrates generalizability through a task-agnostic feature-space. The source code is available at `https://github.com/wazeerzulfikar/alzheimers-dementia`

Index Terms: Alzheimer's Dementia Detection, Affective Computing, Human-Computer Interaction, Computational Paralinguistics, Machine Learning, Speech Processing

1. Introduction

Alzheimer's disease is a progressive disorder that causes brain cells to degenerate and is the most common cause of dementia worldwide. It mainly causes cognitive and behavioural deterioration of the patients [1] which is reflected through memory loss, language impairment [2], and a decreased ability to express their needs. This in turn affects their quality of life, prognosis, and social relationships. Consequently, it has been imposing increased health risks [3] and a significant financial burden to patients, caregivers, families, and healthcare institutions [4]. The number of people with dementia worldwide in 2015 was estimated at 47.47 million, and reaching 135.46 million in 2050 [5]. At the time of writing this paper, someone in the U.S. develops Alzheimer's disease every 66 seconds, and by 2050 it is projected to be 33 seconds [6]. According to the World Health Organization, the global economic burden is nearly a trillion dollars which amounts to 1.1% of the global GDP. [7], with 63% of people with dementia living in low- and middle-income countries [8]. In this work, we aim to take a significant step towards more reliable, cost-effective, scalable, and non-invasive technologies to detect the onset of Alzheimer's dementia (AD) and predict the Mini-Mental State Exam [9] scores to estimate the severity of it.

Dementia can be strongly characterized by cognitive degeneration leading to language impairment which primarily occurs due to decline in semantic and pragmatic levels of language processing [10]. It has been widely reported that AD can be more sensitively detected with the help of a linguistic analysis than with other cognitive examinations [11] and also long before the diagnosis is medically confirmed [12]. The temporal characteristics of spontaneous speech, such as speech tempo, number of pauses in speech, and their length are sensitive detectors of the early stage of the disease [13, 14, 15, 16, 17]. Given the relative ease of collecting balanced and representative data of spontaneous speech and their corresponding transcriptions, they can be utilized in early and robust predictions for the onset of AD.

Consequently, our research work:

1. Presents a novel architecture comprising of domain-specific feature engineering and artificial neural networks for Alzheimer's Dementia (AD) detection and its severity through classification and MMSE score regression (Section 3).
2. Evaluates the system in a subject-independent setting with a carefully curated balanced and stratified dataset matched for age and gender, to help minimize common biases in the tasks (Section 3.1).
3. Achieves state-of-the-art test accuracy, precision, recall, and F1-score for AD classification, and state-of-the-art test RMSE for MMSE score predictions on the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) dataset. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy when evaluated on the full benchmark DementiaBank Pitt database (Sections 4 and 5).
4. Spans a multimodal feature space to increase generalizability and robustness, and uses ensemble mechanisms to leverage individual feature sets and model performances.
5. Reflects upon the transferability and interdependence of the two tasks of AD classification and MMSE regression.

2. Related work

Many current AD detection studies use medical imaging [18, 19, 20] with deep neural networks and random forests. Several studies claim that AD can be sensitively detected in early stages by doing linguistic analysis which leverages speech and language features to train machine learning models for the detection of AD [13, 14, 15, 16, 17, 21].

In study [22], machine learning methods based on image description were used reaching an accuracy of 75% on a limited

[*]Equal Contribution

http://dx.doi.org/10.21437/Interspeech.2020-3137

number of subjects enrolled in a longitudinal study. Study [23] used logistic regression trained with spectrogram features extracted from audio files reaching accuracy of 83.3% and 84.4% on VBSD and Dem@Care datasets respectively. Data used in each of the above works are limited to around 32 to 36 subjects and highly imbalanced between the classes and across age and gender. In study [14], different traditional classification algorithms like logistic regression, SVM, and more were used to learn speech parameters from dialogues in Carolina Conversations Collection. The best of their solutions reached 86.5% leave-one-out cross-validation (LOOCV) accuracy with 38 subjects. Works based on data extracted from DementiaBank have reported scores of around 0.87, 0.85, 0.82, 0.80, 0.79, 0.64, and 0.62 [24, 25, 13, 26, 27, 28, 29] for AD classification. Study [30] used speech related features to get a mean absolute error (MAE) of 3.83 for MMSE scores with longitudinal data derived from DementiaBank. While a number of works have proposed speech and language based approaches to AD recognition through speech, their studies have used different, often unbalanced and acoustically varied data sets, thereby introducing bias and hindering generalization, reproducibility and comparability of the proposed approaches.

3. Methods and materials

3.1. Dataset

The DementiaBank Pitt database [31] consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [32]. The database consists of multiple samples per subject corresponding to multiple visits. The full database contains 242 speech samples from 99 control healthy subjects and 255 speech samples from 168 AD subjects. The dataset also provides Mini-Mental Status Examination (MMSE) scores, ranging from 0 to 30, of the subjects, which offers a way to quantify cognitive function and screen for cognitive loss by testing the individuals' orientation, attention, calculation, recall, language and motor skills [9]. A 10-fold cross-validation was used on this database for fair comparison with previously reported results.

The ADReSS Challenge Dataset [29] is a balanced subset consisting of 156 speech samples, each from a unique subject, matched for age and gender and evenly spread across the two classes, AD and non-AD. A stratified train-test split of around 70-30 (108 and 48 subjects) for this dataset was provided by the challenge. The test set was held out for all experimentation until final evaluation. Any cross-validation mentioned in the paper refers to cross-validation using the train split. Normalized speech segments are also provided, but we only use full audio samples. The MMSE scores provided are used as labels for the regression task.

We first evaluate on the balanced ADReSS dataset and then extend the evaluation to the full DementiaBank Pitt database.

3.2. Feature engineering

People with dementia show symptoms of cognitive decline, impairment in memory, communication, and thinking [17]. To include such domain knowledge and context, our system extracts cognitive and acoustic features using three different strategies, which are then prepared and fed into their respective neural models. Similarly extracted features have been repeatedly used to propose speech recognition based solutions for automated detection of mild cognitive impairment from spontaneous speech [33, 17]. The following features were extracted upon exploring the data to find the most descriptive set of correlated features for detecting AD and its severity:

• *Disfluency:* A set of 11 distinct and carefully curated features from the transcripts, like word rate, intervention rate, and different kinds of pause rates reflecting upon speech impediments like slurring and stuttering. These are normalized by the respective audio lengths and scaled thereafter.

• *Acoustic:* The ComParE 2013 feature set [34] was extracted from the audio samples using the open-sourced openSMILE v2.1 toolkit, widely used for affect analyses in speech [35]. This provides a total of 6,373 features that include energy, MFCC, and voicing related low-level descriptors (LLDs), and other statistical functionals. This feature set encodes changes in speech of a person and has been used as an important noninvasive marker for AD detection [36, 29]. Our system standardizes this set of features using z-score normalization, and uses principal component analysis (PCA) to project the 6,373 features onto a low-dimensional space of 21 orthogonal features with highest variance. The number of orthogonal features was selected by analyzing the percentage of variance explained by each of the components.

• *Interventions:* Cognitive features reflect upon potential loss of train of thoughts and context. Our system extracts the sequence of speakers from the transcripts, categorizing it as subject or the interviewer. To accommodate for the variable length of these sequences, they are padded or truncated to length of 32 steps, found upon analyses and tuning of sequence lengths.

We evaluated each of these features individually and in a combined fashion to highlight the different configurations and compare their performances.

3.3. Model architecture and training

Figure 1 - (1), (2), and (3) illustrate the architecture of the disfluency, acoustic, and interventions models respectively. The disfluency model is a multi-layer perceptron (MLP) that projects the 11-feature input to a higher dimensional space for better separability of the binary classes. The acoustic model is an MLP with a single hidden layer that adds non-linearity and regularizes the PCA decomposed feature space. The interventions model uses a recurrent architecture to learn the temporal relations from the sequence of interventions. These models were trained with corresponding inputs obtained upon feature engineering (Section 3.2), and one-hot encoded binary class labels.

To leverage the features learnt from classification for regression, transfer learning was done on the trained classification models. The regression module, as shown in Figure 1 - (4) replaced the terminal output layer in the models and the remaining original layers were frozen. The resultant models were then trained with MMSE scores as labels.

