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

Comparing Acoustic-based Approaches for Alzheimer's Disease Detection

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

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

Abstract

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

1. Introduction

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

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

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

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

1. **Extracting conventional acoustic features from speech:** with this approach, we extract acoustic features from the audio files for binary AD vs non-AD classification. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection [4, 9].
2. **Using pre-trained deep neural models:** with this approach, we embed the raw audio into representations using the last hidden state of the pre-trained wav2vec 2.0 model [10] and classify audio samples using these embeddings as input to simple classifiers.
3. **A combination of both approaches.** Here, we combine the two approaches and make use of both engineered fesatures and audio representations generated using a pre-trained deep neural model.

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

The main contributions of our paper are as follows:

- We use audio representations extracted from the pre-trained deep neural model wav2vec 2.0 [10] for the task of AD classification from speech. To the best of our knowledge, this method and this model were never applied before for such a classification task.
- We present the model that combines conventional acoustic features with pre-trained representations of speech and outperforms the ADReSSo baseline model by 2.82%.
- We carefully compare three different acoustic-based approaches for AD detection, which allows us to draw more general conclusions on performance and generalizability of acoustic models.

2. Background

2.1. Extracting conventional acoustic features from speech

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

http://dx.doi.org/10.21437/Interspeech.2021-759

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

2.2. Using pre-trained deep neural methods

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

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

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

2.3. A combination of pre-trained and conventional approaches

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

3. Methodology

3.1. Dataset

The ADReSSo dataset we use in this work consists of set of speech recordings of picture descriptions produced by cognitively normal (or healthy) subjects and patients with an AD diagnosis, who were asked to describe the Cookie Theft picture [22] from the Boston Diagnostic Aphasia Examination [22]. There are speech samples from 237 participants in total, out of which 166 are in the training set, and 71 in the test set ($70/30$ split balanced for demographics). Out of the samples in the training set, 83 were cognitively healthy, while 83 had AD. The prediction dataset used is matched for age and gender so as to minimise risk of bias in the prediction tasks using a propensity score. Along with the recorded speech data, the dataset also included segmentation profiles for optional use. For all our acoustic approaches, we use the full unsegmented audio, following the baseline acoustic approach [7] and prior works [18, 23].

3.2. Feature Extraction

3.2.1. Using conventional acoustic features from speech

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

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

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

3.2.3. A combination of both approaches

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

3.3. Evaluation Methods

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

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

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

[1] https://huggingface.co/models

[2] https://librosa.org/

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

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

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

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

3.4. Experiments

3.4.1. Using conventional acoustic features from speech

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

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

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

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

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

4. Results

4.1. AD detection results on the ADReSSo train set

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

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

4.2. AD detection results on the unseen ADReSSo test set

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

5. Discussion

5.1. Classification performance

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

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

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

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

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

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

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

5.2. Generalizability

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

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

Both embedding-based models (LR-embed and DT-embed) have a substantial gap between the test performance and the cross-validated performance on the train set (6.7-13.6% difference, with the test performance always being lower). The combination approaches on average, however, result in a smaller gap, especially in terms of accuracy, recall and F1 score. The SVM-combo model even performs better on the test set, when evaluated using recall and F1 metrics. In terms of accuracy, the SVM-combo model shows only 0.46% lower performance on the test set than on the train set, when cross-validated using the LOSO method. Such a slight difference confirms that the combination approach in general is a more generalizable method to train the AD detection system, comparing to using the embedding-based representations of speech alone.

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

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

6. Conclusions and Future Work

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

7. References

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

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

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

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

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

[6] Z. Zhu, J. Novikova, and F. Rudzicz, "Detecting cognitive impairments by agreeing on interpretations of linguistic features," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1431–1441.

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

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

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

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

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

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

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

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

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

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

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

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

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

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

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

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

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

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

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

Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only

Jun Chen[1], *Jieping Ye*[1], *Fengyi Tang*[2], *Jiayu Zhou*[2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, USA
[2]Department of Computer Science and Engineering, Michigan State University, USA

junnchen, jpye@umich.edu, jiayuz, tangfeng@msu.edu

Abstract

Alzheimer's disease (AD) is a neurodegenerative syndrome which affects tens of millions of elders worldwide. Although there is no treatment currently available, early recognition can improve the lives of people with AD and their caretakers and families. To find a cost-effective and easy-to-use method for dementia detection and address the dementia classification task of InterSpeech 2021 ADReSSo (Alzheimer's' Dementia Recognition through Spontaneous Speech only) challenge, we conduct a systematic comparison of approaches to detection of cognitive impairment based on spontaneous speech. We investigated the characteristics of acoustic modality and linguistic modality directly based on the audio recordings of narrative speech, and explored a variety of modality fusion strategies. With an ensemble over top-10 classifiers on the training set, we achieved an accuracy of 81.69% compared to the baseline of 78.87% on the test set. The results suggest that although transcription errors will be introduced through automatic speech recognition, integrating textual information generally improves classification performance. Besides, ensemble methods can boost both the accuracy and the robustness of models.

Index Terms: Cognitive Decline Detection, Modality Fusion, Alzheimer's Disease, Computational Paralinguistics

1. Introduction

Dementia is a syndrome associated with a deterioration in memory, language, problem-solving, and other cognitive functions to perform daily activities. According to the estimation of WHO, there are around 50 million people having dementia worldwide, and this number is increasing by nearly 10 million every year. Amongst many different forms of dementia, Alzheimer's disease is the most common one and contributes to 60–70% of total cases [1]. Screening of Alzheimer's dementia is typically conducted through paper-and-pencil cognitive tests, such as the Mini Mental Status Examination (MMSE) [2] and the Montreal Cognitive Assessment (MoCA) [3]. Although cheap and quick to administer, the scoring process totally relies on the personal judgment of clinicians, which may introduce errors and result in a high inter-rater variability [4]. To address these issues, extensive studies have been carried out for the purpose of automated cognitive assessment [5]. One promising direction is speech-based screening. Speech signals can be relatively easily collected throughout the day without burdening the participants or the researchers. Moreover, the rapid development of speech technology and machine learning algorithms provides us a good opportunity to utilize those speech data for automatic screening of dementia [6] and finally translate speech-based methods into clinical practice.

There are existing efforts on the acoustic characteristics of AD. In [7], Warnita et al. extracted several sets of paralinguistic features from speech utterances in DementiaBank Pitt Corpus [8]. They trained a gated convolutional neural network for utterance-level AD classification, and then made the final verdict for each subject through majority voting. The best accuracy of 73.6% was achieved from this acoustic-only method. Luz et al. [9] assessed the effectiveness of several other acoustic feature sets with different classifiers for AD detection on the same data set. They showed that the eGeMAPS feature set provided the best single feature set accuracy and simple hard fusion of feature sets could improve the accuracy from 71.34% to 78.70%. [10] used low-level descriptors of IS10-Paralinguistics feature set [11] and Bag-of-Acoustic-Words (BoAW) for feature aggregation, and achieved a leave-one-subject-out (LOSO) accuracy of 76.85% on the training set of InterSpeech 2020 ADReSS challenge. Moreover, there are evidences showing that using manual transcripts of speech or a combination of transcripts and speech audio can generally lead to better performance compared to using audio alone. In [12], Yuan et al. explored disfluencies and language problems in Alzheimer's Disease subjects, and achieved the best accuracy 89.6% on the test set of the ADReSS challenge by fine-tuning Transformer-based pre-trained language models. Syed et al. achieved an accuracy of 85.45% on the same challenge task [10] by using both acoustic features and linguistic features from manual transcription.

In this paper, we conducted a systematic comparison of methods of detecting cognitive impairment based on a narrative speech from a picture description task, and extensively explored different modality fusion strategies. We first investigated the characteristics of acoustic modality and trained machine learning classifiers based on speech paralinguistic feature sets. Next, we generated two sets of transcripts through automatic speech recognition (ASR) and extracted linguistic features, both deep text embedding, and human-defined psychological features, from transcripts for dementia screening. Finally, we compared different modality fusion strategies to boost the performance of our model. Our proposed model outperformed the ADReSSo challenge baseline for AD classification task on both training partition and test partition.

2. Dataset

The ADReSSo challenge [6] released two distinct benchmark datasets for three different tasks. The dataset for AD classification consists of audio recordings of a picture descriptions task from both cognitively healthy people and patients diagnosed with AD. Participants were asked to describe the *Cookie Theft* picture from the Boston Diagnostic Aphasia Examination [8]. Recordings were preprocessed with stationary noise removal and audio volume normalization across audio segments to reduce the variation caused by recording conditions [6].

The resulting dataset includes 237 audio files. To minimize the risk of bias in the prediction, these files are carefully partitioned into training and test sets at a ratio of 7:3 so as to preserve

http://dx.doi.org/10.21437/Interspeech.2021-2002

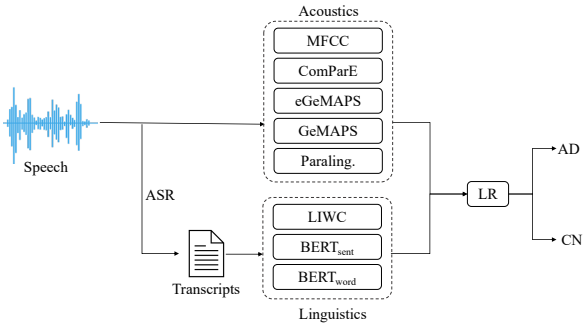


Figure 1: *Framework of automatic screening of AD.*

the balance of gender and age distribution [6]. There are 166 instances allocated to the training set. Among them, 87 subjects are diagnosed with dementia and 79 are elderly normal controls. The other 71 instances are allocated to the test set, among which 35 are with an AD diagnosis and 36 are cognitively normal.

3. Methodology

In this paper, we investigate the efficacy of logistic regression (LR) model on AD screening, as well as modality fusion strategies to boost the performance of the model. The schematic workflow of audio-based and ASR transcripts-based screening is shown in Figure 1.

3.1. Screening based on speech

In this section, we use python library librosa [13] and open-source audio feature extraction toolkit openSMILE [14] for audio preprocessing and paralinguistic acoustic feature extraction. Paralinguistics have been widely used for emotion recognition and detection of some other mental disorders such as depression [15] and bipolar disorder [16]. Evidence shows that AD patients have deterioration in emotional control [1] and may have difficulty in expressing emotions in prosodies [17]. Based on this, we hypothesize that paralinguistic acoustics is a good candidate for AD biomarkers. Five sets of acoustic features which are known to represent paralinguistic characteristics of speech are extracted as follows.

1. Mel-frequency cepstral coefficients (MFCCs)[18]

   The first 13 MFCC bands (0-12), and corresponding 13 delta MFCCs and 13 delta-delta MFCCs, which reflect the rate of change and the acceleration in MFCCs, are extracted. Descriptive statistics functions are applied, totaling 468 features for one utterance.

2. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[19]

   GeMAPS contains the 18 low-level descriptors(LLD), including frequency related parameters pitch, jitter, formants, energy related parameters shimmer, loudness, harmonics-to-noise ratio (HNR), and several spectral parameters. Statistical functionals are applied to each LLD, totalling 62 parameters.

3. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[19]

   An extension set, which contains 7 LLDs of cepstral and dynamic parameters and corresponding functionals, is added to the GeMAPS. In total, 88 features are extracted per utterance.

4. INTERSPEECH 2016 Computational Paralinguistics Challenge Feature Set (ComParE-2016)[20]

   ComParE is the largest standard set of openSMILE with 6373 features. It is a brute-force feature set that has proved to be useful for a variety of speech paralinguistic tasks. ComParE-2016 is the most recent version of the ComParE.

5. INTERSPEECH 2010 Paralinguistics Challenge Feature Set (IS10-Paraling.)[11]

   IS10-Paralinguistics contains 38 LLDs and 38 corresponding delta coefficients. 21 functionals are applied to these LLDs, totaling 76 LLD for one frame and 1582 features for one utterance. This feature set can be considered as a low-dimensional alternate to the ComParE.

Extracted feature sets are normalized by standard min-max scaling and passed down to the logistic regression classifier as illustrated in the machine learning pipeline in Figure 1.

3.2. Screening based on auto-transcription

Language impairment is a distinguishing marker of dementia [5]. In addition to analyzing acoustic characteristics of AD, we also utilize ASR techniques to introduce the second modality, which is transcript text. Three sets of linguistic features are then extracted from each auto-transcript, as illustrated in Figure 1.

3.2.1. Automated transcript generation through ASR

We use two different approaches to generate transcripts from spontaneous speech. The first approach is using a pre-trained English DeepSpeech model from Mozilla [21]. The second is using Google Cloud standard speech-to-text service. Both transcribing processes can be done automatically, with no need for model fine-tuning or human intervention. Linguistic feature sets are extracted from transcripts produce by each approach.

3.2.2. Linguistic feature based screening

The language feature sets can be largely categorized into two classes. One is a transparent linguistic feature set, in which each individual feature has a well-defined meaning and thus having desirable interpretability. Linguistic Inquiry and Word Count (LIWC) [22] is one such text analysis method. It counts words in psychologically meaningful categories. Extensive studies using LIWC demonstrated its ability to detect meaning in a wide variety of experimental settings, such as attentional focus, emotion, and thinking styles. Here we extract a 64-dimensional LIWC vector from each transcript.

Another category of linguistic feature is language embedding which usually a dense numerical representation of a word or a sentence. Those embeddings have already shown great success in a variety of tasks of natural language processing [23]. We investigate the efficacy of pre-trained embeddings from deep language model Bidirectional Encoder Representations from Transformers (BERT) [24]. Specifically, we compute embedding for each word(BERT$_{word}$) and embedding for the whole transcript(BERT$_{sent}$) from pretrained BERT base uncased model using the Huggingface Transformers library[25]. BERT$_{sent}$, as a 768-dim linguistic feature, is passed to the downstream pipeline directly, while BERT$_{word}$ are first aggregated by four kinds of pooling functions (max, min, average, and standard deviation) to generate a transcript-level representation. The outputs of

Table 1: *Summary of results for AD classification using acoustic features on LOSO cross-validation*

Feature Set	Accu.	F1	Spec.
MFCC	67.47	69.32	64.56
GeMAPS	68.67	70.79	64.56
eGeMAPS	**74.10**	**75.98**	69.62
ComParE	71.69	72.83	70.89
Paraling.	71.69	72.19	**73.42**
Maj. voting	**77.11**	**78.41**	**74.68**
Avg. fusion	74.7	75.58	**74.68**
Wgt. fusion	74.7	75.58	**74.68**

pooling operations are concatenated, resulting in a 6373-dim vector, and passed down to the LR classifier (Figure 1).

3.2.3. Modality fusion

To make full use of audio and text modalities, we explored a multimodal framework for the automatic screening of AD. We first apply a straightforward early fusion or data-level fusion [26] by selecting a good performing feature set in each modality, and concatenating features into a single vector for each sample. Different combinations of feature sets are investigated for comparison. Besides, we also applied the late fusion (decision level fusion) strategy which uses input modalities independently followed by fusion at a decision-making stage. It is inspired by the popularity of ensemble methods [27]. Three different rules are used in this paper to combine independently trained classifiers: a) majority voting of predicted class labels, b) average fusion of predicted class probabilities, and c) weighted average fusion of class probabilities with the weight set as the accuracy of the corresponding base classifiers.

4. Experiments and Results

We use LR classifier for the AD classification task. Two hyperparameters are fine-tuned for optimized performance. Regularization strength λ was tuned with grid search between a range of $5e$-5 and $1e2$, and penalty is chosen from {L1, L2}.

4.1. LOSO evaluation of acoustic models

In Table 1, we show the classification results of five acoustic feature sets. eGeMAPS outperforms other sets with a leave-one-subject-out (LOSO) accuracy of 74.10% on the training set. This result is closely followed by the Paraling. and ComParE models which achieve the second-best single model accuracy of 71.69%. Since Paraling. is a low-dimensional alternate to the ComParE set, these tied results are as expected. The model based on GeMAPS achieved an accuracy of 68.67%, which is 5.43% lower than the eGeMAPS based model. This indicates that the 26 additional features in the extended minimalistic set do have some contribution to dementia recognition. Furthermore, the ensemble methods can boost the overall performance of audio modality to 77.11%.

4.2. LOSO evaluation of linguistic models

From the summary in Table 2, we see that models based on transcripts generated by Google Cloud universally outperform models based on DeepSpeech transcripts. Among them, BERT word embedding model achieved the best LOSO accuracy of 75.30%. This result is followed by the result of another embedding based model $\text{BERT}_{\text{sent}}$. LIWC based models from both transcription settings have an accuracy of less than 70%.

Table 2: *Summary of results for AD classification using linguistic features on LOSO cross-validation*

ASR Model	Feature Set	Accu.	F1	Spec.
DeepSpeech	LIWC	67.47	71.28	56.96
	$\text{BERT}_{\text{word}}$	68.07	71.04	60.76
	$\text{BERT}_{\text{sent}}$	70.48	73.22	63.29
	Maj. Voting	69.88	72.53	63.29
	Avg. Fusion	66.87	69.95	59.49
	Wgt. Fusion	67.47	70.33	60.76
GoogleCloud	LIWC	69.88	71.26	68.35
	$\text{BERT}_{\text{word}}$	**75.30**	**76.30**	**74.68**
	$\text{BERT}_{\text{sent}}$	72.89	74.58	69.62
	Maj. Voting	75.30	76.84	72.15
	Avg. Fusion	71.08	72.41	69.62
	Wgt. Fusion	71.08	72.41	69.62
Overall	Maj. Voting	73.49	75.0	70.89
	Avg. Fusion	72.29	74.44	67.09
	Wgt. Fusion	72.29	74.44	67.09

We noticed that the average number of words is 60.9 for transcripts generated by pretrained DeepSpeech and 91.2 for those generated by Google Cloud speech-to-text service, which indicates that the latter may have a better speech recognition performance on this particular data set. This partially explains why models based on the latter generally have better classification performance.