A 5-fold cross-validation setting was adopted for evaluation. The models were also evaluated in a leave-one-out cross validation (LOOCV) setting, which in the case of ADReSS dataset is equivalent to leave-one-subject-out cross validation (LOSO) since each datapoint is an independent subject. Each training run used a batch size of 8; and Adam optimizer with a learning rate of 0.01 to minimize categorical cross-entropy loss for classification, and a learning rate of 0.001 to minimize mean squared error loss for regression. The best models were saved by monitoring the validation loss in each fold.

To leverage all sets of features and models together, a parallel ensemble was performed using the outputs of the three models for each of the two tasks independently. We experimented

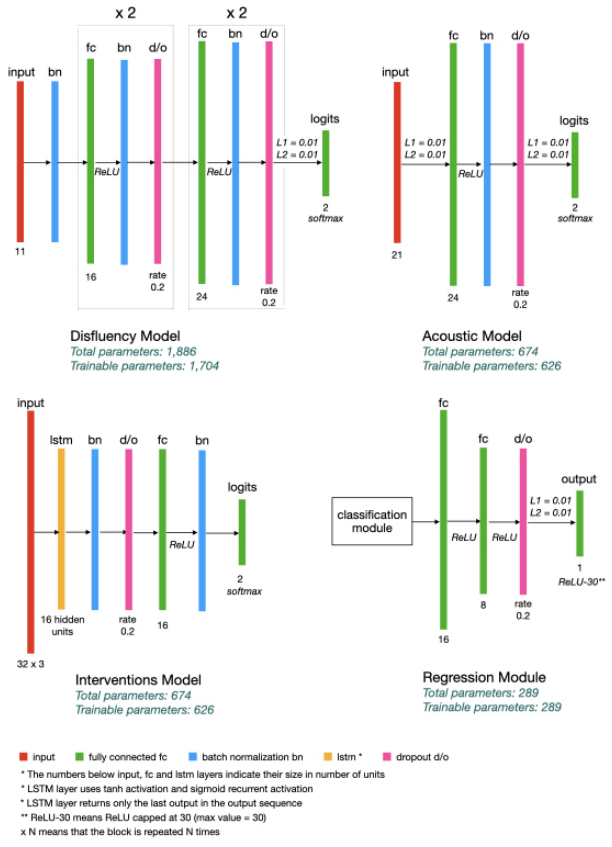


Figure 1: *Architecture of (1) Disfluency, (2) Acoustic, (3) Interventions models, and (4) Regression module.*

with three kinds of ensemble modules for classification:

• *Hard:* A majority vote was taken between the predictions of the three individual models.

• *Soft:* To leverage the confidence of the predictions, a weighted sum of the class probabilities was computed for final decision. The weight used was $1/N$ where N is the total number of models.

• *Learnt:* Instead of weighing the confidence of all the models equally as in soft voting above, we used a logistic regression to learn the weights. A logistic regression voter was trained using class probabilities as inputs.

For regression, the predictions of all the individual models were averaged by the ensemble module.

4. Results

The results of the experiments were recorded using a combination of accuracy, precision, recall and F1-score for classification, and root mean squared error (RMSE) for regression.

4.1. ADReSS Challenge dataset

Table 1 shows the 5-fold cross-validation results for the classification task. The individual features achieved competitive performance, although the acoustic model slightly overfits while the interventions model marginally underfits on the data. The ensemble model counteracted these and achieved an increased 5-fold mean training as well as validation accuracy with comparable variance. The low variance generally observed across all runs signifies high model stability across folds which is essential in small datasets. Similar observations can be seen on the regression task in Table 1, where the ensemble model reduced the train and validation mean RMSE as well as the variance. This is consistent with the intuition behind using transfer learning using the trained classification models through the addition of a regression module.

Table 1: *5-fold cross validation results of the models. Accuracy measures the AD classification performance while RMSE measures the MMSE score regression performance over all 5 folds. Ensemble in this table refers to hard ensemble for classification and the regression ensemble for regression.*

Model	Split	Accuracy	RMSE
Disfluency	Train	0.87 ± 0.08	4.37 ± 0.40
	Val	0.89 ± 0.05	4.87 ± 0.78
Acoustic	Train	0.89 ± 0.03	4.40 ± 0.64
	Val	0.83 ± 0.07	5.63 ± 1.15
Interventions	Train	0.82 ± 0.06	5.05 ± 0.56
	Val	0.89 ± 0.04	4.70 ± 0.96
Ensemble	Train	$\mathbf{0.91 \pm 0.04}$	$\mathbf{3.65 \pm 0.38}$
	Val	$\mathbf{0.92 \pm 0.06}$	$\mathbf{4.26 \pm 0.75}$

Table 2: *5-fold cross-validation accuracies of different ensemble mechanisms for AD classification.*

Ensemble Type	Split	Accuracy
Hard	Train	0.91 ± 0.04
	Val	$\mathbf{0.92 \pm 0.06}$
Soft	Train	0.86 ± 0.04
	Val	0.86 ± 0.04
Learnt	Train	$\mathbf{0.95 \pm 0.03}$
	Val	0.81 ± 0.08

The improvement in performance upon ensembling the three models as compared to the individual models further reflects upon the significance of leveraging acoustic and cognitive features together from multimodal speech and text inputs.

Table 2 shows the 5-fold cross validation results of different parallel ensemble techniques, discussed in Section 3.3, for the classifiation task. The learnt ensemble showed signs of overfitting due to the extra trainable parameters in the model. The soft and hard ensemble helped counter this. However, the hard ensemble proved to be the most competitive by improving training and validation accuracies along with a strong degree of generalization across folds.

Figure 2 shows the receiver operating characteristic (ROC) curve for the individual models on the classification task. The ROC is cumulatively calculated over the validation splits of all 5 folds of cross-validation.

We compare our results with the currently available baseline performance results on this dataset [29]. Amongst our models, the best performing model, the hard ensemble classification model and the ensemble regression model, considerably improved all the metrics on the LOSO as well as the held-out test set on AD classification and regression, as can be seen in Table 3 and Table 4 respectively.

The confusion matrices in Figure 3 provide further insights into the predictions of the hard ensemble classification model that has been compared with the baseline in Table 3.

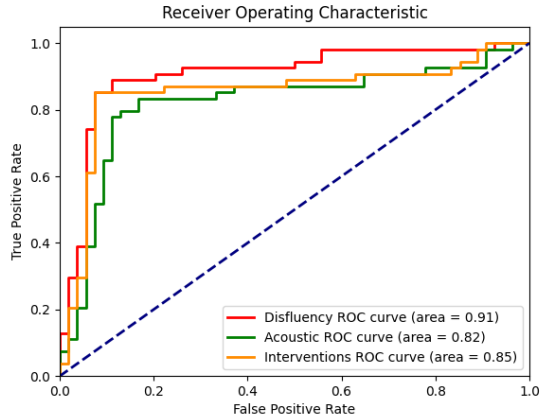


Figure 2: *Receiver Operating Characteristic for Disfluency, Acoustic, and Interventions models, cumulatively calculated over validation splits of all the folds of 5-fold cross-validation.*

Table 3: *Baseline comparison of the AD classification. Our test results below are corresponding to the hard ensemble model.*

	Model	Accuracy	Precision	Recall	F1-Score
LOSO	Luz et al. [29]	0.77	0.77	0.76	0.77
	Ensemble (*ours*)	**0.99**	**0.99**	**1.00**	**0.99**
TEST	Luz et al. [29]	0.75	0.83	0.62	0.71
	Ensemble (*ours*)	**0.83**	**0.83**	**0.83**	**0.83**

Table 4: *Baseline comparison of the MMSE score regression. Our test results are corresponding to the regression ensemble.*

	Model	RMSE
LOSO	Luz et al. [29]	4.38
	Ensemble (*ours*)	**0.82**
TEST	Luz et al. [29]	5.20
	Ensemble (*ours*)	**4.60**

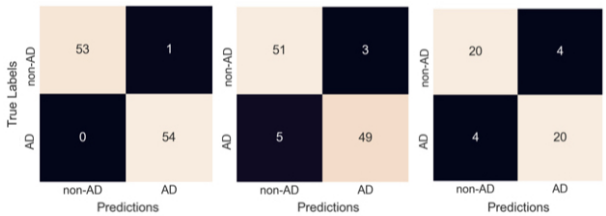


Figure 3: *Confusion matrices for the hard ensemble classification model (1) cumulatively calculated over the validation splits of all the folds of LOOCV and (2) 5-fold cross-validation, and (3) calculated on the held out test set.*

4.2. DementiaBank Pitt database

The same AD classification models were retrained on the DementiaBank Pitt database and a 10-fold cross-validation was performed for fair comparison with previously reported results. To the best of our knowledge, our hard ensemble model achieves state-of-the-art 0.88 $\pm$ 0.04 accuracy, also showing minimal variance across the folds (Table 5).

Table 5: *Comparison of the AD classification on DementiaBank Pitt. All are 10-fold cross-validation results. Our results below are corresponding to the hard ensemble model.*

Model	Accuracy	Precision	Recall	F1-Score
Fraser et al. [13]	0.82	-	-	-
Masrani [25]	0.85	-	-	0.85
Kong et al. [24]	0.87	0.86	**0.91**	**0.88**
Ensemble (*ours*)	**0.88**	**0.92**	0.82	**0.88**

5. Discussion and Future Work

There has been substantial work using spontaneous speech samples and manual transcriptions present in the DementiaBank dataset [31]. Some of the highest reported scores for AD classification are 0.87, 0.85, 0.82, 0.80, 0.79, 0.64, and 0.63 [24, 25, 13, 26, 27, 28, 29]. Many of these previous results were obtained on datasets with variable subject dependencies. In such datasets, a data point corresponds to a session and there can exist multiple sessions per subject. Given the subject independent setting in ADReSS dataset, our LOSO method clearly distinguishes the left-out test subject. Hence, the near perfect LOSO results on classification and regression (Tables 3 and 4) demonstrate that every subject individually can be correctly evaluated with the engineered features. Furthermore, almost all previous results are reported using cross-validation, whereas our work is evaluated on a designated held-out test set as well. This helps overcome 'validation overfitting' which is prone in small dataset settings.

Study [30] used speech related features to obtain a cross-validated mean absolute error (MAE) of 3.83 for MMSE scores with data derived from DementiaBank. Our ensemble regression model recorded a cross-validated MAE of 3.01 on ADReSS dataset.

Through considerable improvements in both the AD classification and MMSE score regression by employing an ensemble of independent models extracting acoustic and cognitive features, our work reveals the potential of multimodal analysis and its applicability to a age and gender balanced subject-independent dataset. Future work would include incorporating automated transcription of speech samples in our system. The continuous range of the MMSE scores can provide more insights into progression of dementia. This can further be leveraged for risk stratification and analyzing potential causal relationships modelling AD with its symptoms and markers, through a longitudinal dataset.

6. Conclusion

We present a novel architecture that uses domain knowledge for inductive transfer learning for AD classification and MMSE score regression. Our work achieves state-of-the-art accuracy, precision, recall, and F1-score of 83.3% each for AD classification, and state-of-the-art RMSE of 4.60 for MMSE predictions on the designated held-out test set of the ADReSS challenge. To the best of our knowledge, the system further achieves state-of-the-art AD classification accuracy of 88.0% when evaluated on the full benchmark DementiaBank Pitt database. Our system spans a multimodal feature space to increase generalization and robustness. We aim to extend our work by adding automated transcription, further textual analysis, and personalized context through longitudinal data.

7. References

[1] J. G. Molinuevo, "Role of biomarkers in the early diagnosis of alzheimer's disease," *Revista espanola de geriatria y gerontologia*, vol. 46, pp. 39–41, 2011.

[2] L. M. V. ESCOBAR and N. P. AFANADOR, "Calidad de vida del cuidador familiar y dependencia del paciente con alzheimer," *Avances en Enfermería*, vol. 28, no. 1, pp. 116–128, 2010.

[3] R. Schulz and S. R. Beach, "Caregiving as a risk factor for mortality: the caregiver health effects study," *Jama*, vol. 282, no. 23, pp. 2215–2219, 1999.

[4] J. M. Atance, A. I. Yusta, and B. G. Grupeli, "Costs study in alzheimer's disease," *Revista clinica espanola*, vol. 204, no. 2, pp. 64–69, 2004.

[5] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis," *Alzheimer's & dementia*, vol. 9, no. 1, pp. 63–75, 2013.

[6] A. Association *et al.*, "2016 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 12, no. 4, pp. 459–509, 2016.

[7] W. H. Organization *et al.*, "The top 10 causes of death. fact sheet no. 310. 2017."

[8] ——, "The epidemiology and impact of dementia. current state and future trends. geneva, switz: World health organization; 2015."

[9] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[10] S. H. Ferris and M. Farlow, "Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors," *Clinical interventions in aging*, vol. 8, p. 1007, 2013.

[11] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[12] M. Mesulam, A. Wicklund, N. Johnson, E. Rogalski, G. C. Léger, A. Rademaker, S. Weintraub, and E. H. Bigio, "Alzheimer and frontotemporal pathology in subsets of primary progressive aphasia," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 63, no. 6, pp. 709–719, 2008.

[13] 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.

[14] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *arXiv preprint arXiv:1811.09919*, 2018.

[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] 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*, 2019.

[17] 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*, p. 113213, 2020.

[18] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[19] A. Ortiz, F. Lozano, J. M. Gorriz, J. Ramirez, F. J. Martinez Murcia, A. D. N. Initiative *et al.*, "Discriminative sparse features for alzheimer's disease diagnosis using multimodal image data," *Current Alzheimer Research*, vol. 15, no. 1, pp. 67–79, 2018.

[20] S. Sarraf and G. Tofighi, "Deep learning-based pipeline to recognize alzheimer's disease using fmri data," in *2016 Future Technologies Conference (FTC)*. IEEE, 2016, pp. 816–820.

[21] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," *arXiv preprint arXiv:1906.05483*, 2019.

[22] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, no. s3, pp. S3–S17, 2014.

[23] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying alzheimer's disease," *Simulation Modelling Practice and Theory*, vol. 99, p. 102023, 2020.

[24] W. Kong, H. Jang, G. Carenini, and T. Field, "A neural model for predicting dementia from language," in *Machine Learning for Healthcare Conference*, 2019, pp. 270–286.

[25] V. Masrani, "Detecting dementia from written and spoken language," Ph.D. dissertation, University of British Columbia, 2018.

[26] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[27] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 260–268, 2018.

[28] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[29] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[30] 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.

[31] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "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.

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

[33] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[34] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[36] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, M. Faundez-Zanuy, M. Ecay-Torres, C. M. Travieso, A. Ezeiza, A. Estanga *et al.*, "Alzheimer disease diagnosis based on automatic spontaneous speech analysis," 2012.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Exploiting Multi-Modal Features From Pre-trained Networks for Alzheimer's Dementia Recognition

Junghyun Koo[1], *Jie Hwan Lee*[1], *Jaewoo Pyo*[2], *Yujin Jo*[3], *Kyogu Lee*[1]

[1]Music & Audio Research Group (MARG), Seoul National University
[2]Electrical and Computer Engineering, Seoul National University
[3]College of Liberal Studies, Seoul National University
{dg22302, wiswisbus, jwpyo, tera-yujin, kglee}@snu.ac.kr

Abstract

Collecting and accessing a large amount of medical data is very time-consuming and laborious, not only because it is difficult to find specific patients but also because it is required to resolve the confidentiality of a patient's medical records. On the other hand, there are deep learning models, trained on easily collectible, large scale datasets such as Youtube or Wikipedia, offering useful representations. It could therefore be very advantageous to utilize the features from these pre-trained networks for handling a small amount of data at hand. In this work, we exploit various multi-modal features extracted from pre-trained networks to recognize Alzheimer's Dementia using a neural network, with a small dataset provided by the ADReSS Challenge at INTERSPEECH 2020. The challenge regards to discern patients suspicious of Alzheimer's Dementia by providing acoustic and textual data. With the multi-modal features, we modify a Convolutional Recurrent Neural Network based structure to perform classification and regression tasks simultaneously and is capable of computing conversations with variable lengths. Our test results surpass baseline's accuracy by 18.75%, and our validation result for the regression task shows the possibility of classifying 4 classes of cognitive impairment with an accuracy of 78.70%.

Index Terms: Multimodal Systems, Cognitve Decline Detection, Pre-trained Model

1. Introduction

Collecting a sufficient amount of electronic health records is a challenging task with various factors [1, 2]. Due to this problem, researchers in the medical field are often provided with only a small amount of data given. Owing to the fact that deep learning techniques perform better on large amounts of data, a number of studies using machine learning techniques have been conducted to solve specific medical problems, regarding a limited number of data [3, 4]. Dementia is also one of many medical symptoms facing this situation.