4.3. Modality Fusion

An ideal feature set should be compact enough to be implemented in a real-time system and robust enough to detect subtle changes of spontaneous speech of people developing dementia. Here we used early and late fusion strategies to investigate the multi-modality classification problem and the results are summarized in Table 3. Three linguistic feature sets based on Google Cloud transcripts are combined with each one of the top three acoustic feature sets in Section 4.1. The best accuracy is achieved by simple early fusion of google-LIWC (g-LIWC) and eGeMAPS. This combination also produces the most compact feature set (152 features in total) in our modality fusion settings. Furthermore, feature sets from both modalities are interpretable, which can help us to have a better understanding of the linguistic and paralinguistic characteristics of the disease. The second-best performance comes from the late average fusion of google-$\text{BERT}_{\text{word}}$ (g-$\text{BERT}_{\text{word}}$) and eGeMAPS, which is higher than the early fusion of the same sets. Late fusion, however, does not always outperform early fusion. Late fusion of g-LIWC and eGeMAPS, for example, has an accuracy 9% lower than the early fusion. Besides, we also applied ensemble methods to combine the predictions of either all of the twenty single classifiers or the selected top ten classifiers. An early fusion model is considered a single classifier here since it is trained only once after feature concatenation. The best ensemble accuracy is 80.72% following the majority voting rule.

Table 3: *Results of multimodal methods on training set LOSO cross-validation*

FS	Feature Set	Accu.	F1	Spec.
Early	g-LIWC+eGeMAPS	**81.93**	**82.56**	**82.28**
	g-BERT$_{word}$+eGeMAPS	75.90	76.74	75.95
	g-BERT$_{sent}$+eGeMAPS	74.10	75.71	70.89
	g-LIWC+ComParE	71.69	73.45	68.35
	g-BERT$_{word}$+ComParE	74.70	75.58	74.68
	g-BERT$_{sent}$+ComParE	75.30	76.02	75.95
	g-LIWC+Paraling.	75.90	76.74	75.95
	g-BERT$_{word}$+Paraling.	77.11	77.91	77.22
	g-BERT$_{sent}$+Paraling.	78.92	80.23	75.95
Late	g-LIWC + eGeMAPS	72.29	74.16	68.35
	g-BERT$_{word}$ + eGeMAPS	**77.71**	**78.61**	**77.22**
	g-BERT$_{sent}$ + eGeMAPS	75.90	77.01	74.68
	g-LIWC + ComParE	71.08	72.09	70.89
	g-BERT$_{word}$ + ComParE	72.89	74.29	70.89
	g-BERT$_{sent}$ + ComParE	72.29	73.26	72.16
	g-LIWC + Paraling.	73.49	74.42	73.42
	g-BERT$_{word}$ + Paraling.	72.29	74.73	65.82
	g-BERT$_{sent}$ + Paraling.	71.08	71.76	72.15
Ens.	Overall Avg.	75.3	76.84	72.15
	Overall Maj.	77.11	79.12	70.89
	Overall Wgt.	80.12	80.92	79.75
	Top-10 Avg.	79.52	80.23	79.75
	Top-10 Maj.	**80.72**	**81.18**	**82.28**
	Top-10 Wgt.	79.52	80.23	79.75

4.4. Predictions for the test partition

ADReSSo challenge allows each team to submit the results of five attempts. A summary of the baseline and our results for the test set is provided in Table 4. For the first attempt, we use predictions from the early fusion of g-LIWC and eGeMAPS, which was the best performing model on the training partition by achieving an accuracy of 81.93%. On the test set, however, this model only achieved an accuracy of 64.79%, which indicates overfitting on the training set. The second attempt is the early fusion of g-BERT$_{sent}$ and Paraling., which is the second-best early fusion model. The performance dropped a little bit on the test set from 77.11% to 74.65%. The third attempt is the majority voting of all the classifiers from audio modality which achieved an accuracy of 67.61% on the test set. This indicates that the information that audio modality offers is not robust and sufficient enough for AD detection. The fourth and fifth attempts used ensemble strategies. Majority voting and average fusion are applied separately on the top 10 best-performing models. The resultant prediction accuracy scores for the test partition are 80.28% and 81.69%, which are both better than the challenge baseline of 78.87%. We also noticed that these two models achieved similar accuracy on the training set and test set, which shows that the ensemble increased the robustness of the models.

5. Discussion

The effectiveness of several paralinguistic feature sets for Alzheimer's recognition was evaluated in Section 4.1. In addition to utilizing those predefined features sets, one future direction could be introducing paralinguistic embeddings generated from a representation model pretrained on a large external dataset to our dementia recognition pipeline. While representation learning models like BERT have achieved great success in the text domain, such methods are underutilized in the speech domain.

Table 4: *Summary of results for AD classification on test set*

Feature Set	Accu.	F1	Spec.
g-LIWC+eGeMAPS	64.79	72.22	61.54
g-BERT$_{sent}$+Paraling.	74.65	80.56	72.73
Audio Maj.	67.61	77.78	63.49
Top-10 Maj.	80.28	88.89	78.13
Top-10 Avg.	**81.69**	**88.89**	**80.00**
Challenge baseline[6]	78.87	77.78	78.87

Our model built on linguistic and acoustic features achieved the best accuracy of 81.69% on the test set, while paralinguistics based model only had an accuracy of 67.61%. This indicates that although paralinguistic changes are potential markers of dementia, acoustic modality alone may not have enough information for the diagnosis of disease. Introducing linguistic features through ASR often leads to a considerable improvement to the predictions accuracy of the disease, despite the fact that the ASR transcripts have a relatively high word error rate.

Another observation is that, from the aspect of accuracy and robustness, ensemble methods generally gives better performance, especially when we have lots of individual classifiers. This is because the errors from multiple models are dealt with independently.

We also noticed that the most promising model based on eGeMAPS and g-LIWC on the training set performs worst among the five attempts on the test partition. One explanation for the performance gap between training and test stages might be the disadvantage of LOSO cross validation. Even though test-error is unbiased in each iteration, LOSO has a high variability as only one observation is predicted for validation. Stratified 10-fold cross-validation or nested cross validation could be applied in future study to alleviate the overfitting issue caused by LOSO. Another possible explanation might be the choice of base classifier. Our logistic regression model marginally outperformed other machine learning classifiers, including SVM, decision tree, and multilayer perceptron, regarding the LOSO performance on the training set, while other classifiers might be more robust to the outliers or have a better generalization capability.

6. Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-1749940, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) RF1AG072449.

7. References

[1] "Dementia," Sep 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of pa-

tients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

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

[4] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[5] 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.

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

[7] 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.

[8] J. T. Becker, "The natural history of alzheimer's disease," *JAMA Neurology*, vol. 51, no. 6, p. 585, 1994.

[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] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[14] 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.

[15] 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.

[16] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 39–45.

[17] G. Tosto, M. Gasparini, G. Lenzi, and G. Bruno, "Prosodic impairment in alzheimer's disease: assessment and clinical relevance," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 23, no. 2, pp. E21–E23, 2011.

[18] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221)*, vol. 1. IEEE, 2001, pp. 73–76.

[19] 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.

[20] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, 2016, pp. 2001–2005.

[21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[23] S. Ghannay, B. Favre, Y. Esteve, and N. Camelin, "Word embedding evaluation and combination," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 300–305.

[24] 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.

[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 preprint arXiv:1910.03771*, 2019.

[26] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[27] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

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

Identifying Indicators of Vulnerability from Short Speech Segments using Acoustic and Textual Features

Xia Cui[1,2], *Amila Gamage*[2], *Terry Hanley*[1], *Tingting Mu*[1]

[1]The University of Manchester, United Kingdom
[2]VoiceIQ Ltd., United Kingdom

{xia.cui, terry.hanley, tingting.mu}@manchester.ac.uk, {xia, amila}@voiceiq.ai

Abstract

In order to protect vulnerable people in telemarketing, organisations have to investigate the speech recordings to identify them first. Typically, the investigation is manually conducted. As such, the procedure is costly and time-consuming. With an automatic vulnerability detection system, more vulnerable people can be identified and protected. A standard telephone conversation lasts around 5 minutes, the detection system is expected to be able to identify such a potential vulnerable speaker from speech segments. Due to the complexity of the vulnerability definition and the unavailable annotated vulnerability examples, this paper attempts to address the detection problem as three classification tasks: age classification, accent classification and patient/non-patient classification utilising publicly available datasets. In the proposed system, we trained three sub models using acoustic and textual features for each sub task. Each trained model was evaluated on multiple datasets and achieved competitive results compared to a strong baseline (i.e. in-dataset accuracy).

Index Terms: vulnerability detection, speech and text processing, age classification, accent classification, patient/non-patient classification, feature extraction, feature selection

1. Introduction

Protecting vulnerable people is a vital part of government regulation bodies and commercial companies in telemarketing [1, 2]. Vulnerability is a complex issue to detect as it is a multifaceted phenomenon that involves considering biological, psychological and social elements. According to [1, 2], everyone can be vulnerable – people with health conditions, older adults and children are arguably more likely to be vulnerable. Further, as the conversations conducted are conducted in English, people from non-English speaking countries may also be more vulnerable to being mis-sold products. When there is no priority information of the vulnerability criteria in the database, it is costly and time-consuming for an investigator to access a large number of recordings to identify the vulnerable people. Given this, there is an increasing demand to develop an automatic vulnerability detection system [2]. To the best of our knowledge, few studies [3, 4] have been conducted in the community on tackling the vulnerability in Speech Processing, addressing a fraction of vulnerability concerns. In order to adapt to a real-time system, this paper reports the development of a detection system that is able to work from a short speech segment (i.e. the average duration of an audio clip is less than 10 seconds). Without directly relying on the annotated vulnerability data to reduce product development cost, we propose a multi-task data-driven approach to detect the vulnerability through speech recordings by decomposing the task into a collection of sub-tasks, each of which can be solved by learning from publicly available data. Automatic Speech Recognition (ASR) techniques have been greatly developed in the recent decade such as Deep Speech [5], we use both the speech transcriptions and acoustic waves to support the detection. More specifically, we investigated acoustic and textual feature selection that can be used for classifying speakers by age (i.e. child, adult or older adult), accent (i.e. native English speaker or non-native English speaker) and health status (i.e. patient with commonplace neurological difficulties or non-patient).

Our main contributions can be summarised as follows:

- We develop a vulnerability detection system for short speech segments with transcription by indicating the speaker's age group, accent group and health status.
- We study the feature extraction to detect the vulnerable people from speech segments. We found using a combination of acoustic and textual features works better than one modality (i.e. either speech or text) in most cases.
- Unlike prior research on investigating each feature, we investigate all possible combinations of feature groups.
- We evaluate three sub models on multiple benchmark datasets. Limited data resources are publicly available for evaluating the patient model, we collected and annotated a set of patient/non-patient speech segments accompanied with transcription from YouTube [1].

2. Related Work

Prior works demonstrated the potential of identifying vulnerable people such as patients with dementia [6], aphasia [7] and older adults [3, 4] using speech-based approaches. Feature extraction is an essential step to traditional approaches and deep learning approaches [8]. With the acquired popularity in ASR, extracting textual features from speech recordings along the acoustic features for speech classification become more reliable. In this section, we review some acoustic and textual features that have been frequently used in the prior works and can be applied to detect the vulnerability. Fundamental Frequency (F0) is a common measure for age and gender detection. Women have a higher F0 compared to men, and children have a higher F0 compared to adults [9, 10]. Spectral features such as Mel-frequency Cepstral Coefficients (MFCCs), Filter Bank Energies (FBEs) and Spectral Centroid Coefficients (SCCs) are frequently used in a number of applications [3, 6]. Voicing features such as the duration and number of unvoiced segments [6], and voiced utterances [9, 11] have shown the effectiveness in detecting language disorders. Jitter is a measure of frequency instability whereas shimmer is a measure of amplitude instability [9]. They are frequently used to detect the fluctuation and perturbation in speech signal respectively [12].

[1]https://www.youtube.com/

http://dx.doi.org/10.21437/Interspeech.2021-1525

Harimonicity refers to Harmonic-to-Noise Ratio (HNR) and Noise-to-Harmonic Ratio (NHR) that measure the voice quality and are reported as a better measure for discriminating older adults and young people [13]. Mean of autocorrelation is another measure of voice quality estimating the pitch period of a given speech signal [6]. The Term Frequency-Inverse Document Frequency (TF-IDF) vectors are used to measure the repetitiveness by the cosine similarity between documents [14]. Part-Of-Speech (POS) features are represented by the frequency of various POS tags, such as interjections (i.e. filler words) were reported frequently in the use of detecting behaviour patterns and personality recognition [15]. Type Token Ratio (TTR) measures the weight of unique words in a document and shows the vocabulary richness (i.e. lexical diversity) of a document. A more advanced measure is moving-average type-token ratio (MATTR) [16], which computes the ratio by moving a fixed-size window within the document. Vulnerable people such as patients with memory problems and second language speakers are expected to have a lower TTR [11]. Psycholinguistic features were used for speech transcripts summarisation [17]. Older adults and people with certain health condition usually have memory problems. Several emotional categories from psycholinguistic features (e.g. depression, anxiety and stress) are often considered as causes for memory problems. In addition, the topical categories from psycholinguistic features can provide some insight to evidence on the speaker's life events.

3. Methods

We developed an automatic vulnerability detection system using a data-driven approach based on feature extraction and classification techniques. Below, we introduce the datasets and preprocessing (Section 3.1), feature extraction (Section 3.2) and classifier training details (Section 3.3).

3.1. Data

All sub-models were created and evaluated using features extracted from three English speaking TalkBank datasets (AphasiaBank [18], DementiaBank [19] and RHDBank [20]) and three large ASR datasets (Common Voice [2], VoxForge [3] and VCTK [4]). For ASR datasets such as Common Voice, we use the official validated subset. TalkBank datasets contain videos conducted and recorded by investigators and students, which are interviews with patients or people from health control group. The original video files were firstly converted into audio files via MoviePy [5]. Then, the audio files were trimmed into short clips by the timestamp and speaker label. In our scenario, models are created without any hand-crafted information, or probably based on the transcription from ASR. Therefore, the transcripts were downsampled. All hand-crafted information within the transcripts (e.g. timestamps for sub-sentences, POS tags, and manually-corrected words) were removed. Due to the recording devices, we found some audio files in the TalkBank are noisy, this was also reported in Al-hameed et al. [6]. Therefore, we used spectral gating [21] to reduce the stationary noise from the audio clips. Furthermore, we extracted available speaker information such as age, accent and gender for annotating the datasets. Depending on the model, we selected 1000 instances from each class to form a validation set for each model.

[2]https://voice.mozilla.org/

[3]http://www.voxforge.org/

[4]https://datashare.ed.ac.uk/handle/10283/3443

[5]https://zulko.github.io/moviepy/

3.2. Feature Extraction

The feature extractor plays an important role in the system. Two sets of features we extract from recording and transcription are shown in Table 2. We implemented an acoustic feature extractor using parselmouth [6] and librosa [7]. Following Al-hameed et al. [6] and Teixeira et al. [22], we extracted acoustic features including 2 F0 variants (mean and covariance), first 42 MFCCs and their skewness, kurtosis, mean with kurtosis and skewness of the mean, 26 FBEs, 26 SCCs, 5 pitch variants (mean, median, standard deviation, minimum and maximum), 4 pulses variants (number of pulses, number of periods, mean of and standard deviation of the periods), 3 voicing (fraction of locally unvoiced frames, number and degree of voice breaks), 5 jitter variants (local, local-absolute, the relative average perturbation, five-point perturbation quotient and the average absolute difference), 6 shimmer variants (local, local-dB, three point amplitude perturbation, five-point amplitude perturbation quotient ,eleven-point amplitude perturbation quotient and the average absolute difference) and 3 harmonicity variants (mean of the autocorrelation, NHR and NHR). We implemented a textual feature extractor using scikit-learn [8]. We extracted textual features including 3000 dimensional TF-IDF features, POS features, TTR and MATTR, psycholinguistic features and sentiment. We used the Universal POS tags [23] to form POS features, other POS tag marks such as Penn Treebank POS tags [9] can also be used. We use the pretrained Convolutional Neural Networks (CNN) based sentiment analyser from stanza [24] to produce the sentiment feature. We use Empath [25] to extract a vector of 200 lexical categories to form the topic and emotion features.

3.3. Training

We address the vulnerability detection problem as three classification tasks to find the related indicators from speech recordings and corresponding transcriptions. A collection of three separate classification models were created: an age model to classify the speaker's age group (below 20 as child, between 20 and 60 as adult, and over 60 as older adult), a non-native model to classify the speaker's accent group (native and non-native English speaker) and a patient model to classify the speaker's health status (patient with aphasia, dementia or RHD and non-patient).

The number of instances used for training each sub model is summarised as (a) age model: child (14,472), adult (18,162) and older adult (9,943); (b) non-native model: native (31,500) and non-native English speaker (31,500); (c) patient model: patient (7,000) and non-patient (7,000). Due to incomplete speaker information available in the six datasets, we use different subsets for training different model. We employ a simple data fusion technique to combine multiple data sources in training. We learn a weight w_i for each training dataset, where w_i maximises the prediction accuracy on a validation set. The weights are learned using Bayesian Optimisation [26].