Dementia, a syndrome in which there is deterioration in cognitive function beyond what might be expected from normal ageing, is mostly affected by Alzheimer's Disease [5]. There were previous researches with various approaches to recognize Alzheimer's Dementia [6, 7, 8, 9], which has shown excellent performance. However, datasets used in these works were sufficient with quantity than the one used in this paper.

The ADReSS challenge [10] at INTERSPEECH 2020 hosts two tasks: Alzheimer's Dementia (AD) classification and Mini Mental Status Examination (MMSE) regression, while providing a refined dataset. The dataset is equally balanced of AD and non-AD participants with the metadata of age and gender. Each data is a conversation in which participants, in both audio and text modalities, spontaneously describes the picture given by the investigator. Participants of the challenge are suggested to solve hosted tasks using only the given data, where the numbers of train and test data are 108 and 48, respectively.

For recognizing AD with small amounts of data, we determined it would be beneficial to use both acoustic and textual features. Furthermore, we leverage models pre-trained on large scale datasets as feature extractor to get better representation. To this end, this paper focus on exploiting various multi-modal features, and design suitable network architecture. We compare 3 and 4 different acoustic and textual features, respectively, and use the hand-crafted (HC) feature and part-of-speech (POS) tagging as additional inputs. The usage of POS and HC is influenced by previous research, which has approved that using these features gained from transcript can improve the performance [8]. The proposed network is a modified version of Convolutional Recurrent Neural Network (CRNN); capable of computing conversations with variable lengths, and implemented with methods to fit with a small amount of data. Also, the model is able to compute using the acoustic feature only, without any metadata, which can be efficient considering the real-world situation. Our experimental results show using features of the pre-trained network leads to performance gain than that of raw, and regression results imply the potential of network classifying classes of cognitive impairment based on MMSE score.

2. Multi-Modal Features

This work compares 3 different acoustic and 4 different textual features. To obtain a speech signal corresponding to each utterance in the transcription, alignment of the transcription and the signal is done by using [11, 12]. Hence, the following multi-modal feature extraction in this section is applied to the aligned data.

2.1. Acoustic Features

- *openSMILE features*: The openSMILE v2.3 toolkit [13] provides multiple features from raw audio files. From the toolkit, we use the ComParE feature [14] and the eGeMAPS feature [15]. For ComParE feature, using one-way ANOVA, we select 393 features ($p \leq 0.05$), out of 6,373 concerning the efficiency of model capacity.
- *VGGish*: We use VGGish [16] which is trained with *Audio Set* [17] for audio classification. The feature is composed of 128 feature dimensions, where each feature is extracted from audio with a length of 960ms. To handle different lengths of utterance, we use the average value of the extracted VGGish features.

http://dx.doi.org/10.21437/Interspeech.2020-3153

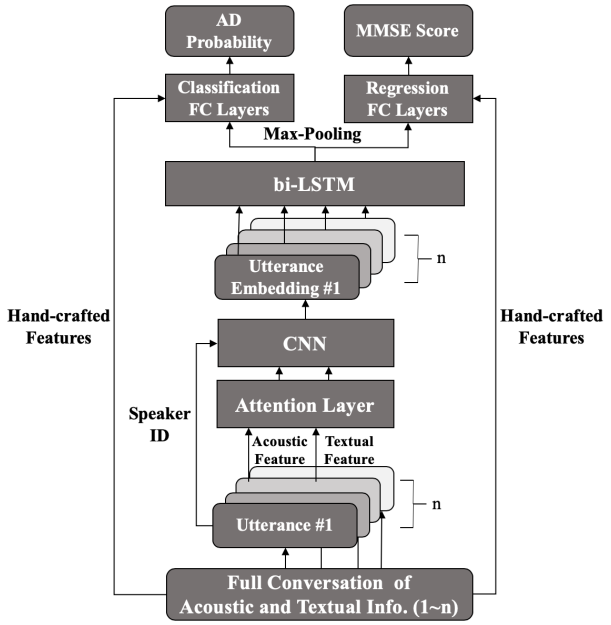


Figure 1: *Overview of the proposed method. The acoustic and textual features extracted from each utterance are fed in to the CRNN network. Then, the hand-crafted features retrieved from the participant's entire conversation are concatenated to the utterance-level features. Finally, the FC layer of each task estimates the AD probability and MMSE score of the participant.*

2.2. Textual Features

- *Pre-trained language model features*: We exploit transformer [18] based language models, GPT [19], RoBERTa [20], and Transformer-XL [21]. Pre-trained on large corpora, these language models have shown the effectiveness to improve performance over a wide range of natural language processing tasks. Sentence representations are obtained by averaging word embeddings via [22]. The specific settings for the language models are as follows, GPT: openai-gpt, RoBERTa: roberta-base, Transformer-XL: transfo-xl-wt103. The feature dimensions of GPT and RoBERTa is 768, and Transformer-XL dimensions of 1024. Besides the aforementioned features, we also use 300-dimensional GloVe vectors [23].
- *Hand-crafted features*: We integrate three categories, psycholinguistic, repetitiveness, and lexical complexity features, as HC features, which reflect the features of Alzheimer's. Psycholinguistic features and repetitiveness[1] are that suggested by [6], and lexical complexity is the Lexical Complexity Analyzer for Academic Writing (LCA-AW)[2]. These token-level HC features are aggregated to the conversational-level by only averaging participant's utterance. We select and use 23 features whose p-value from one-way ANOVA is less than 0.05 amongst a total of 42 features.

3. Proposed Method

While the proposed model can cope with additional inputs such as visual modality, the ADReSS challenge only offers acoustic and textual modalities. Thus, we primarily focus on the network with bimodal inputs. The overview of our model is as Figure 1. In case of unimodal, the network has the same structure, except that only a single modality feature is input.

3.1. Input

An input dialogue consists of its utterances and an extracted HC feature. Each utterance comes along with an acoustic and a textual feature, and a speaker index. The speaker index is a binary feature denoting an investigator or a participant, where it is extended as the size of the largest size of input feature dimension, 1024 in our case, by a single fully connected layer. Input features smaller than 1024 are also expanded the same way by a fully connected layer.

We apply dropout [24] to the input features before they are inserted into the network. This way, the model can be provided with more opportunity to learn independent representations, because each dimension can convey significant information, especially for the features extracted from pre-trained models.

3.2. Model Architecture

The proposed network is a modified version of CRNN, where an attention layer is a forefront layer of the network, and fully connected layers followed after the recurrent layer. Here, we use a bidirectional Long Short-Term Memory Network (bi-LSTM) [25] as the recurrent network.

Each modality input is individually inserted and computed through an attention layer. Our attention layer is implemented as the *Scaled Dot-Product Attention* mechanism introduced in [18]. We use a self-attention mechanism, where an individual feature is used as a query, key, and value during the attentional computation.

Outputs of the attention layer and embedded speaker index of a single utterance are channel-wise concatenated then inserted into the one-dimensional Convolutional Neural Network (CNN). After a convolutional layer expands channel dimension to 32, 6 *Squeeze-and-Excitation* (SE) [26] blocks are followed in the CNN. Each SE block consists of 2 convolution layers with a SE layer in between them. The last convolutional layer of every 2 SE blocks reduces feature dimension by convolutional stride factor of 4 and increments channel dimension. The expanding sizes of the channel dimension are 128, 512, 1024 respectively. Ultimately, CNN outputs 1024-dimensional channels with a global max pooled value.

After every utterance from the input dialogue is each computed through the CNN, the processed utterance embeddings are sequentially inputted into the bi-LSTM. The recurrent network consists of 3 bi-LSTM layers with 512 hidden units. Ultimately, the recurrent network outputs the max-pooled state from the results of the last layer's hidden states and is concatenated with HC.

Three fully connected (FC) layers follow after the bi-LSTM layers. Both the first two FC layers are followed by a rectified linear unit (ReLU) activation and reduce the input dimension by a factor of 4. The last activation function for classification and regression tasks are softmax and sigmoid, respectively. Ground truth MMSE score is scaled from 0 to 1 for regression loss computation.

3.3. Training and Inference

We use different numbers of utterances per batch during the training phase for the network to have opportunities to interpret

[1] https://github.com/vmasrani/dementia_classifier

[2] https://github.com/Maryam-Nasseri/LCA-AW-Lexical-Complexity-Analyzer-for-Academic-Writing

Table 1: *Validation Results of Acoustic Unimodal Network*

Feature	Accuracy	F1	RMSE
eGeMAPS	61.82%	71.98%	6.7178
ComParE	68.27%	74.62%	6.7852
VGGish	**85.27%**	**86.28%**	**5.1144**

various sequences of dialogue. The size is randomly selected between 5 and the minimum number of utterances among the dialogues in each batch. Since the minimum number of utterances of dialogue in the training data is 7, it was reasonable to set the minimum length to 5. If the length is too short, the network could be vulnerable to utterances with less meaningful data such as the investigator's "okay" or "mhm". A single batch is used during the inference phase to analyze every utterance in an input dialogue.

Our training loss for classification and regression tasks are binary cross-entropy error and mean squared error, respectively. The total cost function is a summation of these two values. We use the Adam optimizer [27] with a learning rate of 0.0002 and momentum parameters $\beta 1 = 0.5$, $\beta 2 = 0.9$.

4. Experiments

In this section, we evaluate model performances for both classification and regression tasks. Recorded performances are averaged value from measurements of 5-fold cross validation, where each fold contains 86 training and 22 validation conversations, except for the last fold containing 88 training and 20 validation conversations.

Prior experiments were conducted for optimizing several hyperparameters in the proposed network. First, we compared model performance with a one-dimensional convolutional kernel size of 3, 5, 10, 15. Through observations, larger kernel sizes led to performance gain; thus, we set the kernel size to 15. Attempts to ascertain the ideal dropout rate among 0 to 50% at 10% intervals could not be determined. Yet, we adopted a 20% dropout rate for data augmentation and prevention of overfitting. Finally, we discovered using 6 instead of 3 stacked convolutional blocks achieved better performance. Experimental results of each model shown in this section share above achieved hyperparameter values.

4.1. Feature Comparison

4.1.1. Unimodal Network

Table 1 is validation results of unimodal networks using acoustic features. The accuracy using VGGish exceeds openSMILE's by over 17%, which conveys a significant difference in these audio features on performance. Hence, this result establishes a strong point that using an acoustic feature extracted from a pretrained network outperforms features extracted from scratch.

Textual feature comparing experiment is further conducted by including combinations of using POS and HC features as input. Upon using POS, it is concatenated to the input textual feature to fed into the network. The best performing features for classification and regression are Transformer-XL and GloVe, respectively, according to Table 2.

4.1.2. Bimodal Network

We choose VGGish as a fixed auditory input feature for the bimodal network, considering its leading validation performance among other audio features. The use of POS and HC features is

Table 2: *Validation Results of Textual Unimodal Network*

Feature		Accuracy	F1	RMSE
GloVe	+ None	90.73%	0.9158	3.9282
	+ POS	90.73%	0.9122	3.8959
	+ HC	92.55%	0.9303	**3.3493**
	+ POS + HC	93.55%	0.9389	3.3650
GPT	+ None	91.55%	0.9224	3.7825
	+ POS	92.55%	0.9303	4.0275
	+ HC	91.55%	0.9246	3.6695
	+ POS + HC	89.82%	0.9076	3.7684
RoBERTa	+ None	92.45%	0.9312	3.7622
	+ POS	91.64%	0.9231	3.8437
	+ HC	93.45%	0.9391	3.3852
	+ POS + HC	93.45%	0.9391	3.3773
Transformer -XL	+ None	92.55%	0.9296	4.0078
	+ POS	93.45%	0.9382	4.0588
	+ HC	**94.36%**	**0.9469**	3.4866
	+ POS + HC	92.55%	0.9325	3.6602

Table 3: *Validation Results of Bimodal Network*

Feature		Accuracy	F1	RMSE
GloVe	+ None	92.55%	0.9288	4.0743
	+ POS	93.55%	0.9398	3.7091
	+ HC	90.73%	0.9122	3.9138
	+ POS + HC	93.55%	0.9382	3.4989
GPT	+ None	93.45%	0.9398	3.5503
	+ POS	93.45%	0.9398	3.9910
	+ HC	91.64%	0.9231	3.6334
	+ POS + HC	92.55%	0.9318	3.5182
RoBERTa	+ None	91.64%	0.9231	3.7842
	+ POS	92.55%	0.9318	3.6860
	+ HC	93.45%	0.9375	**3.4977**
	+ POS + HC	92.55%	0.9311	3.5182
Transformer -XL	+ None	91.64%	0.9201	4.0703
	+ POS	92.55%	0.9288	4.0546
	+ HC	90.73%	0.9114	3.7820
	+ POS + HC	**94.45%**	**0.9454**	3.6099

performed in the bimodal network as well. Acknowledging the results of unimodal networks, Transformer-XL feature is also well performed in the classification tasks, where RoBERTa feature scores the best root mean squared error (RMSE).

From the observation of this experimental result, it was reasonable to ascertain the best text feature. Notably, using Transformer-XL produced the highest performance in the classification task. Moreover, while comparing the average RMSE scores by feature, RoBERTa outputs the lowest score for both unimodal and bimodal networks. On the other hand, when analyzing performance between additional inputs, only little tendency could be observed. This can be an implication that the quantity of given data may not be sufficient for the additional inputs to exert influence.

4.2. Analysis of Regression Task

Figure 2 illustrates a graph comparing regression outputs from a bimodal network and the actual patient's corresponding MMSE score during validation stage. The severity classes, each shaded area in the figure, were categorized based on the MMSE score presented in [28].

In this example, VGGish and RoBERTa were used as input features, and the RMSE and r^2 value between network outputs and ground truths are 3.5182 and 0.7361, respectively. The

Table 4: *Results of Test Set*

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

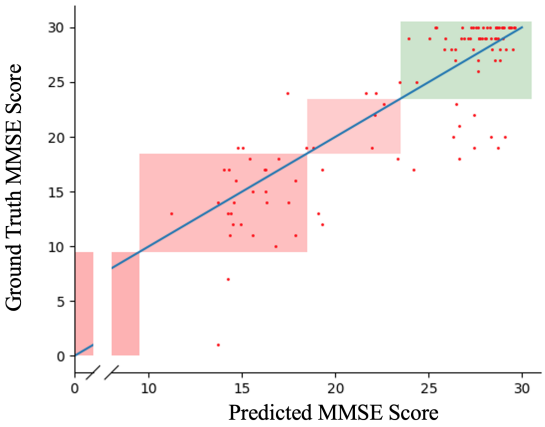


Figure 2: *MMSE comparison graph between regression outputs and ground truth scores - The linear line is a representation of the ideal output with zero error for the task. Each rectangular regions represent dementia severity based on the MMSE score and is shaded respectively as classes of normal, mild, moderate, and severe ranging from high to low values.*

output plot shows that the distribution of the network output decreases as the MMSE score decreases, which is inferred to follow the distribution of the given training data. Even though there was no network output below a score of 11, 78.70% of the points are included in the shaded area. This indicates that classifying severity classes of dementia is possible to some extent, based on regression outputs.

4.3. Test Set Results

The test dataset of the ADReSS challenge consists of 48 conversations and can be scored with a total of 5 different submissions. Taking this into account, we use two different models for unimodal and bimodal networks each and an ensembled output of bimodal networks to infer the test data. In the case of the unimodal network, VGGish and Transformer-XL are adopted to represent acoustic and textual modality, respectively. For the bimodal network, GloVe and Transformer-XL are adopted as textual modality regarding on their performance from the validation results. Besides, we select models using POS and HC features along with the bimodality inputs. Lastly, the outputs of the top 5 bimodal networks with high validation results are ensembled and used as the final submission.

The final result for each conversation was deduced by five different models with the same configurations used during the training and validation stage. Combining these results, the final output was concluded using majority voting for AD classification and the median value for the MMSE regression task. The baseline and our test results are presented at Table 4. When using only audio modality, our test accuracy surpasses baseline's by 10%, where accounting textual modality contributes another 8% performance gain. Although our textual unimodal model performed the best classification result among the single models, our bimodal network's ensembled output indicates that other bimodal models were able to achieve better performance. Furthermore, the test RMSE implies using both modalities is more advantageous for the regression task.

We could infer from the experimental results that the auditory information led to some performance degradation compared to the textual. This matter can be attributed to the low quality of the audio files provided. In particular, the participant's voice was barely hearable, while it was clear for the investigator's comments in some audio files. Even so, with the methodology to infer AD possible with only recorded audio files, the proposed model can be utilized as a real-world application reflecting on the difficulty of acquiring transcriptions and the target's metadata. The metadata is not dealt with in this work; this is because there was little difference when conditioning age and gender into our model in our prior empirical results.