The age model is trained on a combination of six datasets: Common Voice, VCTK, VoxForge, AphasiaBank, DementiaBank and RHDBank. Their audio clips and corresponding transcripts are categorised into three age groups. We found the datasets are strongly imbalanced, we adjusted the class weight for training. In addition to the original age model, we train

[6]https://parselmouth.readthedocs.io/en/stable/

[7]https://librosa.org/doc/latest/index.html

[8]https://scikit-learn.org/

[9]https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Table 1: *Average classification accuracy (acc) with standard deviation (std), train and test time (in seconds) over 5-fold cross-validation using various learning algorithms for the age model using a vaildation dataset from Common Voice.*

Classifier	acc±std	Train Time	Test Time
Nearest Neighbors	0.6190±0.0227	1.2412	1.1662
Decision Tree	0.6390±0.0332	2.5636	0.1759
Random Forest	0.7565±0.0131	1.5734	0.2663
MLP	0.6060±0.0883	17.5251	0.1973
AdaBoost	0.7070±0.0058	15.5314	0.3440
Naïve Bayes	0.6110±0.0312	0.8323	0.2026
QDA	0.5090±0.0394	8.5438	0.9927
Logistic Regression	0.6885±0.0333	1.6408	0.1595
Linear SVM with SGD	0.5565±0.0446	1.2600	0.1667

a variant with separating training data into gender-age groups (e.g. female child, elderly male) and then map the gender-age groups back to their age groups. The non-native model is trained on a combination of three datasets: Common Voice, VCTK and VoxForge, of which the audio clips and transcripts are categorised into the two groups of native and non-native English speakers. We define native speakers by a list of native English speaking countries [10]. Unlike prior works that studied individual long-term illness and tried to differentiate patients from people in the health control group, we explored the possibility to discriminate the patients from non-patients in a more general sense. We train a patient model that aims at discriminating patients and non-patients by combining the three TalkBank datasets. In this scenario, we mix the data that are labelled as patients from the three datasets and use them as positive instances for training. Negative training instances are randomly selected from Common Voice.

4. Experiments

We conducted experiments on classification (Section 4.1), feature selection (Section 4.2) and evaluated each sub model on multiple datasets (Section 4.3).

4.1. Classification

To choose a proper classifier, we experimented several learning algorithms: Nearest Neighbors, Decision Tree, Random Forest, Multi-layer Perceptron (MLP), AdaBoost, Naïve Bayes, Quadratic Discriminant Analysis (QDA), Logistic Regression and Linear Support Vector Machine (SVM) with Stochastic Gradient Descent (SGD). We randomly selected 1000 older adult and 1000 non-older adult examples from Common Voice to train a binary classifier. We used the classifier implementation from scikit-learn [11]. Table 1 shows the average results using different classifiers over 5-fold cross validation. Random Forest classifier has shown improvements in a couple of speech classification tasks such as speech/non-speech discrimination [27] and speech emotion recognition [28]. We observed Random Forest classifier also shows a promising performance in speech age group classification and reaches competitive time efficiency in both training and testing.

4.2. Feature Selection

We conducted a comprehensive study into feature extraction and selection. More specifically, we run 5-fold cross-validation on each model over all possible combinations (32,767 combinations in total for 15 feature groups). Table 3 shows the top

[10] https://www.gov.uk/english-language/exemptions

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

Table 2: *Features with their dimensionality (dim).* $(\cdot)$ *denotes the shorthand name for each feature.*

	Features	dim
Acoustic Features	Mel-frequency Cepstral Coefficients (mfcc)	506
	Filter Bank Energy (fbank_energy)	26
	Spectral Centroid Coefficients (spectral_centroid)	26
	Fundamental Frequency (f0)	2
	Pitch (pitch)	5
	Pulses (pulses)	4
	Vocing (vocing)	3
	Jitter (jitter)	5
	Shimmer (shimmer)	6
	Harmonicity (harmonicity)	3
Text Features	TF-IDF (tf_idf)	3000
	Part-of-Speech Tags (pos_counts)	17
	Type Token Ratio (ttr)	2
	Topic and Emotion (empath)	194
	Sentiment (sentiment)	1

5 feature combination candidates for the age model and the top combination candidates are sorted by classification accuracy descendingly. In contrast to the prior works using either text [29] or audio features [4, 30] for estimating the age, we observe that most of the top candidates are combinations of both text and audio features (4 out of 5). Table 4 shows the ablation study of feature selection, the classification accuracy falls around 0.02 when we remove the text features such as TTR and sentiment. SCCs improve the performance significantly (i.e. around 0.07). Furthermore, we rank each feature by its occurrence in the top 10 combination candidates. Table 5 shows the frequently-occurred candidate features in top 10 combinations, the first row is the most frequently-occurred candidate feature and we add others to the following rows by their occurrence in the top 10 combinations. The results indicate audio features take a major role in the feature extraction for the age model. Using FBEs alone gains a good performance on the classification accuracy (0.574). MFCCs are frequently used as a promising feature for the audio classification, however, we find using MFCCs alone achieves around 0.59 in accuracy, which is more computational costly (i.e. the dimensionality of MFCCs is 506) and does not perform as well as a combination of the other audio features with lower dimensionality. In our preliminary experiments, we found TF-IDF feature had a strong impact on the performance. TF-IDF usually fails if a test sentence contains many out-of-vocabulary words. To expand the feature space to overcome this issue, one of the possible solutions is to use word embeddings. Table 6, we compare the TF-IDF feature (tf_idf), topic and emotion feature (empath) with some popular word embeddings such as FastText (crawl and news) [31], Extended Dependency Skipgram (extvec) [32], GloVe (glove) [33], Skip-gram (twitter) [34] and Turian (turian) [35]. We use the implementation from flair [36] and each sentence is represented by a fix-length 100 dimensional embedding. The age model is a three-class classifier, both TF-IDF and word embeddings do not improve the classification accuracy significantly (i.e., close to 0.3333). Due to the page limit, we presented the results on one of the sub models, similar trend is also observed in the other models.

4.3. Model Evaluation

Table 9 shows the average classification accuracy of three sub models with an additional age model variant evaluated on multiple datasets. We first evaluate the trained age model on all six datasets. Table 7 shows the classification accuracy on each dataset's test set. In-dataset accuracy denotes the classification accuracy using the given dataset, and it is often considered as a

Table 3: *Top 5 feature combination candidates for the age model with the acc and std on 5-fold cross validation, sorted by the accuracy descendingly.*

Features	acc±std	dim
pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	0.662± 0.023	64
pitch + voicing + jitter + harmonicity + fbank_energy + spectral_centroid + f0 + ttr	0.660± 0.019	72
pitch + fbank_energy + spectral_centroid + sentiment	0.660± 0.014	58
pitch + pulses + harmonicity + fbank_energy + spectral_centroid + f0	0.660±0.017	66
pitch + harmonicity + fbank_energy + spectral_centroid + f0	0.659±0.018	62

Table 4: *Ablation study of the feature combination (age model).*

Features	acc±std	dim
pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	0.662±0.023	64
(-ttr)	0.648±0.021	62
(-sentiment)	0.642±0.016	63
(-f0)	0.650±0.020	62
(-spectral_centroid)	0.597±0.019	38
(-fbank_energy)	0.624±0.017	38
(-harmonicity)	0.647±0.023	61
(-pulses)	0.646±0.027	60

Table 5: *Frequently-occurred candidate features (age model).*

Features	acc±std	dim
fbank_energy	0.574±0.013	26
fbank_energy + spectral_centroid	0.623±0.023	52
fbank_energy + spectral_centroid + harmonicity	0.636±0.010	55
fbank_energy + spectral_centroid + harmonicity + f0	0.648±0.024	57
fbank_energy + spectral_centroid + harmonicity + f0 + pitch	0.659±0.018	62
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses	0.660±0.017	66
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment	0.648±0.014	67
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr	0.653±0.012	69
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr + voicing	0.637±0.024	72
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr + voicing + jitter	0.655±0.014	77

Table 6: *Evaluation on embedding features (age model).*

Embedding	crawl	extvec	glove	news
acc+std	0.3647±0.0266	0.3690±0.0191	0.3643±0.0087	0.3473±0.0302
Embedding	turian	twitter	tf_idf	empath
acc+std	0.3643±0.0284	0.3507±0.0203	0.3530±0.0128	0.3393±0.0182

Table 7: *Classification accuracy tested on various datasets for the age model. #test denotes the number of test examples.*

Dataset	#test	In-dataset	Data Fusion	Data Fusion + Gender Separation
Common Voice	9000	0.7249	0.7414	0.7460
VoxForge	3579	0.8178	0.8128	0.8268
VCTK	4000	0.9423	0.9398	0.9480
AphasiaBank	1099	0.6016	0.6497	0.6261
DementiaBank	73	0.8493	0.7808	0.7945
RHDBank	772	0.9391	0.9443	0.9313

strong baseline for evaluating the model generalisation. Common Voice and VoxForge contain data from all three age groups. We observe a slight improvement by data fusion and gender separation compared to the in-dataset accuracy. VCTK does not contain any data from older adult class and the age range is narrow (speakers are 18 to 30 years old). In this case, a binary in-dataset classifier is trained. AphasiaBank, DementiaBank and RHDBank have a similar situation that there is no or few data from the child class and the age range is close to the pre-defined boundary. We observe the proposed model still performs competitively under this challenging condition. In general, by using data fusion to introduce additional data sources, a few improvements can be observed in the classification accuracy across these datasets. With the help of gender separation, 4 out of 6 datasets perform slightly better than the original age model. Table 8 shows the classification accuracy for the non-native model tested on ASR datasets. Common Voice contains a large number of non-native speakers that are Indian, whereas VoxFroge contains a large number of non-native speakers are European. VCTK was claimed as a native English speaker dataset in the original publication. However, we found it contains one speaker from India. Considering the diversity of the three datasets, the proposed data fusion model is relatively robust that the test accuracy is slightly lower than the in-dataset accuracy. For evaluating the patient model, we retrieved 59 videos from YouTube using some patient related keywords: *aphasia+example*, *dementia+example*, *mental+ill+patient* and *patient+voices+nhs*. "+" denotes an AND relation in a search query. We also retrieved the associated transcription. We converted these videos into audio and trim them into short clips by the timestamp in the transcription. However, no speaker information is available for this dataset. We randomly select 510 audio clips from this dataset and manually annotate them as patient (124) or non-patient (386) voice clips. Due to the transcription quality, this model was evaluated under a challenging noisy condition and the results can be treated as a baseline for future development. The patient model achieved a classification accuracy of 0.684 (Table 9). We observed a relatively low false positive rate (0.1891) but a low true positive rate (0.2903).

Table 8: *Classification accuracy tested on various datasets for the non-native model.*

Dataset	#test	In-dataset	Data Fusion
Common Voice	18000	0.8066	0.8042
VoxForge	6000	0.860	0.8584
VCTK	3000	0.967	0.9553

Table 9: *Average classification accuracy evaluated on multiple datasets for all trained models with top feature combination and dimensionality.*

Model	acc	Top Feature Combination	dim
Age (+Gender Separation)	0.8121	jitter + shimmer + fbank_energy + spectral_centroid + f0 + sentiment + ttr	64
Age	0.8115	pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	68
Non-Native	0.8726	pitch + voicing + jitter + shimmer + fbank_energy + spectral_centroid	71
Patient	0.6840	shimmer + harmonicity + mfcc + fbank_energy + spectral_centroid + f0 + pos_counts + ttr	588

5. Conclusion and Future Work

We studied the features extracted from short speech segments and their transcription. We address the detection problem by dividing it into three separate classification tasks: age classification, accent classification and patient/non-patient classification. We trained an age model, a non-native model and a patient model respectively. We evaluated the age and non-native models on multiple benchmark datasets. The patient model was evaluated on a manually-annotated dataset collected from YouTube. We presented a data-driven approach to address the vulnerability detection problem. The models were trained using supervised learning algorithms on extracted features. We plan to extend this work using semi-supervised learning and pre-trained deep learning models for speech to reduce the number of labelling data required for training. In the future, we will adapt the vulnerability detection system to the practice.

6. References

[1] "Consumer vulnerability," Financial Conduct Authority (FCA), Tech. Rep. 8, Feb. 2015.

[2] "Consumer vulnerability: challenges and potential solutions," Competitions & Markets Authority (CMA), Tech. Rep., Feb. 2019.

[3] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[4] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[6] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for alzheimer's disease," in *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 32–36.

[7] A. Shivkumar, J. Weston, R. Lenain, and E. Fristed, "Blabla: Linguistic feature extraction for clinical analysis in multiple languages," *arXiv preprint arXiv:2005.10219*, 2020.

[8] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[9] C. L. Lortie, M. Thibeault, M. J. Guitton, and P. Tremblay, "Effects of age on the amplitude, frequency and perceived quality of voice," *Age*, vol. 37, no. 6, p. 117, 2015.

[10] A. Heidari, A. Moossavi, F. Yadegari, E. Bakhshi, and M. Ahadi, "Effects of age on speech-in-noise identification: subjective ratings of hearing difficulties and encoding of fundamental frequency in older adults," *Journal of audiology & otology*, vol. 22, no. 3, p. 134, 2018.

[11] 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.

[12] M. L. B. Pulido, J. B. A. Hernández, M. Ángel Ferrer Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems with Applications*, vol. 150, p. 113213, 2020.

[13] C. T. Ferrand, "Harmonics-to-noise ratio: an index of vocal aging," *Journal of voice*, vol. 16, no. 4, pp. 480–487, 2002.

[14] V. Masrani, G. Murray, T. S. Field, and G. Carenini, "Domain adaptation for detecting mild cognitive impairment," in *Canadian Conference on Artificial Intelligence*. Springer, 2017, pp. 248–259.

[15] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 955–959.

[16] M. Covington and J. D. McFall, "The moving-average type-token ratio," *Linguistics Society of America, Chicago, IL*, 2008.

[17] S. K. Barnwal and U. S. Tiwary, "Using psycholinguistic features for the classification of comprehenders from summary speech transcripts," in *Intelligent Human Computer Interaction*, P. Horain, C. Achard, and M. Mallem, Eds. Cham: Springer International Publishing, 2017, pp. 122–136.

[18] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[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, "Understanding spoken language through talkbank," *Behavior research methods*, vol. 51, no. 4, pp. 1919–1927, 2019.

[21] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[22] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[23] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *Proceedings of the International Conference on Language Resources and Evaluation*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2089–2096.

[24] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of ACL: System Demonstrations*, 2020.

[25] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4647–4657.

[26] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *NIPS*. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 2951–2959.

[27] S. V. Thambi, K. Sreekumar, C. S. Kumar, and P. R. Raj, "Random forest algorithm for improving the performance of speech/non-speech detection," in *2014 First International Conference on Computational Systems and Communications (ICCSC)*. IEEE, 2014, pp. 28–32.

[28] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[29] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10 805–10 816, 2017.

[30] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors." in *INTERSPEECH*, 2016, pp. 1402–1406.

[31] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[32] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *NAACL-HLT*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1490–1500.

[33] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[34] 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.

[35] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 384–394.

[36] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

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

Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition

Jiajun Deng[1*], *Fabian Ritter Gutierrez*[1*], *Shoukang Hu*[1], *Mengzhe Geng*[1], *Xurong Xie*[2], *Zi Ye*[1], *Shansong Liu*[1], *Jianwei Yu*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

{jjdeng,ritter,skhu,mzgeng,zye,ssliu,jwyu,xyliu,hmmeng}@se.cuhk.edu.hk xr.xie@siat.ac.cn

Abstract

Automatic recognition of elderly and disordered speech remains a highly challenging task to date. Such data is not only difficult to collect in large quantities, but also exhibits a significant mismatch against normal speech trained ASR systems. To this end, conventional deep neural network model adaptation approaches only consider parameter fine-tuning on limited target domain data. In this paper, a novel Bayesian parametric and neural architectural domain adaptation approach is proposed. Both the standard model parameters and architectural hyper-parameters (hidden layer L/R context offsets) of two lattice-free MMI (LF-MMI) factored TDNN systems separately trained using large quantities of normal speech from the English LibriSpeech and Cantonese SpeechOcean corpora were domain adapted to two tasks: a) 16-hour DementiaBank elderly speech corpus; and b) 14-hour CUDYS dysarthric speech database. A Bayesian differentiable architectural search (DARTS) supernetwork was designed to allow both efficient search over up to 7^{28} different TDNN structures during domain adaptation, and robust modelling of parameter uncertainty given limited target domain data. Absolute recognition error rate reductions of 1.82% and 2.93% (13.2% and 8.3% relative) were obtained over the baseline systems performing model parameter fine-tuning only. Consistent performance improvements were retained after data augmentation and learning hidden unit contribution (LHUC) based speaker adaptation was performed.
Index Terms: speech recognition, domain adaptation, Bayesian learning, neural architecture search

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) technologies targeting normal speech [1–5] in the past few decades, accurate recognition of atypical speech task domains represented by, for example, elderly and dysarthric speech, remains a highly challenging task to date [6–9].

Ageing presents enormous challenges to health care and current speech technologies. Neurocognitive disorders (NCDs), such as Alzheimer's disease (AD), are often found among older adults [10] and manifest themselves in speech and language impairments including weakened neuro-motor control in speech production and imprecise articulation [11, 12]. Speech disorders such as dysarthria can also be caused by a range of other conditions including cerebral palsy, amyotrophic lateral sclerosis, stroke or traumatic brain injuries [13]. People with speech impairment often experience co-occurring physical disabilities and mobility limitations.