5. Conclusion

This paper demonstrates extracted features from pre-trained networks are satisfactory for handling small amounts of data, to recognize Alzheimer's Dementia. The proposed model can compute variable lengths of dialogue and also introduce productive methods to fit the network with a little amount of data. Furthermore, our model does not require any metadata and also can perform well without transcript, which may be practical in real-world situations. Our test result outperforms baseline's with both tasks, and our regression results imply the potential of network classifying classes of cognitive impairment based on the MMSE score.

For future work, with the expectation of performance gain, mechanisms effectively fusioning different modality features [29] [30] can be applied in the model architecture.

6. Acknowledgements

This work was supported partly by Institute for Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-01367, BabyMind) and partly by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(NRF-2017M3C4A7078548). Also, the authors would like to thank Jeonghyun Yoon, and Ayoung Choi for their fruitful comments, and inspiration.

7. References

[1] M. Adibuzzaman, P. DeLaurentis, J. Hill, and B. D. Benneyworth, "Big data in healthcare–the promises, challenges and opportunities from a research perspective: a case study with a model database," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 384.

[2] T. Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh, "Barriers to retrieving patient information from electronic health record data: failure analysis from the trec medical records track," in *AMIA annual symposium proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 180.

[3] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, 2017.

[4] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Machine learning for predictive modelling based on small data in biomedical engineering," *IFAC-PapersOnLine*, vol. 48, no. 20, pp. 469–474, 2015.

[5] A. Burns and S. Iliffe, "Alzheimer's disease." *BMJ*, vol. 338, p. b158, 2009.

[6] 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.

[7] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.

[8] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," *arXiv preprint arXiv:1906.05483*, 2019.

[9] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.

[10] 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

[11] F. Schiel, "Automatic phonetic transcription of non-prompted speech," 1999.

[12] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[15] 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 transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] R. M. Crum, J. C. Anthony, S. S. Bassett, and M. F. Folstein, "Population-based norms for the mini-mental state examination by age and educational level," *Jama*, vol. 269, no. 18, pp. 2386–2391, 1993.

[29] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4193–4202.

[30] T. Yilmaz, A. Yazici, and M. Kitsuregawa, "Non-linear weighted averaging for multimodal information fusion by employing analytical network process," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 234–237.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automated Screening for Alzheimer's Dementia through Spontaneous Speech

Muhammad Shehram Shah Syed[1], *Zafi Sherhan Syed*[2], *Margaret Lech*[1], *Elena Pirogova*[1]

[1]RMIT University, Australia
[2]Mehran University, Pakistan

muhammad.shehram.shah.syed@rmit.edu.au

Abstract

Dementia is a neurodegenerative disease that leads to cognitive and (eventually) physical impairments. Individuals who are affected by dementia experience deterioration in their capacity to perform day-to-day tasks thereby significantly affecting their quality of life. This paper addresses the Interspeech 2020 Alzheimer's' Dementia Recognition through Spontaneous Speech (ADReSS) challenge where the objective is to propose methods for two tasks. The first task is to identify speech recordings from individuals with dementia amongst a set of recordings which also include those from healthy individuals. The second task requires participants to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech alone. To this end, we investigated characteristics of speech paralinguistics such as prosody, voice quality, and spectra as well as VGGish based deep acoustic embedding for automated screening for dementia based on the audio modality. In addition to this, we also computed deep text embeddings for transcripts of speech. For the classification task, our method achieves an accuracy of 85.42% compared to the baseline of 62.50% on the test partition, meanwhile, for the regression task, our method achieves an RMSE = 4.30 compared to the baseline of 6.14. These results show the promise of our proposed methods for the task of automated screening for dementia based on speech alone.

Index Terms: Social signal processing, Computational paralinguistics, Alzheimer's disease

1. Introduction

Dementia is an umbrella term for diseases which causes significant and continual cognitive and physical impairments. Individuals who are affected by dementia experience decline in language, thinking ability, and memory along with deterioration in their ability to perform day-to-day tasks in order to take care of themselves at a level which is beyond what is expected for ageing. According to the World Health Organization (WHO), there are around 50 million people worldwide who suffer from dementia and this number is increasing, with 10 million new cases every year [1]. Although there are various causes of dementia, Alzheimer's disease is the most prominent one, accounting for $60 - 70\%$ of total cases [1]. Alzheimer's disease is also known to adversely affect the mental health of care givers [2] such that they may require psychiatric interventions themselves.

It is known that cognitive impairments such as those caused by dementia affect the speech production system [3]. In [4], Yu et al. reported the use of vocal biomarkers for prediction of cognitive decline in the elderly population. They investigated the efficacy of a variety of acoustic features such as pitch variance, syllable rate, phoneme-based measures, and formant-based articulatory coordination features for automated cognitive impairment diagnosis. Ivanov et al. [5] developed phoneme-conditioned statistical models for cognitive impairment diagnosis and found them to be useful for the task at hand. Fraser et al. [6] consider a large number of features (370 in total) such as part-of-speech information, grammatical constituents, and vocabulary richness to capture linguistic phenomena which can identify subjects with dementia amongst a corpus which also includes healthy subjects. Luz et al. [7] used turn-taking patterns, speech rate, and other speech parameters which are essentially "content-free" for Alzheimer's disease recognition and report that their method achieves better accuracy than lexical, syntactic and semantic features.

In [8], Mirheidari et al. explored the use of word vector representations based on word2vec and GloVe embeddings for dementia recognition based on speech-transcripts and reported high accuracy. The authors hypothesized that since these embeddings can capture the semantics and syntax of words in a text, they will be useful for detecting diminished articulation from subjects with dementia. Haider et al. [9] investigate the efficacy of various types of speech paralinguistic features for voiced based screening from spontaneous speech. We find that the ADReSS challenge baseline closely follows the methodology proposed in [9].

In this paper, we propose methods for speech based screening of Alzheimer's dementia. To this end, we first train machine learning models which seek to model differences in characteristics of speech paralinguistics between subjects with dementia and those from the control group. Next, we conduct an exploratory analysis to generate numerical representations for speech transcripts based on recently developed deep language models. Our proposed models perform significantly better than the ADReSS challenge baselines for classification and regression tasks.

2. Dataset

The dataset for the Interspeech 2020 ADReSS challenge consists of speech recordings elicited for the Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam [10]. This data was explicitly balanced by the organizers in terms of age, gender, and the distribution of labels between the training and test partitions in order to minimize the risk of bias in the prediction tasks. The dataset has labels for machine learning tasks of binary classification and regression. As the name suggests, labels for the binary classification include Alzheimer's dementia and healthy control, whereas the labels for the regression task are Mini-Mental State Examination (MMSE) scores [11] which provide a means for dementia diagnosis based on linguistic tests. For further details regarding the dataset, we refer the reader to the ADReSS challenge baseline paper [12].

http://dx.doi.org/10.21437/Interspeech.2020-3158

3. Methodology

As part of our investigation into automated recognition of dementia with spontaneous speech as the input, we follow a two-pronged approach which includes voice-based screening and speech transcripts based screening as illustrated in Figure 1. For voice-based screening, we investigate the efficacy of acoustic features which are known to represent paralinguistic characteristics of prosody, voice quality, and spectra. Such categorization has previously proved to be useful for automated recognition of depression [13, 14] and bipolar disorder [15]. Meanwhile, our work on speech-transcripts based screening is largely exploratory such that we investigate the efficacy of deep language embeddings such as Bidirectional Encoder Representations from Transformers (BERT) [16] and its derivatives for generating a numerical representation of speech-transcripts.

3.1. Voice based screening

Here, we hypothesize that subjects with dementia have unique characteristics to their voice, given that the disease causes cognitive impairments, which can be quantified using acoustic descriptors of speech-paralinguistics. Following the approach of Horwitz et al. [13] for depression recognition, we propose to investigate the efficacy of acoustic features which characterize prosody, voice quality, and voice spectra. Prosody defines patterns of stress and intonation and is likely to be affected due to cognitive impairments. Voice quality analysis seeks to quantify changes at the vocal source level (glottis). It has been shown that the perceptual quality of voice changes on a scale between breathy and tense depending on the available cognitive resources [17]. Finally, acoustic descriptors of voice spectra have the potential to provide vital insights into muscular changes due to dementia at the vocal-tract level.