∗ Equal contribution

Elderly and dysarthric speech exhibit a wide spectrum of challenges for current deep neural networks (DNNs) based ASR technologies that predominantly target normal speech. First, a large mismatch between such data and non-aged, healthy adult voice is often observed. Such difference manifests itself across many fronts including articulatory imprecision, decreased volume and clarity, changes in pitch, increased dysfluencies and slower speaking rate [14, 15]. State-of-the-art ASR systems designed for normal speech often produce very high recognition error rate above 40% when being applied to elderly or impaired speech [9, 16, 17]. Second, the co-occurring disabilities, mobility or accessibility limitations often found among elderly and disordered speakers lead to the difficulty in collecting large quantities of such data that are essential for current data intensive deep learning based ASR system development.

To this end, a range of techniques designed to address the above domain mismatch and data sparsity issues have been studied in recent years primarily in the context of dysarthric speech recognition. Motivated by the spectral-temporal level differences of disordered speech from normal speech such as slower speaking rates, recent research in data augmentation has been largely focused on tempo-stretching [18], vocal tract length perturbation (VTLP) [19], and speed perturbation [20] of normal speech recorded from healthy control speakers. The resulting "disordered like" speech carrying a slower speaking rate and modified overall vocal tract spectral shape is then used to augment the limited dysarthric speech training data. Alternative approaches based on cross-domain DNN model or feature adaptation [21–23], domain adversarial training [24], transfer learning [25, 26], knowledge distillation [27], and voice conversion [28, 29] have also been investigated.

Among the above, model based domain adaptation approaches benefit not only from a tight integration of domain dependently estimated parameters with the underlying speech recognition error cost based on, for example, the lattice-free maximum mutual information (LF-MMI) criterion [25], or sequence to sequence learning objective functions, for example, used in recurrent neural network (RNN) transducers [22], but also fine modelling granularity in adapted parameters when sufficient target domain data is available.

However, there are two issues associated with model based domain approaches when being applied to elderly or disordered speech recognition tasks. First, due to the difficulty in collecting large quantities of such data, and the often limited amounts of existing elderly [30] or dysarthric speech datasets [31], direct fine-tuning of large numbers of out of domain, normal speech data estimated DNN model parameters on limited elderly or dysarthric speech data is generally problematic. The severe data sparsity issue and the resulting modelling uncertainty need to be

http://dx.doi.org/10.21437/Interspeech.2021-289

addressed. Second, the underlying neural architecture designs in current ASR systems are often designed using expert knowledge and empirical evaluation within individual task domains, for example, conversational telephone speech [32], or meeting transcription [33]. For example, the left and right splicing context offsets in the hidden layers of state-of-the-art LF-MMI trained time delay neural network (TDNN) systems [3, 34] represent the range of temporal contexts that can be exploited in modelling. The DementiaBank Pitt corpus [30], the largest publicly available elderly speech database, contains 4.8 words per utterance on average, in contrast to the normal speech data from the LibriSpeech corpus [35] of approximately 31 words per utterance. Similar designs based on shorter sentences also feature in current dysarthric speech corpora [31].

In order to address these issues, a novel Bayesian parametric and neural architectural domain adaptation approach is proposed in this paper. Both the standard model parameters and architectural hyper-parameters (hidden layer left and right context offsets[1]) of two LF-MMI factored TDNN systems separately trained using large quantities of normal speech from the English LibriSpeech and Cantonese SpeechOcean corpora were domain adapted to two tasks: a) 16-hour DementiaBank elderly speech corpus; and b) 14-hour CUDYS dysarthric speech database. Bayesian learning of differentiable architectural search (DARTS) [38] super-network was employed to allow both efficient search over up to 7^{28} different TDNN structures during domain adaptation, and robust modelling of parameter uncertainty given limited target domain data. Absolute recognition error rate reductions of 1.82% and 2.93% (13.2% and 8.3% relative) were obtained over the baseline systems performing model parameter fine-tuning only. Consistent performance improvements were retained after data augmentation and learning hidden unit contribution (LHUC) based speaker adaptation was performed. To the best of our knowledge, this is the first work to consider both parametric and architectural cross-domain adaptation for elderly and dysarthric speech recognition. In contrast, the majority of previous researches on domain adaptation for the same tasks have been focused on direct parameter fine-tuning [22, 23, 25, 26] while the data sparsity and architecture mismatch issues remain unsolved.

The rest of this paper is organized as follows. Section 2 presents Bayesian domain adaptation of LF-MMI trained TDNN systems. A novel differentiable architecture search approach automatically learning the L/R context offsets hyper-parameters of Bayesian TDNN systems is proposed in Section 3. Section 4 presents the experiments and results. Finally, the conclusions are drawn in Section 5.

2. Bayesian TDNN Adaptation

In contrast to conventional model adaptation methods performing fixed-value, deterministic parameter fine-tuning given limited target domain data, Bayesian adaptation approaches address the data sparsity issue by modelling parameter uncertainty using the following predictive distribution. Given an adaptation data set $\mathcal{D} = \{\mathbf{O}_r, \mathbf{H}_r\}$, where $\mathbf{O}_r$ and $\mathbf{H}_r$ are the r-th speech utterance and the reference word sequences, respectively. The prediction over the r-th test utterance $\mathbf{O}_r^*$ is given by

$$p(\mathbf{H}_r^*|\mathbf{O}_r^*, \mathcal{D}) = \int p(\mathbf{H}_r^*|\mathbf{O}_r^*, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \quad (1)$$

where $\mathbf{H}_r^*$ denotes the predicted word sequence for the test utterance r, $\mathbf{w}$ is the Bayesian adaptation parameters and $p(\mathbf{w}|\mathcal{D})$ is its posterior distribution learned from the adaptation data.

LF-MMI Trained TDNNs: TDNNs [39] produced state-of-art performance on different tasks [34, 40, 41]. TDNN is an instance of 1-dimension convolutional neural networks with parameters tying over different time steps. The lower TDNN layers are designed to learn narrower, local temporal contexts, while the higher layers learn wider, longer range contexts. The TDNN hidden left and right splicing context offsets are important hyper-parameters controlling its hierarchical temporal modelling ability. This paper adopted the factored TDNN [40].

In contrast to the conventional cross entropy criterion, sequence level error costs more closely related recognition accuracy, for example, the MMI [1] criterion, is widely used in state-of-the-art ASR systems [3, 42, 43].

$$\mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}) = \sum_r \log \frac{p(\mathbf{O}_r|\mathbf{H}_r)^\kappa P(\mathbf{H}_r)}{\sum_{\hat{\mathbf{H}}_r} p(\mathbf{O}_r|\hat{\mathbf{H}}_r)^\kappa P(\hat{\mathbf{H}}_r)} \quad (2)$$

where $\boldsymbol{\Theta}$ contains both hyper-parameters such as hidden layer context offsets and normal TDNN weight parameters, κ is the acoustic scaling factor and $\hat{\mathbf{H}}_r$ is the possible word sequence in the decoded speech lattice for utterance r. The efficient lattice-free MMI training [3] that alleviates the explicit denominator lattice generation is considered in this paper.

Bayesian TDNN Model Adaptation: During domain adaptation, the parameter posterior distribution $p(\mathbf{w}|\mathcal{D})$ required in the form of Bayesian prediction in Eqn. (1) can be learned by maximising the following MMI criterion marginalisation over all parameter estimates.

$$\mathcal{F} = \log \int \exp\{\mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta})\} P_r(\mathbf{w}) d\mathbf{w} \quad (3)$$

where $\mathbf{w} \in \boldsymbol{\Theta}$ and $P_r(\mathbf{w})$ is the prior distribution of adaptation parameters. Direct optimisation of the above integral is nontrivial. An alternative more efficient variational inference is utilized to learn the adaptation parameter posterior distribution by optimising the following lower bound,

$$\begin{aligned} \mathcal{F} &\geq \int q(\mathbf{w}) \mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}) d\mathbf{w} - KL(q(\mathbf{w})||P_r(\mathbf{w})) \\ &= \mathcal{L}_1^{MMI} - \mathcal{L}_2^{MMI} = \mathcal{L}^{MMI} \end{aligned} \quad (4)$$

where $q(\mathbf{w})$ is the variational approximation of the posterior distribution $p(\mathbf{w}|\mathcal{D})$ and $KL(q(\mathbf{w})||P_r(\mathbf{w}))$ is the Kullback-Leibler (KL) divergence between $q(\mathbf{w})$ and $P_r(\mathbf{w})$. For efficiency, and based on the previous research findings [41], both $q(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and $P_r(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu_r}, \boldsymbol{\sigma_r^2})$ are assumed to be Gaussian distributions. The first term $\mathcal{L}_1^{MMI}$ is approximated with Monte Carlo sampling method, which is given by

$$\mathcal{L}_1^{MMI} \approx \frac{1}{N} \sum_{k=1}^{N} \mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon_k}) \quad (5)$$

where ϵ_k is the k-th Monte Carlo sampling value drawn from the standard normal distribution $\mathcal{N}(0, 1)$ and $\odot$ is the Hadamard product. The KL divergence based second term $\mathcal{L}_2^{MMI}$ in Eqn. (4) can be explicitly calculated as

$$\mathcal{L}_2^{MMI} = \frac{1}{2} \sum_i \left(\frac{\sigma_i^2 + (\mu_i - \mu_{r,i})^2}{\sigma_{r,i}^2} + 2 \log \frac{\sigma_{r,i}}{\sigma_i} - 1 \right) \quad (6)$$

[1] Prior researchers suggested [36, 37] that TDNN context offset settings significantly affect the resulting system's temporal modelling resolution and recognition performance, while other hyper-parameters, e.g. the hidden layer dimensionality, were used to control the overall system complexity, thus not considered here.

where $\{\mu_{r,i}, \sigma_{r,i}\}$ and $\{\mu_i, \sigma_i\}$ are the i-th hyper-parameters of the prior distribution $\{\boldsymbol{\mu_r}, \boldsymbol{\sigma_r}\}$ and variational distribution $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$, respectively. The variational distribution parameters are updated during back propagation.

Implementation Details over several crucial settings are: 1) The first TDNN layer parameters practically exhibit more uncertainty due to the larger input data variability than those observed at higher layers designed to produce more invariant features. Based on our previous findings [41, 44], Bayesian domain adaptation was applied to the first layer of all TDNN systems in this paper, while the other higher layers parameters were fine-tuned to the target domain data. 2) The prior for all Bayesian adapted TDNN systems is based on the comparable fully converged standard fixed-parameter fine-tuning adapted TDNN systems. Other parameters in the Bayesian adapted TDNN systems are initialized using those of the halfway fine-tuned TDNN systems during adaptation. 3) The variational distribution variance is shared among all nodes of the first layer, which allows the number of parameters in Bayesian adapted TDNN system to be comparable to that of the standard fixed-parameter adapted system. 4) For efficiency, only one parameter sample is drawn in Eqn. (5) to ensure the computational cost in Bayesian adaptation to be comparable to that of the standard fine-tuning adapted TDNN system. During recognition time, the predictive inference integral in Eqn. (1) is efficiently approximated by the expectation of Bayesian adapted TDNN model parameters.

3. TDNN Architecture Adaptation

The general problem of TDNN hyper-parameter domain adaptation is transformed into a domain adaptive neural architecture search [36, 45] task within the DARTS [38] framework that allows both the architecture hyper-parameters and TDNN, or Bayesian TDNN parameters to be optimized consistently during adaptation to elderly or disordered speech data. An over-parameterized super-network containing paths connecting all neural architecture candidates is trained first, before the selection weights over each neural architecture candidate within the super-network are learned in the search stage. On convergence of the super-network model, the optimal architecture is obtained by pruning lower weighted paths. For example, the output $\mathbf{h}^l$ of l-th layer in the DARTS super-network is given by

$$\mathbf{h}^l = \sum\nolimits_{i=0}^{N^l-1} \lambda_i^l \phi_i^l(\mathbf{W}_i^l \mathbf{h}^{l-1}) \tag{7}$$

where N^l denotes the number of architecture candidate selections in the l-th layer and λ_i^l is the weight of the i-th architecture candidate in the l-th layer. $\mathbf{W}_i^l$ and ϕ_i^l are the linear transformation parameter matrix and activation function for the i-th candidate system in the l-th layer, respectively.

Pipelined Gumbel-Softmax DARTS: In order to minimize the confusion between different architectures found in conventional Softmax based DARTS [38], a Gumbel-Softmax distribution [46] is used to produce approximately a one-hot vector, categorical architecture weights as the following

$$\lambda_i^l = \frac{\exp((\log \alpha_i^l + G_i^l)/T)}{\sum_{j=0}^{N^l-1} \exp((\log \alpha_j^l + G_j^l)/T)} \tag{8}$$

where α_i^l is the parameter in the Gumbel-Softmax distribution, $G_i^l = -\log(-\log(U_i^l)))$ is the Gumbel variable, U_i^l is a random variable sampled from a uniform distribution. T is the temperature hyper-parameter annealed from 1 to 0.03 in this paper.

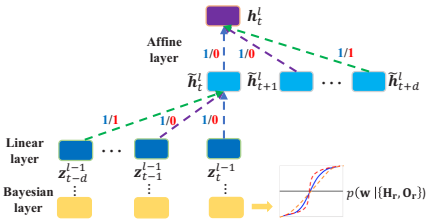


Figure 1 *Example DARTS super-network for Bayesian TDNNs (Bayesian layer in yellow square). Dashed lines in different colors are different Left/Right context offsets. The blue integers denote the super-network system using all context offsets, while the red integers represent a candidate offset choice of* $\pm$ *2.*

Following [36, 46], the update of TDNN parameters and architecture weights were performed in two stages, in a pipelined fashion, to avoid sub-optimal selection of architectures. In order to prevent overfitting to the training data, a separate held-out data set taken out of the original training data is used. In the first stage, the TDNN parameters are updated to convergence using the training data first, while randomly sampled one-hot architecture weights drawn from a uniform distribution are used in back-propagation. In the second stage, the TDNN parameters estimated in the first stage in the super-network are fixed and the architecture weights are updated using the held-out data.

TDNN-F Context Offset Search Space: The context offsets of TDNNs are crucial for modeling long temporal information in speech. Manually setting these hyper-parameters by evaluating a large number of possible system configurations is impractical. To this end, parameter sharing within the super-network can be used [47]. As shown in Fig. 1, all possible choices of context offsets to the left $(\{\text{-}d, 0\}, \cdots, \{1, 0\}, \{0, 0\})$ and right $(\{0, 0\}, \{0, 1\}, \cdots, \{0, d\})$ at each layer are incorporated into the TDNN-F super-network. The super-network designed for a L hidden layers TDNN contains $(d{+}1)^{2L}$ possible candidate models, each of which is indicated by setting the corresponding connecting weights as 1, while others as 0.

Architectural and Parametric Adaptation of Bayesian TDNNs is performed in three stages: a) **Architecture adaptation** is performed by first constructing a Bayesian TDNN super-network shown in Fig. 1 that contains all possible hidden layer context offset settings using the source domain data alone, before being adapted to the target domain. In this process, the very large number of standard TDNN parameters, often in tens of millions, are Bayesian adapted to ensure robustness on limited target domain data as described in Section 2, while the comparatively much smaller number of architecture selection weights, $2L(d{+}1)$ in total, linearly related to the number of hidden layers L and maximum context offset d, are fine-tuned during adaptation. b) **Architecture search** performed over the resulting domain adapted Bayesian TDNN super-network will then be searched over to produce the 1-best TDNN context offset settings. c) **Bayesian model adaptation** is finally performed by first constructing a TDNN system that features the above adapted architecture configurations but uses the source domain training data. The standard model parameters of this prior TDNN system are then further adapted in a Bayesian fashion to the target domain speech to produce the full architecture plus parameter adapted system.

4. Experiments

The proposed Bayesian parametric and architectural domain adaptation approach was investigated on two tasks for LF-

MMI factored TDNN systems: 1) from the English LibriSpeech speech corpus to DementiaBank elderly database; 2) from the Cantonese SpeechOcean corpus to the CUDYS dysarthric data.

Experimental Setup: The data sets and the baseline systems used in the two adaptation tasks are described below.

English Elderly Adaptation Task: A 1000 Hour LibriSpeech data set [35] is adopted as the source domain data. The DementiaBank database [30] is the target domain data, which includes 15.74-hour training set (9.72-hour elderly participant and 6.03-hour investigator data) and 3.14-hour test set (1.93-hour elderly participant and 1.21-hour investigator data) after silence stripping [8]. The word and duration per utterance on average in LibriSpeech (DementiaBank) corpus are 31 (4.8) and 11.3 (1.9) second, respectively. The training set was expanded to 59 hours when speed perturbation was performed. A 4-gram language model described in [8] was used.

Cantonese Dysarthric Adaptation Task: A Cantonese CUDYS dysarthric speech corpus [31] containing a 14.09-hour training set and 3.61-hour testing set with low and high intelligibility groups after silence stripping and speed perturbation [19] is utilized as the target domain data. 19.77-hour external data extracted from 163-hour Cantonese SpeechOcean normal speech corpus was mixed with 14.09-hour CUDYS training set for source-domain acoustic model training. The word and duration per utterance on average in the SpeechOcean (CUDYS) data set are 9.3 (1.6) and 4.0 (1.6) second, respectively. A 80k word 4-gram language model in [19] was used in recognition.

Baseline TDNN Systems: For the two domain adaptation tasks, LF-MMI factored TDNNs of 14 (English) and 7 (Cantonese) hidden layers were used[2], with the GMM-HMM system configuration the same as [8]. 40-dim filter-bank input features were used in both tasks. 100-dim i-vector features were appended for Librispeech and DementiaBank systems, while 3-dim pitch features were used for Cantonese SpeechOcean and CUDYS systems. For both tasks, the Bayesian architecture and parameter adaptation procedure in Section 3 was performed[3].

Performance of English Elderly DementiaBank: Table 1 demonstrates the performance of the DementiaBank corpus. Several trends are observed. First, the systems considering both architectural and parametric adaptation (Sys. 5, 6) outperform the corresponding systems only considering parameter adaptation (Sys. 3, 4) by up to **2.55%** absolute word error rate (WER) reductions. Second, further improvement by **0.38%** absolute WER reduction was obtained in the Bayesian architectural and parametric adapted systems (Sys. 6) over the corresponding architectural and parametric adapted systems without Bayesian estimation (Sys. 5). Finally, consistent performance improvements were retained after data augmentation and LHUC based speaker adaptation. In the cross-domain adapted systems, the largest absolute WER reduction up to **2.93%** was achieved by the Bayesian parametric and architectural adapted system (Sys. 6) over the parameter fine-tuning system (Sys. 3).

Table 1 *WERs (%) of TDNN systems trained using LibriSpeech or DementiaBank data alone, before domain adaptation of model parameters and optionally architecture (context offsets) w/o Bayesian estimation. (a,b) in the "context offsets" column denotes the context offsets {-a, 0} to left and {0, b} to right. † denotes a statistically significant difference obtained over the parametric fine-tuning baseline system (Sys. 3, 8, 13).*

Sys.	Data sets	Domain adaptation Arch.	Domain adaptation Para.	Context offsets 1-th to 14-th layer	Data aug.	LHUC SAT	DEV. PAR	DEV. INV	Eval. PAR	Eval. INV	ALL
1	LIBRI	✗	✗	1,1 1,1 0,0 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3	✗	✗	-	-	-	-	99.59
2	DEMEN	✗	✗	same as Sys. (1)	✗	✗	51.16	22.01	38.78	21.53	36.34
3	LIBRI → DEMEN	✗	FineTune				49.70	20.77	38.97	21.31	35.28
4		✗	Bayes [8, 41]				47.48	20.01	36.72	19.09	33.65†
5		DARTS	FineTune	5,0 0,4 0,2 0,4 5,4 2,6 0,6 0,6 0,5 6,5 6,5 6,6 6,5 6,6			46.45	**19.16**	35.78	**18.87**	32.73†
6		Bayes	Bayes	1,1 4,3 5,3 5,2 4,3 6,2 4,4 4,5 4,5 4,6 5,6 5,6 6,6 6,6			**45.31**	19.86	**34.35**	19.53	**32.35**†
7	DEMEN	✗	✗	same as Sys. (1)	✓	✗	46.94	20.06	36.97	19.98	33.53
8	LIBRI → DEMEN	✗	FineTune				46.91	19.29	36.64	20.09	33.15
9		✗	Bayes[8, 41]				45.90	19.84	35.15	19.53	32.71
10		DARTS	FineTune	0,6 0,5 5,5 4,5 6,6 6,5 0,6 0,6 0,6 0,6 0,5 0,6 0,6 6,5			45.25	**18.94**	35.46	21.09	32.19†
11		Bayes	Bayes	1,1 4,6 4,5 4,4 4,4 3,3 3,6 4,3 4,5 6,6 6,6 6,6 6,5 6,6			**44.56**	19.66	**33.68**	**17.87**	**31.81**†
12	DEMEN	✗	✗	same as Sys. (1)	✓	✓	44.95	18.52	35.33	**17.54**	31.77
13	LIBRI → DEMEN	✗	FineTune				44.16	19.12	34.16	19.42	31.56
14		✗	Bayes[8, 41]				44.08	19.11	34.22	18.87	31.52
15		DARTS	FineTune	same as Sys. (10)			43.75	**18.37**	33.84	19.53	31.04†
16		Bayes	Bayes	same as Sys. (11)			**43.36**	19.07	**32.08**	17.98	**30.83**†

[2]The DARTS systems perform the search over 7^{28} (English) and 7^{14} (Cantonese) TDNN-F choices with the maximum contexts of ± 6.

[3]A matched pairs sentence-segment word error based statistical significance test was performed at a significance level α=0.05.

Performance of Cantonese Dysarthric CUDYS: Results conducted on the CUDYS corpus are presented in Table 2 with similar trend to the DementiaBank task, absolute character error rate (CER) reductions of up to **1.61%** were obtained in the systems considering both architectural and parametric adaptation (Sys. 6, 7) over the corresponding parameter fine-tuning adapted systems (Sys. 4, 5). Second, the Bayesian architectural and parametric adapted systems (Sys. 7, 12) perform the best among other systems before and after speaker adaptation. In the cross-domain systems, the greatest absolute CER reduction up to **1.82%** was obtained by the Bayesian parametric and architectural adapted TDNN system (Sys. 7) over the parameter fine-tuning TDNN system (Sys. 4).

Table 2 *CERs (%) of TDNN systems trained using SpeechOcean or CUDYS data alone, before domain adaptation of model parameters and optionally architecture (context offsets) w/o Bayesian estimation. † denotes a statis. sig. diff. obtained over the parametric fine-tuning baseline system (Sys. 4, 9).*

Sys.	Data sets	Domain adaptation Arch.	Domain adaptation Para.	Context offsets 1-th to 7-th layer	LHUC SAT	DEV. High	DEV. Low	Eval. High	Eval. Low	ALL
1	SPOC CUDY	✗	✗	1,1 1,1 0,0 3,3 3,3 6,6 6,6	✗	32.92	98.51	14.24	94.97	36.12
2						9.94	88.09	1.30	85.89	19.80
3	SP. & CU. → CUDY	✗	✗	same as Sys. (1)	✗	4.97	85.32	**0.72**	70.36	15.37
4		✗	FineTune			5.67	79.47	1.02	57.14	13.74
5		✗	Bayes			**4.13**	74.68	0.90	52.30	12.22†
6		DARTS	FineTune	5,6 6,6 6,6 5,6 6,6 6,5 6,6		5.14	71.17	1.17	**48.88**	12.13†
7		Bayes	Bayes	6,4 5,6 6,6 6,6 6,6 6,6 6,6		4.77	**66.06**	1.49	49.05	**11.92**†
8	SP. & CU. → CUDY	✗	✗	same as Sys. (1)	✓	1.85	76.91	0.60	63.12	12.74
9		✗	FineTune			1.91	75.74	0.44	52.07	11.15
10		✗	Bayes			**1.15**	71.81	0.44	50.92	10.51†
11		DARTS	FineTune	same as Sys. (6)		2.01	**67.87**	**0.40**	45.49	9.96†
12		Bayes	Bayes	same as Sys. (7)		1.51	69.47	0.69	**45.51**	**9.41**†

5. Conclusions

The paper proposed a Bayesian parametric and neural architectural domain adaptation approach to rapidly port LF-MMI trained TDNNs based state-of-the-art ASR systems developed using large amounts of normal speech data to elderly and disordered speech task domains of more limited quantities. Experimental results suggest Bayesian adaptation can effectively mitigate the risk of overfitting when directly cross domain fine-tuning systems containing a large number of parameters. Architecture adaptation can further improve the generalization of systems using parameter adaptation only. Future research will focus on the adaptation of more advanced neural architectures.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200218, 14200220, TRS T45-407/19N, ITF grant No. ITS/254/19, and SHIAE grant No. MMT-p1-19.

7. References

[1] L. Bahl, P. Brown, P. De Souza *et al.*, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP*, 1986.

[2] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[3] D. Povey, V. Peddinti, D. Galvez *et al.*, "Purely sequence-trained neural networks for asr based on lattice-free MMI," in *INTERSPEECH*, 2016.

[4] W. Chan, N. Jaitly, Q. Le *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[5] Y. Wang, A. Mohamed, D. Le, C. Liu *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP*, 2020.

[6] H. Christensen, S. Cunningham, C. Fox *et al.*, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012.

[7] J. Yu, X. Xie, S. Liu *et al.*, "Development of the CUHK dysarthric speech recognition system for the ua speech corpus," in *INTERSPEECH*, 2018.

[8] Z. Ye, S. Hu, J. Li *et al.*, "Development of the CUHK elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *ICASSP*, 2021.

[9] D. Wang, J. Yu, X. Wu *et al.*, "Improved End-to-End dysarthric speech recognition via meta-learning based model re-initialization," in *ISCSLP*, 2021.

[10] A. Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

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

[12] A. König, N. Linz, J. Tröger *et al.*, "Fully automatic speech-based analysis of the semantic verbal fluency task," *Dementia and Geriatric Cognitive Disorders*, vol. 45, no. 3-4, pp. 198–209, 2018.

[13] W. Lanier, "Speech disorders." in *Greenhaven Publishing LLC*, 2010.

[14] T. Hixon and J. C. Hardy, "Restricted motility of the speech articulators in cerebral palsy." *J SPEECH HEAR DISORD*, vol. 29, pp. 293–306, 1964.

[15] R. D. Kent, J. F. Kent, G. Weismer *et al.*, "What dysarthrias can tell us about the neural control of speech," *J. Phonetics*, vol. 28, pp. 273–302, 2000.

[16] H. Albaqshi and A. Sagheer, "Dysarthric speech recognition using convolutional recurrent neural networks," *INT J INTELL SYST*, vol. 13, pp. 384–392, 2020.

[17] E. Hermann and M. M. Doss, "Dysarthric speech recognition with lattice-free MMI," in *ICASSP*, 2020.

[18] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*, 2019.

[19] M. Geng, X. Xie, S. Liu *et al.*, "Investigation of data augmentation techniques for disordered speech recognition," in *INTERSPEECH*, 2020.

[20] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *INTERSPEECH*, 2018.

[21] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *SLPAT*, 2015.

[22] J. Shor, D. Emanuel, O. Lang *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," in *INTERSPEECH*, 2019.

[23] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *INTERSPEECH*, 2020.

[24] D. Woszczyk, S. Petridis, and D. Millard, "Domain adversarial neural networks for dysarthric speech recognition," in *INTERSPEECH*, 2020.

[25] F. Xiong, J. Barker, Z. Yue *et al.*, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*, 2020.

[26] Y. Takashima, R. Takashima, T. Takiguchi *et al.*, "Dysarthric speech recognition based on deep metric learning," in *INTERSPEECH*, 2020.

[27] Y. Lin, L. Wang, S. Li *et al.*, "Staged knowledge distillation for End-to-End dysarthric speech recognition and speech attribute transcription," in *INTERSPEECH*, 2020.

[28] D. Wang, J. Yu, X. Wu *et al.*, "End-to-End voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *ICASSP*, 2020.

[29] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang *et al.*, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," in *INTERSPEECH*, 2020.

[30] J. T. Becker, F. Boller, O. L. Lopez *et al.*, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Arch. Neurol.*, vol. 51, no. 6, pp. 585–594, 1994.

[31] K. H. Wong, Y. T. Yeung, E. H. Chan *et al.*, "Development of a Cantonese dysarthric speech corpus," in *ISCA*, 2015.

[32] W. Xiong, J. Droppo, X. Huang *et al.*, "Toward human parity in conversational speech recognition," *IEEE TASLP*, vol. 25, pp. 2410–2423, 2017.

[33] T. Hain, L. Burget, J. Dines *et al.*, "Transcribing meetings with the amida systems," *IEEE TASLP*, vol. 20, pp. 486–498, 2012.

[34] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[35] V. Panayotov, G. Chen, D. Povey *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[36] S. Hu, X. Xie, S. Liu *et al.*, "Neural architecture search for LF-MMI trained time delay neural networks," in *ICASSP*, 2021.

[37] T. Moriya, T. Tanaka, T. Shinozaki *et al.*, "Evolution-strategy-based automation of system development for high-performance speech recognition," *IEEE TASLP*, vol. 27, no. 1, pp. 77–88, 2018.

[38] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *ICLR*, 2018.

[39] A. Waibel, "Consonant recognition by modular construction of large phonemic time-delay neural networks," in *ICASSP*, 1989.

[40] D. Povey, G. Cheng, Y. Wang *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018.

[41] S. Hu, X. Xie, S. Liu *et al.*, "Bayesian learning of LF-MMI trained time delay neural networks for speech recognition." *IEEE TASLP*, vol. 29, pp. 1514–1529, 2021.

[42] H. Hadian, H. Sameti, D. Povey *et al.*, "End-to-end speech recognition using lattice-free mmi," in *INTERSPEECH*, 2018.

[43] W. Michel, R. Schlüter, and H. Ney, "Comparison of lattice-free and lattice-based sequence discriminative training criteria for LVCSR," in *INTERSPEECH*, 2019.

[44] S. Hu, X. Xie, S. Liu *et al.*, "LF-MMI training of Bayesian and Gaussian process time delay neural networks for speech recognition." in *INTERSPEECH*, 2019.

[45] S. Xie, H. Zheng, C. Liu *et al.*, "SNAS: Stochastic neural architecture search," in *ICLR*, 2019.

[46] C. J. Maddison, A. Mnih, and Y. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR*, 2017.

[47] H. Pham, M. Y. Guan, B. Zoph *et al.*, "Efficient neural architecture search via parameter sharing," in *ICML*, 2018.

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

Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models

Lara Gauder[1,2]*, Leonardo Pepino*[1,2]*, Luciana Ferrer*[1]*, Pablo Riera*[1]

[1]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[2]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Argentina
{mgauder,lpepino,lferrer,priera}@dc.uba.ar

Abstract

This paper describes our submission to the ADreSSo Challenge, which focuses on the problem of automatic recognition of Alzheimer's Disease (AD) from speech. The audio samples contain speech from the subjects describing a picture with the guidance of an experimenter. Our approach to the problem is based on the use of embeddings extracted from different pre-trained models — trill, allosaurus, and wav2vec 2.0 — which were trained to solve different speech tasks. These features are modeled with a neural network that takes short segments of speech as input, generating an AD score per segment. The final score for an audio file is given by the average over all segments in the file. We include ablation results to show the performance of different feature types individually and in combination, a study of the effect of the segment size, and an analysis of statistical significance. Our results on the test data for the challenge reach an accuracy of 78.9%, outperforming both the acoustic and linguistic baselines provided by the organizers.
Index Terms: computational paralinguistics, ADreSSo challenge, Alzheimer's Disease recognition

1. Introduction

For many health problems, like speech pathologies, Parkinson's disease, Alzheimer's Disease (AD), and respiratory problems, the patient's speech is routinely used by doctors as one of the tools for diagnosis and monitoring of disease progression [1]. In particular, AD is characterized by a progressive decline of cognitive and functional abilities over time [2] often including language impairment, even at early stages [3]. As a consequence, many studies rely on the analysis of the speech signal as a source of clinical information for AD [4, 5].

In this work, we present results and analysis of our submission to the ADreSSo (Alzheimer's Dementia Recognition through Spontaneous Speech only) Challenge [6]. This challenge is focused on the automatic detection of AD using recordings of interviews with the subjects. A previous version of this challenge, called ADreSS, took place last year. In that case, manual transcriptions of the speech signals were provided to the participants along with the recordings. In this year's challenge, manual transcriptions are not provided, so systems have to rely solely on the speech signal for classification. The challenge includes three tasks: AD classification, Mini-Mental State Examination (MMSE) score regression, and cognitive decline inference. In this work, we present results on the AD classification task.

AD may affect the patient's speech production in terms of paralinguistic aspects like the prosodic patterns, pause patterns or quality of speech, and in terms of linguistic aspects, like choice of words or grammatical forms. Previous works have found that both acoustics and linguistic information can be used for automatic prediction of AD. In a paper about the ADreSS challenge [7], a comparison of acoustic and linguistic features showed that acoustic features resulted in an accuracy of 64.5% while linguistic features from manual transcriptions resulted in an accuracy of 85.42%. A similar trend is observed in the baseline results for this year's challenge [6], although with relatively poorer performance for the linguistic features due to the absence of manual transcriptions, which are replaced by automatic ones. In our work for this challenge, we focus on the use of acoustic features, without extracting automatic word transcriptions. Further, considering the sparsity of the available training data, we propose to use transfer learning approaches. To this end, we leverage recently released speech-based embedding models that aim to represent different aspects of the speech signal.

Pre-trained speech-based embeddings are currently being used in several speech recognition tasks, such as speech emotion recognition [8, 9, 10, 11] and automatic speech translation [12]. These compact representations can encode different speech attributes depending on the way the models are trained. Information about prosody, phonetic or lexical content may be emphasized in the representations, depending on the task used to train the models. The use of these representations, in combination with neural networks for the modeling stage, often provides an improvement over directly using signal processing features like mel frequency cepstral coefficients (MFCC) or Mel-spectrograms.

In this paper, we present our results using different types of embeddings and traditional prosodic features for the task of AD classification. We use a simple deep neural network for modeling each individual feature and their combinations. The model takes relatively short segments of speech as input and averages the resulting scores over all segments in an audio sample to create the final score. We show different analysis, including an ablation study to find the most useful features, a study of statistical significance, a comparison of the effect of the window length, and an analysis of the effect of the presence of experimenter speech in the signals. Our results on the challenge data are significantly better than the acoustic-only baseline results implemented by the organizers and also outperform the linguistic baseline that uses automatic transcriptions [6].