To this end, we compute prosody, voice quality, and spectral features using the openSmile [18] and COVAREP [19] toolkits. These toolkits have become the standard tools for computation of acoustic features for tasks related to social signal processing. These are not only open source but also freely available for academic research. In addition to the mentioned features, we use the (a) ComParE-2016 feature-set, (b) IS10-Paralinguistics feature-set, and (c) VGGish acoustic embeddings as part of our investigation of acoustic descriptors. The Computational Paralinguistics Challenge 2013 feature set (ComParE) is a brute-force feature set which has proved to be useful for a variety of speech paralinguistic tasks and is regularly used to set a baseline for Interspeech ComParE challenges [20, 21, 22]. The most recent version of the ComParE feature set was released as part of the 2016 edition of the ComParE challenge. The IS10-Paralinguistics feature set was introduced as part of the 2010 edition of Interspeech ComParE challenge and can be considered as a low-dimensional alternate to the ComParE feature set (6373 features vs 1582 features). Recently, we have found this feature set to be useful for tasks related to the recognition of bipolar disorder from speech [15] and emotion recognition [23]. Finally, we use VGGish embeddings [24] since they provide an alternative to domain-knowledge features such as those computed using openSmile and COVAREP toolkits.

The six types of acoustic features are computed as low-level-descriptors which means that they only represent the acoustic characteristics of a small chunk of the audio file. There is a need for these features to be aggregated using an appropriate method in order to generate a global acoustic representation for the speech recording. For this purpose, we use three types of feature aggregation methods: (a) functionals of descriptive statistics, (b) Bag-of-Audio-Words (BoAW), and (c) Fisher Vector encoding. These feature aggregation approaches are relatively well known in the research community and (mainly due to a requirement of brevity here) we refer the reader to [25, 26, 27, 28, 29] for details.

3.2. Screening based on Speech-Transcripts

The availability of speech transcripts provides a second modality which can be used alongside voice for the development of a multimodal framework for automated screening for dementia. This has been our objective, as illustrated in Figure 1. To this end, we conduct an exploratory analysis in order to determine the efficacy of pre-trained embeddings from deep language models for the task at hand. It must be mentioned here that these embeddings have already been shown to be useful for a large variety of tasks in the field of natural language processing [30, 31]. More specifically, we compute embeddings from eight models i.e. BERT base cased, BERT large cased, BERT large uncased, distilbert cased, distilbert uncased, distilroberta base, roberta base, and the biomed roberta base using the Huggingface Transformers library [32]. These embeddings are computed for each word of every transcript. In order to generate a transcript-level representation for transcripts we use four types of pooling functions which are average pooling (AvgPool), maximum value pooling (MaxPool), outlier-robust percentile-based range pooling (RangePool), and the coefficient of deviation (StdDevNormPool). The resultant feature vector is passed down to the machine learning pipeline as shown in Figure 1.

4. Experiments and Results

In this section, we present results for our experiments on speech based screening for Alzheimer's dementia. We used two types of algorithms each in order to predict labels for the classification and regression tasks. For the classification task, we used support vector machine classifier with a linear kernel (SVC) and logistic regression classifier. A grid search was carried out to optimize the model using leave-one-subject-out (LOSO) cross-validation whilst using the training partition. The optimization parameter *complexity* was tuned for both of these methods between a logarithmically-spaced range of 10^{-7} and 10^{3}. For the regression task, we used support vector machines based regression (SVR) (again with a linear kernel) whose hyperparameters were tuned using the same method as the classifier. In addition to SVR, we used a partial least squares regressor (PLSR) which has been shown to be useful for tasks related to speech paralinguistics [33]. A grid search was carried out to optimise the number of components for PLSR between 1 and 20. The results summarized in this section report the best performing models.

4.1. Voice based screening

A summary of classification results for voiced based screening has been provided in Table 1, where one finds that the IS10-Paraling.-BoAW model achieves the highest classification accuracy of the training partition with 76.85%, which is significantly better than the challenge baseline of 56.50%. This result is closely followed by the VGGish-BoAW model which achieves the second-best performance with an accuracy of 75.00%. Furthermore, the best performing models for Prosody, Voice Quality, and Spectra achieve a classification accuracy of 67.59%, 72.22%, and 71.30% respectively. This suggests that demen-

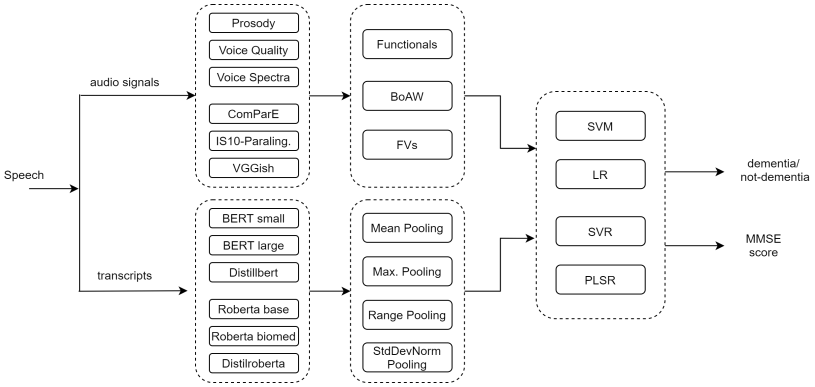


Figure 1: *Multimodal framework for automated screening of Alzheimer's' dementia*

tia may cause changes at voice source and vocal tract level, although a detailed investigation across datasets is required to support this observation. The best performing model based on ComParE features achieves an accuracy of 69.44%. It is important to note that all of these models achieve a better performance than the challenge baseline. The most interesting result from this table is that VGGish features provide better accuracy than most models trained on domain-knowledge based acoustic features such as Prosody features, Voice Quality features, Spectral features, and the ComParE features.

Table 1 also provides a summary of results for the regression task. Here one finds that the best performing model i.e. VGGish-BoAW achieves an RMSE = 5.95 which is better than the challenge baseline of 7.28. Furthermore, while MAE metric was not provided as part of the ADReSS challenge baseline, we find that the VGGish-FV BoAW model also achieves the smallest MAE of 4.49. These results are particularly interesting since they show that deep-learning based acoustic embedding can achieve a better performance than domain-knowledge based features and compliments our observation from the classification task. The performance of VGGish-BoAW is closely followed by BoAW and FV models based on IS10-Paralinguistic features. These models achieve an RMSE = 6.02 and RMSE = 6.04 respectively. The ComParE-FV model also achieved an RMSE = 6.04. Amongst the models which explicitly focus on characteristics of speech paralinguistics, we found that the Voice Quality-BoAW model achieved the smallest RMSE of 6.22, the Spectra-BoAW model achieved an RMSE = 6.12, and the Prosody-functionals model achieved an RMSE = 7.17 – all of these models achieve a smaller RMSE than the challenge baseline. This shows that modelling speech paralinguistics for recognition of dementia speech has promise, although, if the aim is to minimize the error between MMSE scores then the VGGish features with BoAW feature aggregation should be chosen.

4.2. Screening based on speech-transcripts

In Table 2 we provide a summary of classification results for the top-10 performing models based on text modality. Here, one can observe a notable improvement in the classification accuracy as compared to the challenge baseline accuracy of 62.5%, although it needs to be reminded that the challenge baseline was computed using audio modality [1]. The best performing model

Table 1: *Summary of results for classification and regression tasks using acoustic features for the training partition with LOSO cross-validation*

Feature Class	Feat. Agg.	Acc. (%)	RMSE	MAE
Prosody	Functionals	67.59	7.18	6.20
Voice Quality	Functionals	63.89	7.08	6.10
	BoAW	69.44	6.22	5.17
	FVs	72.22	6.52	5.49
Voice Spectra	Functionals	60.19	7.74	6.70
	BoAW	71.30	6.12	5.24
	FVs	71.30	6.12	4.89
IS10-Paraling.	Functionals	70.37	6.66	5.74
	BoAW	**76.85**	6.02	5.04
	FVs	66.67	6.04	5.21
ComParE	Functionals	68.52	7.16	5.69
	BoAW	65.74	6.90	6.13
	FVs	69.44	6.04	5.21
VGGish	BoAW	75.00	**5.92**	4.69
	FVs	62.96	6.75	5.53
Challenge baseline		56.50	7.29	—

i.e. *biomed roberta base* embedding with RangePool achieves an accuracy of 89.81%, which is followed by *roberta base* with RangePool which achieves an accuracy of 87.96%. Interestingly, we do not observe a difference in performance due to case and uncased versions of deep language models. For example, both *distilbert uncased* and *distilbert cased* models achieve the same accuracy, and the cased and uncased versions of the *BERT large* models achieve the same accuracy.