2. Dataset

The development dataset provided by the ADreSSo challenge consists of 166 recordings of 87 patients with AD diagnosis and 79 cognitively normal subjects. All the subjects were asked to

http://dx.doi.org/10.21437/Interspeech.2021-753

describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. Audio files contain both the speech from the subjects and the experimenter conducting the interview. A test set with audio files from 71 subjects was used for blind evaluation of the models. Challenge participants are not provided AD labels for this test data. The complete dataset description is available in [6]. The challenge includes three tasks: an AD classification, an MMSE score regression and a cognitive decline (disease progression) inference task. We participated on the classification task, where the goal is to determine whether a subject is a control (CN) subject or a patient with AD, based only on the speech signal from the interview.

Since recordings include both the speech from the subject and the experimenter, the dataset includes segmentation information indicating where each of the two speakers speaks. In our initial inspection of the development data, though, we found that this information was inaccurate for several of the audio files. Further, we found a case where the recording included speech from more than two speakers and was also not accurately segmented to identify the subject's speech. For these reasons, we decided to work with the full audio files, without using the provided manual segmentation, assuming that the speech from the experimenters and any other speakers represent only a relatively small portion of the speech present in the signal. In Section 5.4, we show results that indicate that including the segments from the experimenters did not degrade the performance of our system.

3. Acoustic Features

Our approach for AD classification is to use embeddings, i.e., vector representations, extracted from a set of pre-trained models. These models are deep neural networks (DNN) trained on large speech dataset to solve different tasks. The embeddings are then extracted from the output of some layer of the DNN. In general, embeddings are extracted over relatively short regions of the signal and may contain only local information or include contextual information about the rest of the signal. The details on the embeddings used for this paper are described below. Further, we also include traditional features, designed for tasks like emotion recognition. All features were normalized by subtracting the mean and dividing by the standard deviation of that feature over each recording. This normalization approach resulted in better performance than global normalization where every feature is normalized with the mean and standard deviation obtained over all the training data.

3.1. eGeMAPS features

The extended Geneva minimalistic acoustic parameter set (eGeMAPS) [13] is a set of features designed specifically for affective speech tasks and includes pitch, loudness, formants and voice quality features among others. The set includes both low-level features, extracted every 10 ms of speech over windows of 25 ms, and high-level features that correspond to different statistics extracted from the low-level features. We use only the low-level descriptors of the eGeMAPS v2.0 set which contains 25 features for every time step.

3.2. Trill features

The trill model was trained to generate a non-semantic representation of speech [14]. The model minimizes a triplet loss designed to solve the task of classifying whether a segment of audio comes from the same or from a different original audio file as another segment. The resulting embeddings were evaluated in many different non-semantic tasks including speaker identification, emotion recognition and others. Authors also tested the model on AD recognition using the Dementia Bank dataset [15], showing good results when fine-tuning the model to this task. We used the distilled version of the trill model, which generates embeddings of size 2048.[1] For this work, we resampled the trill embeddings, which are produced every 167 ms to 100 vectors per second (one every 10 ms) to match the resolution of the other features.

3.3. Allosaurus features

Allosaurus[2] is a universal phone recognition model that includes pre-trained acoustic and language models [16]. It can be used to generate phonetic transcriptions and phone logits given by $\log(p/(1-p))$, where p is the phone posterior probability, producing one set of logits every 10 ms. The model was trained with 11 languages and over 2000 utterances. For English, the output logits are 39-dimensional. The information contained in these logits could help us model specific pronunciation issues in the AD subjects, as well as some indirect information about word usage which could be found in the frequency of certain phones.

3.4. Wav2vec 2.0 features

Wav2Vec 2.0 (from now on wav2vec2) is a framework for self-supervised learning of representations from raw audio [17]. The model can generate contextualized embeddings that can be later integrated in end-to-end speech-to-text models to generate transcriptions. Thus, these embeddings preserve the phone content of the signal among other information. We used the model available from the Transformer Python library which gives embeddings of size 768 and was fine-tuned with 960 hours of LibriSpeech.[3] Large audio files had to be segmented in 20 seconds long fragments to compute the features since they could not otherwise be processed by the model. This may be suboptimal since it prevents the model from extracting contextual information from the full signal. The wav2vec2 embeddings are produced every 20 ms. As for the trill features, we resampled these vectors to obtain 100 per second, matching the resolution of the other features.

4. Classification model

We use deep neural networks as models for the different individual features and their combinations. The input to the networks are 5-second segments extracted from the original audio with 1-second overlap between segments. This results in 3178 segments extracted from the development data. The final score for each audio file is obtained by computing the mean over the scores for all the segments in the file. As explained above, we use the full audio file, which means that some of the segments contain speech from the experimenter. Section 5.4 shows results on the effect of the experimenter's speech on the performance of the system.

The architecture is depicted in Figure 1 for the configuration where all features are used. Each feature set has a corresponding branch that performs a first reduction of the embedding with a 1D convolution with kernel size 1 (equivalently, a time-distributed dense layer) followed by a 1D convolution with

[1] https://tfhub.dev/google/nonsemantic-speech-benchmark/trill-distilled/3

[2] https://github.com/xinjli/allosaurus

[3] https://huggingface.co/facebook/wav2vec2-base-960h

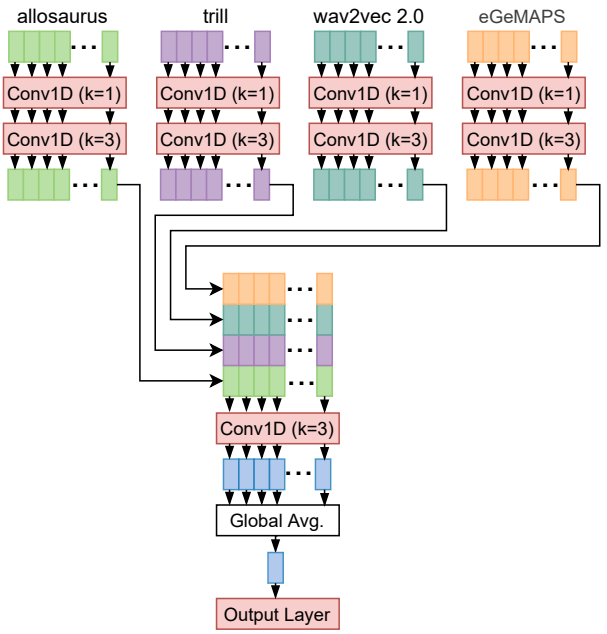


Figure 1: *Neural network architecture for the configuration in which all the features are used. The value k indicates the size of the kernel in the time dimension.*

kernel size 3. After the second convolution, the dimension of the output for each branch is 128. Then the activations from the four branches are concatenated and a second 1D convolution is performed, reducing the dimension back to 128. Finally, the output of this layer is averaged across time and a dense layer computes the prediction scores. Batch normalization and ReLu activations are used in every layer. When considering only a subset of the features, the same model is used with only the corresponding branches included.

5. Results

In this section we show results for the proposed systems, including an ablation study, statistical significance results, an analysis of the influence of the segment size and of the effect of the experimenter speech on the performance of the system. We run experiments using 6-fold cross-validation (CV) on the development data provided to challenge participants. The folds are determined by subject to prevent segments from the same subject being in the training and the test set for a certain fold, which would make the CV results overly optimistic. The models used to obtain scores for the challenge's evaluation data were trained on the full training set.

5.1. Ablation Results

The middle column in Table 1 shows the results on the development set obtained with cross-validation, for several systems including single feature sets, 2-way combinations, and the 4-way combination. The best individual features are trill and wav2vec2. We hypothesize that this may be partly because these two features sets have large dimensionality, 2048 for trill and 768 for wav2vec2 (compared to the other two which have 25 features for eGeMAPS and 39 for allosaurus), allowing these features to contain a richer representation of the audio. Larger dimensions could also result in the model overfitting the training data, which would lead to poor results, but this effect is discouraged by the small architectures we use.

Feature set	Dev	Test
eGeMAPS	63.9%	-
trill	72.9%	69%
allosaurus	66.3%	-
wav2vec2	**75.3%**	**78.9%**
eGeMAPS, allosaurus	63.9%	-
eGeMAPS, wav2vec2	72.3%	-
eGeMAPS, trill	71.1%	-
trill, allosaurus	72.9%	70.4%
trill, wav2vec2	75.2%	69%
allosaurus, wav2vec2	70.5%	-
all	74.7%	70.4%

Table 1: *Accuracy values for different combinations of features for development and test data. The five best performing models on the development set were submitted to the challenge. Results on the test set are shown for those cases.*

Fusion results show no gains with respect to the best individual system, wav2vec2. This could imply that the other three sets of features are redundant given the wav2vec2 features. That is, that wav2vec2 features contain all the information in eGeMAPs, trill and allosaurus features that is important for AD classification. In fact, we would expect allosaurus and wav2vec2 features to be somewhat redundant since they are trained to solve similar tasks: phone recognition and speech recognition, respectively. On the other hand, we would also expect trill or eGeMAPs to provide some complementary information to those two set, since they are designed to contain information beyond the phonetic content. Hence, a more likely explanation for the lack of gain from fusion is that our downstream model is not able to effectively combine the information from all these sets. In the future, we will continue exploring different architectures for the combination of these features.

Finally, the right column in Table 1 shows the results for the 5 best systems based on the development results, which were the ones selected for submission to the challenge. In the test results, as in the development results, wav2vec2 alone was the best performing model, with an accuracy of 78.9%. This result is superior to the acoustic baseline results presented in [6], which have an accuracy of 64.79%. Further, they are also superior to the linguistic baseline results in that paper, which has an accuracy of 77.46%. This is not too surprising since wav2vec2 features are designed to contain phonetic information and, hence, are probably able to implicitly represent some information about word-usage, as well as pronunciation patterns. Further, wav2vec2 embeddings have a very distinct pattern over non-speech regions. Hence, our downstream model could potentially be learning patterns of usage of pauses, which are likely to be useful for differentiating AD from control subjects.

5.2. Statistical Significance Study

Given the relatively small number of samples available both in the development and the evaluation sets, we conducted a bootstrapping analysis on the development scores to determine confidence intervals for each of the systems submitted to the challenge. We sampled with replacement the 166 development scores obtained with CV to get 5000 new sets of scores, each with 166 samples. For each of these bootstrap sets, we computed the accuracy. The purple bars in Figure 2 show the 5% and 95% percentiles of the resulting set of accuracy values. We can see that the intervals are wide: all systems overlap with the others making it impossible to conclude whether there is, in-

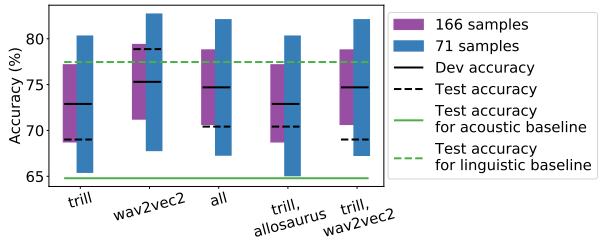


Figure 2: *Confidence intervals from bootstrapping experiment for the 5 best-performing models on the development set. The purple bars show the confidence intervals with bootstrap sets of size 166, the same as the original set, while the blue bars show the intervals with sets of size 71, the size of the test set. Black lines show the accuracy for each model on the development and test sets. Further, the green lines correspond to the test accuracy for the two baseline systems in [6].*

deed, a significant difference between them.

Further, since the test set is smaller than the development set, containing 71 subjects instead of 166, we repeated the bootstrap analysis on the development scores, but this time selecting only 71 samples per bootstrap set. The resulting confidence intervals are, of course, wider, and reflect the variability we could expect when testing these systems on a dataset of that size. Notably, the actual test results (shown in dashed black lines in Figure 2) fall within the estimated blue intervals suggesting that the test data is well represented by the training data.

Finally, the green lines in Figure 2 show the baseline results provided by the organizers in [6]. We can see that results for all our systems are significantly better than the acoustic baseline results. The wav2vec2 results are also better than the linguistic baseline results, though not by a significant margin.

5.3. Effect of the segment size

Our downstream model takes relatively short segments as input, extracted from the original audio with some overlap. The final score for each audio file is then given by an average of the segment-level scores. In this section, we study the effect of the segment size. Figure 3 shows the accuracy results at audio level (i.e., one sample per subject, as in all other results in this paper) and at segment level, using varying segment sizes. For this figure, segments are shifted by 2 seconds instead of 4, as in previous results, so that no speech is lost when using 2-second segments. Note that, since the shift is fixed, the number of segments for a certain audio file is approximately the same for all segment sizes. Further, to improve the stability of the results, we run each model with 3 different seeds to determine the cross-validation folds. The results shown in the plot correspond to the average accuracy over those 3 runs.

Figure 3 shows an interesting trend. As intuition would suggest, segment-level results improve as the segment size increases, since more information is available to make the classification decision. On the other hand, when averaging the scores from all segments in an audio file, the optimal segment size is around 5 seconds; longer segments degrade performance. We hypothesize that this is because, in longer segments, the effect of some short-term phenomena that might be a strong indicator for classification may be *washed out*. On the other hand, when using shorter segments, the model may be able to focus on these local phenomena and produce more discriminative scores for the segments that contain them. Further analysis is necessary to prove or disprove this hypothesis. If proven true, this may

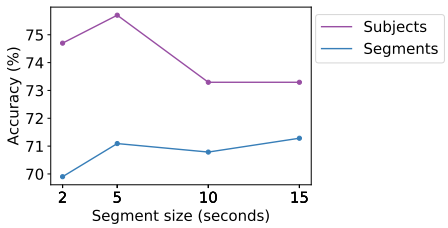


Figure 3: *Average accuracy over three seeds for the wav2vec2-only model, varying the segment size, for segment- and subject-level scores.*

suggest an interesting research direction: the development of hierarchical models that would take the output of our segment-level scores and effectively combine their outputs to emphasize the more informative scores, for example, using attention mechanisms.

5.4. Analysis of the influence of the experimenter speech

As mentioned above, our systems obtain the final scores for each subject as an average over all segments in an audio file. A portion of these segments contain at least some speech from the experimenter. To explore whether these segments had a negative impact on our results, we performed the following experiment. We computed the average accuracy over three seeds using only the audio files for which the manual segmentation had no obvious issues (139 out of the 166 files). Using the wav2vec2-only model from the previous section, with segment size of 5-seconds and shifts of 2-seconds, this gave an accuracy of 76.97%. We then discarded the scores from all the segments with any speech from the experimenter (33% of them) and recomputed the average score for each audio file. The accuracy for these new average scores did not significantly change. Given this result, we can conclude that the effect of the experimenter's speech is not harmful once the model is fixed. On the other hand, it is possible that a model trained without segments including speech from the experimenter would work better. This analysis is left for future work.

6. Conclusions

We presented our work on Alzheimer's Disease recognition, using the data from this year's ADreSSo challenge. Our approach uses speech-based embeddings from three different pre-trained models recently released to the public: trill, allosaurus and Wav2vec 2.0. We also include eGeMAPS, a set of features traditionally used for emotion recognition and related tasks. The features are modeled with a simple neural network that takes short segments of audio and generates scores which are then averaged to obtain the final score for each audio file. Word transcriptions are not used by our system. We show that the best results are obtained using Wav2vec 2.0 features, though all features perform similarly, considering the wide confidence intervals. Our results significantly outperform the acoustic baseline provided by the organizers, reaching an accuracy of 78.87% on the challenge's test set.

7. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation for the donation of a Titan Xp GPU.

8. References

[1] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

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

[3] G. W. Ross, J. L. 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] K. Lopez-de Ipina, U. Martinez-de Lizarduy, P. M. Calvo, J. Mekyska, B. Beitia, N. Barroso, A. Estanga, M. Tainta, and M. Ecay-Torres, "Advances on automatic speech analysis for early detection of alzheimer disease: a non-linear multi-task approach," *Current Alzheimer Research*, vol. 15, no. 2, pp. 139–148, 2018.

[5] 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, and U. M. d. Lizardui, "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. [Online]. Available: https://www.mdpi.com/1424-8220/13/5/6730

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," in *Submitted to INTERSPEECH 2021*, 2021. [Online]. Available: https://edin.ac/31eWsjp

[7] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[8] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 373–380.

[9] Y. Zhao, D. Yin, C. Luo, Z. Zhao, C. Tang, W. Zeng, and Z.-J. Zha, "General-Purpose Speech Representation Learning through a Self-Supervised Multi-Granularity Framework," *arXiv:2102.01930*, 2021.

[10] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive Unsupervised Learning for Speech Emotion Recognition," *arXiv:2102.06357*, 2021.

[11] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based unsupervised pre-training for acoustic representation learning," *arXiv:2007.14602*, 2021.

[12] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating Self-Supervised Pre-Training for End-to-End Speech Translation," in *Proc. Interspeech 2020*, 2020, pp. 1466–1470. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1835

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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, 2016.

[14] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," *arXiv preprint arXiv:2002.12764*, 2020.

[15] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

[16] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

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

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

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

[1]Usher Institute, Edinburgh Medical School, The University of Edinburgh, UK
[2]Department of Psychology, Carnegie Mellon University, USA

{S.Luz, fasih.haider, sofia.delafuente}@ed.ac.uk, {fromm, macw}@andrew.cmu.edu

Abstract

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

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

1. Introduction

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

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

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

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

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

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

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

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

2. Related work

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

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

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

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

3. The ADReSS$_o$ Datasets

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

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

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

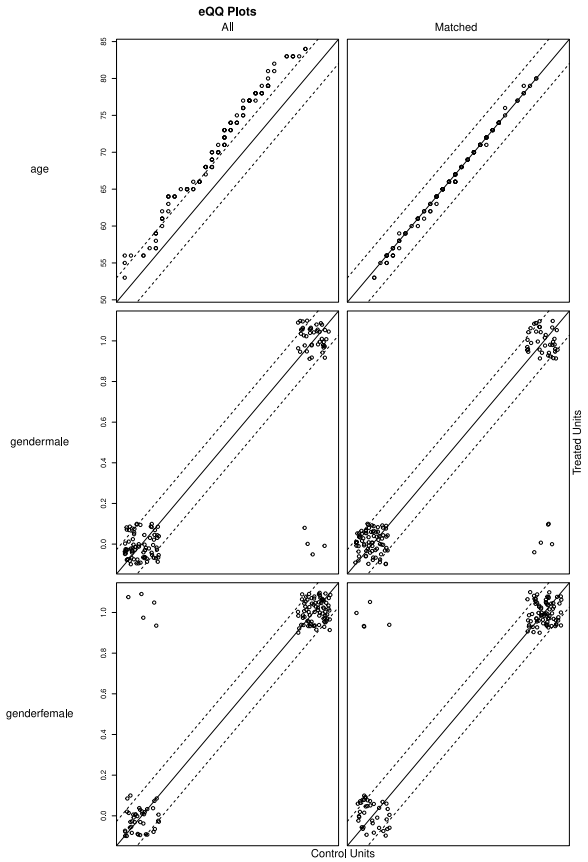


Figure 1: *Quantile-quantile plots for data before (left) and after matching (right) by age and gender.*

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

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

Table 1: *Composition of the datasets.*

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

4. Data representation

4.1. Acoustic features

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

4.2. Linguistic Features

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

5. Diagnosis baseline

5.1. Task 1: AD Classification

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

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

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

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

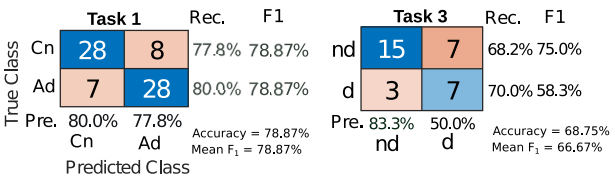


Figure 2: *Late (decision) fusion of the best results of acoustic and linguistic models for Task 1 (left) and Task 3 (right). Precision (Pre) , recall (Rec), accuracy (Accu) and mean F_1 scores are shown on the margins.*

5.2. Task 2: MMSE prediction

For this task we used five types of regression models: linear regression (LR), DT, with leaf size 20 and CART algorithm, sup-

port vector regression (SVR, with a radial basis function kernel, box constraint of 0.1 and sequential minimal optimisation solver), Random Forest regression ensembles (RF), and Gaussian process regression (GP, with a squared exponential kernel). The regression methods are implemented in MATLAB [34].

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

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

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

6. Prognosis baseline (Task 3)

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

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

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

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

7. Discussion

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

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

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

8. Conclusions

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

9. Acknowledgements

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

10. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Identifying cognitive impairment using sentence representation vectors

Bahman Mirheidari[1]*, Yilin Pan*[1]*, Daniel Blackburn*[2]*, Ronan O'Malley*[2]*, and Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
{b.mirheidari,heidi.christensen}@sheffield.ac.uk

Abstract

The widely used word vectors can be extended at the sentence level to perform a wide range of natural language processing (NLP) tasks. Recently the Bidirectional Encoder Representations from Transformers (BERT) language representation achieved state-of-the-art performance for these applications. The model is trained with punctuated and well-formed (writ-ten) text, however, the performance of the model drops significantly when the input text is the – erroneous and unpunctuated– output of automatic speech recognition (ASR). We use a sliding window and averaging approach for pre-processing text for BERT to extract features for classifying three diagnostic categories relating to cognitive impairment: neurodegenerative dis-order (ND), mild cognitive impairment (MCI), and healthy controls (HC). The in-house dataset contains the audio recordings of an intelligent virtual agent (IVA) who asks the participants several conversational questions prompts in addition to giving a picture description prompt. For the three-way classification, we achieve a 73.88% F-score (accuracy: 76.53%) using the pre-trained, uncased base BERT and for the two-way classifier (HCvs. ND) we achieve 89.80% (accuracy: 90%). We further improve these by using a prompt selection technique, reaching the F-scores of 79.98% (accuracy: 81.63%) and 93.56% (accuracy:93.75%) respectively.

Index Terms: Dementia detection, language representation, speech recognition, processing of pathological speech

1. Introduction

Dementia affects the cognitive and communication abilities of people which impairs their speech and language as well as their general ability to undertake daily activities. The number of people living with dementia in the UK is over 850,000 and it is expected to rise to one million by 2025 [1]. Dementia and Alzheimer's is one of the top leading causes of death in the UK [2]. Detecting the early stages of dementia is a challenging task due to overlapping symptoms with normal ageing, limited capability of existing screening tools, and the high risk associated with some procedures such as exposure to radiation. Current cognitive tests routinely assess the speech and communication ability of people but these are inaccurate and difficult to interpret for non-specialists. Therefore, developing automatic assessments methods is of great importance. Recently, we have developed an automatic system analysing people's conversation with an intelligent virtual agent (IVA) to detect cognitive decline [3, 4]. The IVA prompts the users to answer a number of questions as well as to perform some standard cognitive tests including the Cookie Theft description task [5]. In our previous work, we have extracted a number of acoustic, lexical, conceptual and conversational analysis features from the whole conversations between the participants and the IVA, and we showed how the features can be used to reliably distinguish betwee neurodegenerative disorder (ND), mild cognitive impairment (MCI) and healthy controls (HCs) [6, 7, 4]. In this paper, we explore the role of each prompt in the classification accuracy. We use the word vector and language representation (sentence level vectors) to extract the context of the answers given to each question.

Word embedding techniques are widely used in natural language processing (NLP) applications. The early techniques were bag-of-words (BOW) [8] and Frequency Inverse Document Frequency (TF-IDF) [9] which did not consider the order of words nor the context. The next techniques were trained using neural networks such as W2Vec [10] and GloVE [11] which captured the co-occurrences of words and the context of the text. However recently, state-of-the-art performances have been achieved by using transformer-based models like the BERT language model [12]. BERT has been widely used in a variety of natural language processing (NLP) applications like question answering, topic detection, summarisation and semantic search.

Common for those types of applications is that the input text is well-formed and well-punctuated, i.e., matching what the embedding model has been trained on. However, if using BERT in a *speech-driven* pipeline, the input text will no longer be a good match but instead be a continuous string of predicted words. In this paper, we investigate how the automatic speech recognition (ASR) errors and lack of punctuation in the text may affect the performance of BERT model for a system aiming to detect cognitive impairment in spontaneous speech. Using the word vectors and language representation (sentence level vectors) as features, we can train efficient classifiers with high performance to distinguish between different levels of cognitive impairment related to dementia. Using the features we determine a subset of questions that are more useful for the classification (question selection process).

The remainder of the paper contains the following: Section 2 is a brief introduction to related work of using word vector and sentence/text vector techniques. Section 3 contains our experimental setup, especially how we train the ASR, extract the sentence vectors and train the classifiers. Section 4 and 5 cover the results and conclusions.

2. Related work

BERT models have been recently introduced for dementia detection and state-of-the-art results have been reported [13]. There has been a number of studies using BERT in the ADReSS challenge 2020 [14] containing Cookie Theft descriptions, in which the authors shared a training set containing 78 audio files with the corresponding human transcripts of Alzheimer's Disease (AD) participants as well as 78 transcript for non-AD people, and a test set with 48 recordings (24 for each group). The data was taken from the Dementia Bank Corpus [15] to perform two tasks: classification between AD and non-AD, and

http://dx.doi.org/10.21437/Interspeech.2021-915

predicting the mini-mental state examination (MMSE) scores. The successful models in the challenge were mostly trained on the manual transcripts of the recordings. In reality it is not desirable to rely on the manual transcripts as it would be costly and time-consuming to provide them. Instead fully automated systems should be explored, which can take as input the audio recordings. This means they need to be able to deal with the challenges of using ASR systems and combining the automatic transcripts with models such as BERT.

Different groups in the challenge reported different performances. In particular, the models using BERT achieved good accuracies ranging between 75% and 85.4%, e.g. [16] (combining x-vectors and the manual transcripts) reported 75% accuracy; [17] got 81.25% and [18] 83.3%; [19] (adding pauses and disfluency to the transcripts) achieved 85.4% accuracy, similar to the accuracy gained by [20] with a multi-modal (using both audio signals and manual transcripts) BERT.

The ASR errors can affect the word vector techniques significantly. In our previous work, we used GloVe word vectors on 473 recordings of the Dementia Bank dataset and we achieved an accuracy of 75.6% using the manual transcripts without the punctuation. However, the accuracy dropped significantly to 62.3% when we replaced the manual transcripts with the ASR outputs (45.3% WER) [21]. Therefore, we can expect that when using BERT adding the erroneous text from the ASR affects the results considerably. This paper analyses the effect of using ASR transcripts with BERT in a cognitive impairment application and proposes a way of mitigating the effects.

3. Experimental setup

The IVA dataset was collected between 2016 and 2020 at the Department of Neurology, University of Sheffield, UK (Royal Hallamshire Hospital). A total of 168 participants were recorded of which 98 were chosen for this study (61 HC, 19 ND and 18 MCI [1]). The other recordings were only used for training the ASR. The IVA conversation includes nine questions, and the Cookie Theft description task.

3.1. ASR

The LIBRISPEECH dataset was used for training a time delay neural networks (TDNN) acoustic model based on Kaldi's LIBRISPEECH recipe [22]. Then using a 10 fold cross-validation approach, the base acoustic model was adapted to the IVA dataset following the transfer learning technique of [23] (transferring all layers). One epoch of training was carried out on the target dataset (IVA) to adapt the acoustic model. For the language model, the four-gram model was used with Turing smoothing interpolated with the language model of the LIBRISPEECH text. An average 27.9% WER was achieved using the 10-fold cross validation approach.

3.2. Classifiers

Two types of classifiers were chosen for the experiments in this study: the conventional Logistic Regression (LR) classifier and a Transformer based sequence classifiers. The LR classifier is efficient and quick and produces deterministic results, while the Transformer classifiers need longer time to be trained and tuned and each time produce different results depending on the initialisation of the layers' weights. To gain stabilised results, the Transformer classifiers were run five times and then, using a voting approach, the labels were predicted. LR needs a fixed-length feature vector as its input, whereas the Transformers are fed with a fixed-length sequence of vectors.

[1]These are the numbers that we have collected so far. The numbers of each classes are not balanced, so we did calculate weighted F-score, precision and recall.

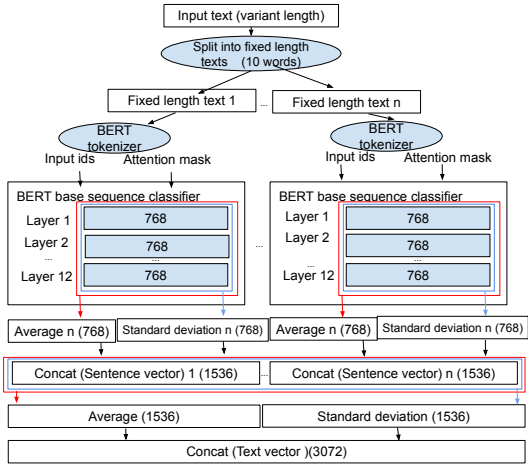


Figure 1: *Averaging technique to build the sentence vectors on the the BERT model: each input text is divided into a number of fixed length text (using a sliding window). Then the average of the layers and the standard deviations are combined to build a sentence and whole text vector.*

3.3. Sentence vectors

The Transformer based classifiers can handle input text of varying lengths. If the input text is shorter than a maximum length of words, the spaces are padded with a specific token, while the longer sentences are truncated. BERT models were trained on punctuated corpora. The authors of BERT trained both cased and uncased [2] models of BERT with different sizes of corpora (tiny, small, medium and large). In this study, we use the pre-trained model with a normal size corpus, as well as the large size: uncased BERT base (for convenience we refer to it as BERT), and uncased BERT large (BERT-large).

A sentence can be passed to the pre-trained BERT model as its input and the corresponding sentence vector can be extracted from the weights of the network's layers. However, as mentioned before, the LR classifier works on fixed-length input features. So for BERT with 12 layers and each layer having 768 neurons, per each input word we can extract a vector of $12 \times 768 = 9216$ dimensions. Assume that the maximum word length is 150, then the BERT model can represent the whole input with a vector of $150 \times 9216 = 1,382,400$ dimensions (3.7 million dimensions for BERT-large), which is not feasible to use with an LR classifier. We therefore need to construct a compact version of the sentence/text vectors. To this end, we firstly calculate the average of the vectors as the representation for the sentence. We have tried different approaches and found that averaging the weights on all layers of the model, combined with the standard deviation of the weights (so representation with $2 \times 768 = 1536$ dimensions instead of 1.4 million), was the best; see Figure 1. We also observed that better representation can be achieved by using smaller sentences. So instead of using the whole text we split it into smaller sentences, and us-

[2]dealing with uppercase and lowercase words, respectively.

Table 1: *Two-way classification results on the ADReSS dataset using manual transcript with punctuation (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	80.00	79.17	79.02	79.17
LR (BERT)	**81.75**	**81.25**	**81.18**	**81.25**
LR (BERT-large)	79.37	79.17	79.13	79.17
Transformer (GloVe)	75.17	75.00	74.96	75.00
BERT classifier	82.67	81.25	81.04	81.25
BERT-large classifier	**88.57**	**87.50**	**87.41**	**87.50**

Table 2: *Two-way classification results on the ADReSS dataset using manual transcript without punctuation (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	**84.16**	**81.20**	**80.84**	**81.25**
LR (BERT)	**84.16**	**81.20**	**80.84**	**81.25**
LR (BERT-large)	**84.16**	**81.20**	**80.84**	**81.25**
Transformer (GloVe)	75.00	75.00	75.00	75.00
BERT classifier	**81.75**	**81.25**	**81.18**	**81.25**
BERT-large classifier	78.31	77.08	76.83	77.08

ing the same averaging technique, we calculate the average of those small sentence representations to represent the whole text (i.e. $4 \times 768 = 3072$ dimensions for BERT, $4 \times 1024 = 4048$ demotions for BERT-large, and $4 \times 300 = 1200$ dimensions for GloVe).

If the input text contains punctuation, the whole text can be divided into smaller pieces using the positions of full stops, exclamation or question marks. However, if there is no punctuation (as is the case when working with the output of conventional ASRs) we can split the whole text into a fixed-length sequence of words using the sliding window technique. We found that a window size of 10 and two-word steps for training the LR classifiers resulted in the best best performing classifiers.

4. Results

Before investigating which questions of the IVA dataset are more informative for the classification, it is interesting to first quantify the effect of removing the punctuation and ASR errors. We will do this by investigating this in the ADReSS dataset as well as in our IVA Cookie Theft descriptions.

4.1. Effect of removing punctuation

To demonstrate the effect of the punctuation on the classifiers' performance, we first used the manual transcripts of the ADReSS challenge with and without punctuation by applying the sliding window technique. Table 1 and 2 show the performance of the classifiers in terms of precision, recall, F-score and accuracy measures. Using the sentence word representation based on the punctuation marks, the best performance, achieved by the LR classifier trained on vectors from BERT, was an F-score of 81.18% and an accuracy of 81.25%. However, the best classifier was BERT-large which achieved 87.41% F-score and 87.50% accuracy (comparable to the highest results reported by the ADReSS challenge [14]). Removing the punctuation from the manual text reduced the performance of the BERT-large based classifier (an almost 10% drop to 77.08%). However, the LR classifiers trained on the features, obtained using the sliding window and averaging, were more robust; all reached almost 81% F-score and accuracy. Since we observed more robustness from the LR classifiers than the Transform based ones, we only used the LR classifiers for the following experiments.

4.2. IVA three-way classification

Our IVA contains Cookie Theft picture descriptions that are directly comparable to those in the ADReSS dataset. The IVA has three diagnostic classes though, ND, MCI and HC, and so we first performed the three-way classification tasks. Section 4.3 contains the two-way classifications (HC vs. ND). The tree-way classification task is naturally more difficult than the two-way classification not just because of a lower chance-level, but the classes are also more confusable in terms of how the speech and language is affected with MCI having some resemblance to both HC and ND classes. To show the effect of using the ASR-generated transcripts, we will perform the classification on the manual transcripts then compare them with the classifications using the ASR outputs. Table 3 shows that using the manual transcripts, LR (BERT-large) achieved 72.19% F-score (74.49% accuracy) (LR (BERT) gained a slightly lower F-score). However, as can be seen from the table, using the erroneous ASR transcripts reduced the performance of the three LR classifiers. The best three-way LR classifier was LR (BERT-large) with 68.76% F-score and 71.43% accuracy.

Table 3: *Three-way classification results on IVA Cookie Theft dataset using manual (Man) transcript vs. ASR outputs (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Man/ASR	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	Man.	72.11	72.45	67.75	72.45
LR (BERT)	Man.	72.78	74.49	72.12	74.49
LR (BERT-large)	Man.	**72.27**	**74.49**	**72.19**	**74.49**
LR (GloVe)	ASR	65.60	66.33	60.99	66.33
LR (BERT)	ASR	66.57	70.41	66.62	70.41
LR (BERT-large)	ASR	**70.26**	**71.43**	**68.76**	**71.43**

4.3. IVA two-way classification

As the next step, the LR classification tasks were repeated on only the two classes (ND vs. HC, which is similar to the ADReSS challenge task). Table 4 shows the LR classifier performances using only the Cookie Theft part of the recordings versus using Cookie Theft and questions. As can be seen given having only Cookie Theft the binary LR (GloVe) classifier outperformed the two other classifiers with 83.60% F-score and 83.75% accuracy. However, adding the nine questions, LR (BERT) achieved 89.80% F-score (90% accuracy).