Table 3 summarizes the results for the top-10 performing models for MMSE scores prediction from the text modality. Here, one finds that the *BERT base uncased* embedding with MaxPool provides the best results in terms of the RMSE, achieving an RMSE = 4.32 which is better than the challenge baseline for regression of RMSE = 7.28. This is followed by the same BERT model but with RangePool which achieved an RMSE = 4.39. One can also note that all of the top-10 models based on text modality achieve a significantly better performance than the challenge baseline. It must be mentioned here for the sake of clarity that the baseline RMSE was computed us-

[1] A text modality baseline was added in the final version of the baseline paper with a classification UAR for train/test = $77.00\%/75.00\%$ and regression RMSE for train/test = $4.38/5.20$. As the reader shall note, our proposed methods still beat the updated challenge baseline.

ing audio features only (the organizer did not provide an RMSE computed using text features). Nevertheless, a comparison of results from Tables 1 and 3 makes it clear that the text modality is better for the task at hand.

Table 2: *Summary of results for top-10 performing models based on text modality for the classification task*

Feature Class	Pooling meth.	Accuracy (%)
biomed roberta base	RangePool	**89.81**
roberta base	RangePool	87.96
distilbert uncased	MaxPool	86.11
distilbert cased	MaxPool	86.11
BERT base uncased	MaxPool	86.11
BERT large uncased	AvgPool	86.11
BERT large cased	AvgPool	86.11
biomed roberta base	MaxPool	85.19
BERT base uncased	RangePool	85.19
BERT large cased	MaxPool	85.19

4.3. Predictions for the test partition

The ADReSS challenge baseline for the test partition is 62.50% and each participant has five attempts at predicting the labels of the test partition. A summary of the baseline and our results for the classification task is provided in Table 4. For our first attempt, we use predictions from the *biomed roberta base RangePool* model which was the best performing model for the training partition by achieving an accuracy of 89.81%. On the test partition, this model achieved an accuracy of 77.08% only which suggests that the model may have overfitted the training partition.

The second attempt used label fusion from the top-5 performing models from the text modality for the training partition (see Table 2). The resultant predictions for the test partition achieved an accuracy of 85.45%. This is not only our best result but also a large improvement from the challenge baseline of 62.50%. Our third attempt used label fusion from the top-5 performing models from the audio modality for the training partition (see Table 1). The resultant predictions for the test partition achieved an accuracy of 64.58% which is slightly better than the challenge baseline, although it does show that the audio modality offers weaker classification performance than the text modality. The fourth attempt used label fusion from the top-5 performing models from audio and text modalities (top-5 from each modality). The resultant predictions for the test partition achieved an accuracy of 79.17% which is an improvement over the results from the first and third attempt. For the final attempt, we used label fusion from the top-10 performing models overall (see Tables 1 and 2). Incidentally, all ten models are based on text modality. The resultant predictions for the test partition achieved an accuracy of 85.45% which is the same as the accuracy achieved by a fusion of top-5 models for text modality.

Table 3: *Summary of results for top-10 performing models based on text modality for the regression task*

Feature class	Pool meth.	RMSE	MAE
BERT base uncased	MaxPool	**4.32**	3.57
BERT base uncased	RangePool	4.39	3.62
distilbert uncased	RangePool	4.49	3.62
roberta base	AvgPool	4.49	3.48
BERT large cased	MaxPool	4.49	3.64
BERT large uncased	MaxPool	4.49	3.64
distilbert uncased	MaxPool	4.51	3.70
allenai biomed roberta base	AvgPool	4.51	3.68
allenai biomed roberta base	MaxPool	4.55	3.69
distilbert cased	AvgPool	4.57	3.51

Table 4: *Summary of results on the test partition for our proposed methods*

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

Similar to the classification task, each participant of the regression task has five attempts at predicting the MMSE scores. The challenge baseline for the regression task is an RMSE = 6.14. Our first attempt used predictions from the *BERT base uncased MaxPool* model, which was the best model on the training partition with an RMSE = 4.32. We find that this model achieved an RMSE = 4.83 on the test partition. The second attempt used test partition predictions from the *VGGish-BoAW* model which achieved an RMSE = 5.92 on the training partition but ends up achieving an RMSE = 6.91 on the test partition. This result is poorer than the challenge baseline.

Our third attempt used prediction for the test partition from the *BERT base uncased RangePool* model. This model was the second-to-best performing model for the training partition by achieving an RMSE = 4.39 and ends up achieving an accuracy of 5.18 on test partition which is still better than the challenge baseline. For the fourth attempt, we submitted the average value of predictions from our first and third attempt, the resultant predictions achieved an RMSE = 4.91 on the test partition. It is important to note that this score is slightly larger than the RMSE achieved from the first attempt. Finally, for our last attempt, we submitted an average of MMSE score predictions for the test partition from the top-10 performing models for the training partition. Interestingly, this setup produced our best RMSE score for the test partition with 4.30. This score easily beats the challenge baseline of 6.14.

5. Conclusions

In this paper, we investigated the efficacy of speech based automated screening of Alzheimer's dementia, a disease which significantly deteriorates the quality of life of affected individuals. From voiced based analysis we report that voice quality and voice spectral features perform better than features which characterise speech prosody. However, the best performing model from voice modality was based on VGGish deep acoustic embeddings. Overall, we report that the text modality which is available in the form of speech-transcripts perform the best by achieving an accuracy of 89.91% for the training partition. On training and test partitions, our methods outperformed the challenge baselines for both classification and regression tasks.

6. References

[1] World Health Organisation, "Dementia: Key Facts," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] A. S. Alfakhri, A. W. Alshudukhi, A. A. Alqahtani, A. M. Alhumaid, O. A. Alhathlol, A. I. Almojali, M. A. Alotaibi, and M. K. Alaqeel, "Depression among caregivers of patients with dementia," *Inquiry (United States)*, vol. 55, no. 1, pp. 1–6, 2018.

[3] G. W. Ross, J. Cummings, and D. F. Benson, "Speech and language alterations in dementia syndromes: Characteristics and treatment," *Aphasiology*, vol. 4, no. 4, pp. 339–352, 1990.

[4] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *INTERSPEECH*, 2015, pp. 3734–3738.

[5] A. V. Ivanov, S. Jalalvand, R. Gretter, and D. Falavigna, "Phonetic and anthropometric conditioning of MSA-KST cognitive impairment characterization system," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 228–233.

[6] 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, 2015.

[7] S. Luz, S. de la Fuente, and P. Albert, "A Method for Analysis of Patient Speech in Dialogue for Dementia Detection," in *International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 35–42.

[8] 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. 1–5.

[9] 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, 2019.

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

[11] 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.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[13] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *IEEE International Conference on Body Sensor Networks (BSN)*, 2013, pp. 1–6.

[14] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[15] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint:1810.04805v2*, vol. 1, no. 1, pp. 1–16, 2018.

[17] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech Deficits in Serious mental Illness: A Cognitive Resource Issue?" *Schizophrenia research*, vol. 160, no. 0, pp. 173–179, 2014.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM international conference on Multimedia*, 2013, pp. 835–838.

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[20] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[21] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Noth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *INTERSPEECH*, 2019, pp. 1–5.

[22] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[23] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the Urdu-Sindhi Speech Emotion Corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1–6, 2020.

[24] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 131–135.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lecture Notes in Computer Science*, vol. 6314, 2010, pp. 143–156.

[27] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[28] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[29] Z. S. Syed, J. Schroeter, K. Sidorov, and D. Marshall, "Computational Paralinguistics: Automatic Assessment of Emotions, Mood, and Behavioural State from Acoustics of Speech," in *INTERSPEECH*, 2018, pp. 511–515.

[30] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 1, no. 1, pp. 100 943 – 100 953, 2019.

[31] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challengesand New Directions in Sentiment Analysis Research," *arXiv:2005.00357*, vol. 1, no. 1, pp. 1–26, 2020.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv:1910.03771*, pp. 1–11.

[33] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 37–43.