4.4. Prompt selection

In our previous work, we have always included all of the prompts (questions and various cognitive tasks) given by the IVA. Here we explore whether this is warranted or perhaps different prompts contribute different amounts and so, using only a subset might improve performance. Table 5 shows the results

Table 4: *Two-way classification results on IVA dataset Cookie Theft (CT) using ASR outputs vs. Cookie Theft and all questions (CT + Q's) (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Prompts	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	CT	**83.47**	**83.75**	**83.60**	**83.75**
LR (BERT)	CT	79.96	81.25	80.20	81.25
LR (BERT-large)	CT	79.33	80.00	79.61	80.00
LR (GloVe)	CT+Q's	87.04	87.50	86.96	87.50
LR (BERT)	CT+Q's	**89.76**	**90.00**	**89.80**	**90.00**
LR (BERT-large)	CT+Q's	89.79	90.00	89.57	90.00

of the three-way classifiers using all questions plus the Cookie Theft text produced by the ASR. As can be seen, adding the text from the questions to the picture descriptions, has improved the performance of the three classifiers. More specifically, LR (BERT) achieved 73.88% F-score and 76.53% accuracy.

Table 5: *Three-way classification results on the IVA dataset (all questions and Cookie Theft) using ASR outputs (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Prompts	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	CT+Q's	65.79	69.39	66.58	69.39
LR (BERT)	CT+Q's	**73.76**	**76.53**	**73.88**	**76.53**
LR (BERT-large)	CT+Q's	73.82	75.51	73.13	75.51

For the two types of classification (three-way and two-way) the five most important prompts were selected on the best-performing classifiers from Tables 5 and 4 (we calculated all possible combination of nine questions and Cookie Theft description and then selected the combination with the highest F-score). The results are in Table 6. The prompts selection improved the three-way F-score of the LR (BERT) classifier from 73.88% to 79.98% (accuracy rose from 76.53% to 81.63%). The five best prompts were questions 3 (asking who's most worried about the condition), 4, 5 & 8 (recent memory) in combination with the Cookie Theft prompt. For the two-way scenario, the LR (BERT) classifier achieved 93.56% F-score (93.75% accuracy), and the five best prompts were questions 3 & 6 (distant memory), 7 & 8 (recent memory), and 9 (who manages finances). The two common best prompts amongst the three-way and two-way classifiers were questions 3 and 8.

To show the effect of question selection on the individual classes, confusion matrices of the classifiers were analysed; Figures 2 and 3. For the three-way classification, the question selection slightly decreased the percentage/number of correctly recognised HC participants (from 96.72% (59) to 95.08% (58)),

Table 6: *Prompt (Prmpt) selection for the best three and two way classifiers (Cl.) (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Cl.	5 best Prmpt's	WPr %	WRc %	WFs %	Ac %
3-way	**3**, 4, 5, **8**, CT	81.14	81.63	79.98	**81.63**
2-way	**3**, 6, 7, **8**, 9	93.75	93.75	93.56	**93.75**

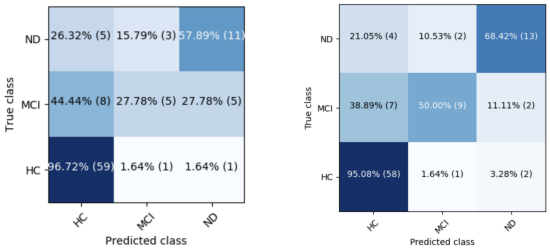


Figure 2: *Confusion matrix of the three-way LR (BERT) classifier using all questions vs. the best 5 questions.*

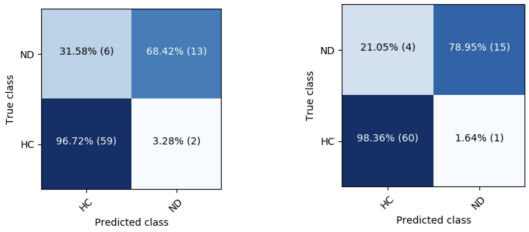


Figure 3: *Confusion matrix of the two-way LR (BERT-Large) classifier using all questions vs. the best 5 questions.*

however, the percentage/number of correctly recognised ND increased from 57.89% (11) to 68.42% (13) and for MCI more improvement can be seen from 27.78% (5) to 50% (9). However, the MCI group was the most confused group having more overlap with the other groups. For the two-way classification, the question selection improved the percent/number of correctly recognised ND participants from 68.42%(13) to 78.95%(15), and the number of correctly recognised HC from 96.72%(59) to 98.36% (60).

5. Conclusions

In this paper, we have shown that even though a BERT model can outperform other approaches such as GloVe embedding in many applications, their performance can be affected by the lack of punctuation and the errors introduced when using ASR in fully automated systems. We proposed using a sliding window and averaging technique to produce a sentence representation that can be successfully used as a feature to train a robust LR classifier to identify cognitive decline with high accuracy. Even more improvement was achieved by selecting a subset of the prompts which are more informative for the classification tasks. This demonstrates that the clinically diagnostic conversational questions asked by the IVA should be carefully chosen and that even quite a small set of prompts is capable of giving a very high performance, when used in conjunction with a model like BERT and careful pre-processing of ASR transcripts. For future work, we plan to further investigate technique to mitigate the effect of ASR transcripts on the quality of BERT output.

6. Acknowledgements

This work is supported by the European Union's H2020 Marie Skłodowska-Curie programme (TAPAS; Grant Agreement No. 766287), the Rosetrees Trust and the Stoneygate Trust (COMPASS, Grant Agreement No. M934 and the Fast ASsessment and Treatment in Healthcare funded by EPSRC (Reference. EP/N027000/1).

7. References

[1] D. UK, "What is dementia?" https://www.dementiauk.org, 2021, accessed on March 25, 2021.

[2] Dementia Statistics, "Deaths due to dementia," 2018, accessed on October 12, 2021. [Online]. Available: https://www.dementiastatistics.org/statistics/deaths-due-to-dementia

[3] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2732–2736.

[4] B. Mirheidari, D. Blackburn, R. O'Malley, A. Venneri, T. Walker, M. Reuber, and H. Christensen, "Improving cognitive impairment classification by generative neural network-based feature augmentation," *Proc. Interspeech 2020*, pp. 2527–2531, 2020.

[5] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[6] 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," *Proc. Interspeech*, pp. 3147–3151, 2017.

[7] B. Mirheidari, Y. Pan, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Data augmentation using generative networks to identify dementia," *arXiv preprint arXiv:2004.05989*, 2020.

[8] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[9] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

[12] 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.

[13] J. Glass *et al.*, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 3833, 2020.

[14] 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.

[15] J. T. Becker, F. B. F, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Arch Neurol*, vol. 51, pp. 585–594, 1994.

[16] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[17] E. L. Campbell, L. Docio-Fernandez, J. Jiménez-Raboso, and C. Gacia-Mateo, "Alzheimer's dementia detection from audio and language modalities in spontaneous speech."

[18] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[19] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[20] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[21] 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.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[23] V. Manohar, D. Povey, and S. Khudanpur, "Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.

ADReSSo Challenge

Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-based Alzheimer's Dementia Detection through Spontaneous Speech

Yilin Pan[1], *Bahman Mirheidari*[1], *Jennifer M Harris*[2,3], *Jennifer C Thompson*[2,4], *Matthew Jones*[2,4], *Julie S Snowden*[2,4], *Daniel Blackburn*[5], *Heidi Christensen*[1]

1. Department of Computer Science, University of Sheffield, UK
2. Division of Neuroscience and Experimental Psychology, University of Manchester, UK
3. Department of Psychology, University of Exeter, UK
4. Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, Salford UK
5. Department of Neuroscience, University of Sheffield, UK

The first two authors have equal contribution

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287., the Rosetrees Trust and the Stoneygate Trus (COMPASS, Grant Agreement No.M934 and the Fast ASsessment and Treatment in Healthcare funded by Engineering and Physical Science Research Council (EPSRC) (Reference.EP/N027000/1)

Proposed five models

Input: audios from training set
target: predict the label of the audios in the test set

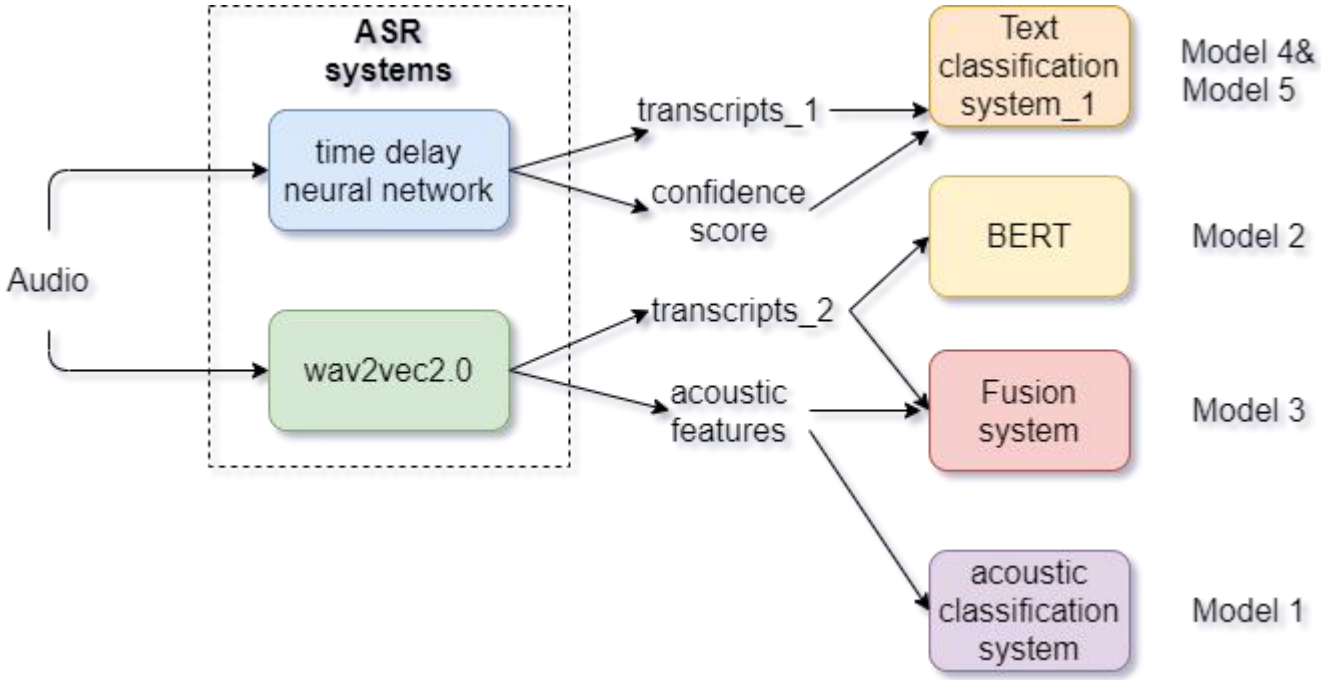
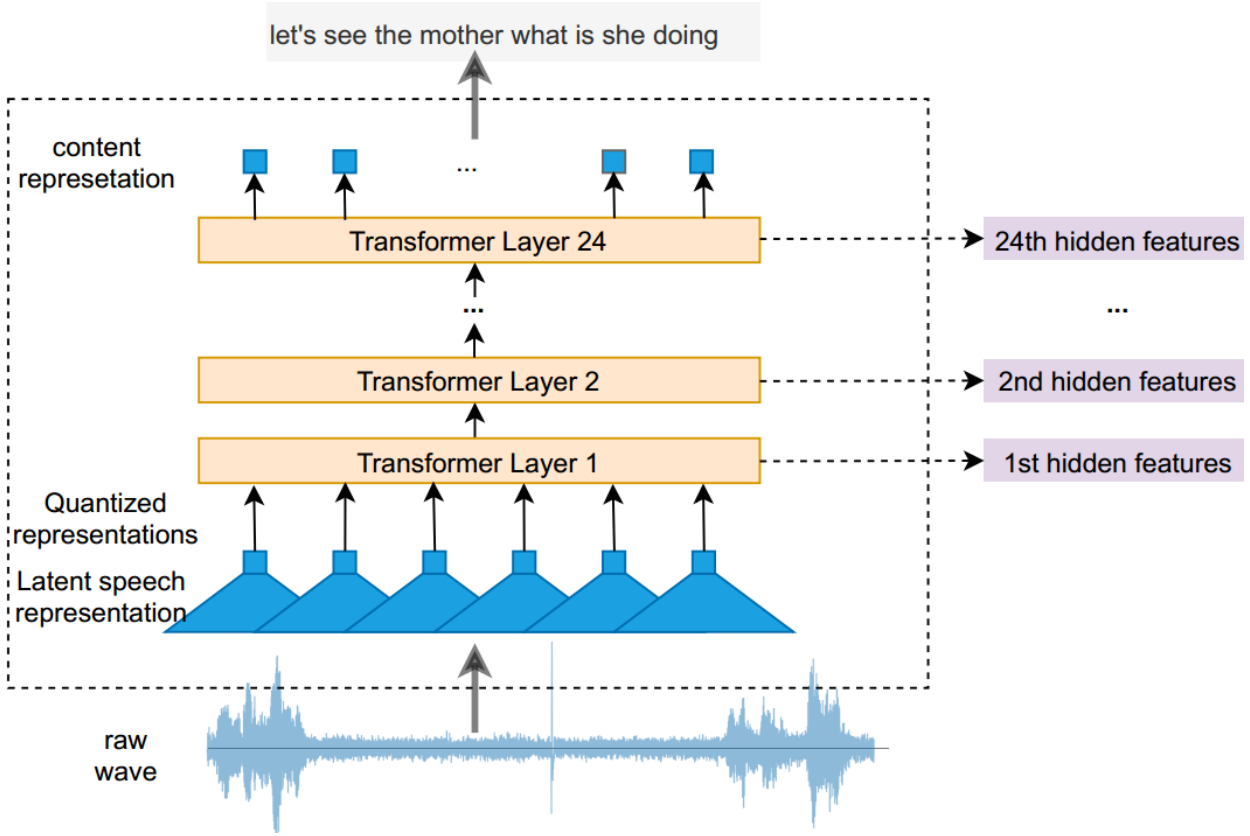


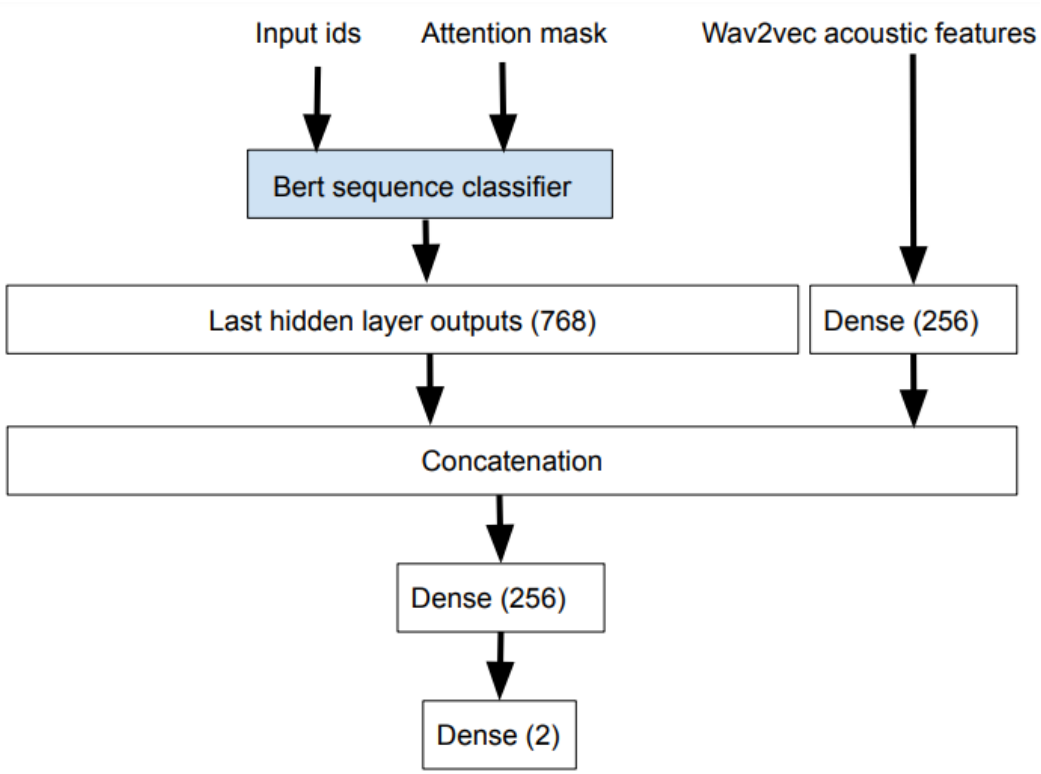
Table 3: *AD detection models, TB=Tree Bagger, CS=confidence scores, model parameter: e=epochs, mwl=max word length, ne=number of estimators, bs=batch size]*

Alias	Information Input	Classifier	Parameters
Model 1	w2v acoustic feat.	TB	ne=10
Model 2	fine-tuned w2v ASR	$\text{BERT}_{base.}$	mwl=512; e=8;bs=4
Model 3	w2v outputs fusion	BERT_{base}	mwl=512; e=8;bs=4
Model 4	ASR hypotheses+CS	BERT_{large}	mwl=105; e=1; bs=64
Model 5	ASR hypotheses+CS	BERT_{large}	mwl=100; e=1; bs=64

Wav2vec2.0 ASR System



Model 1

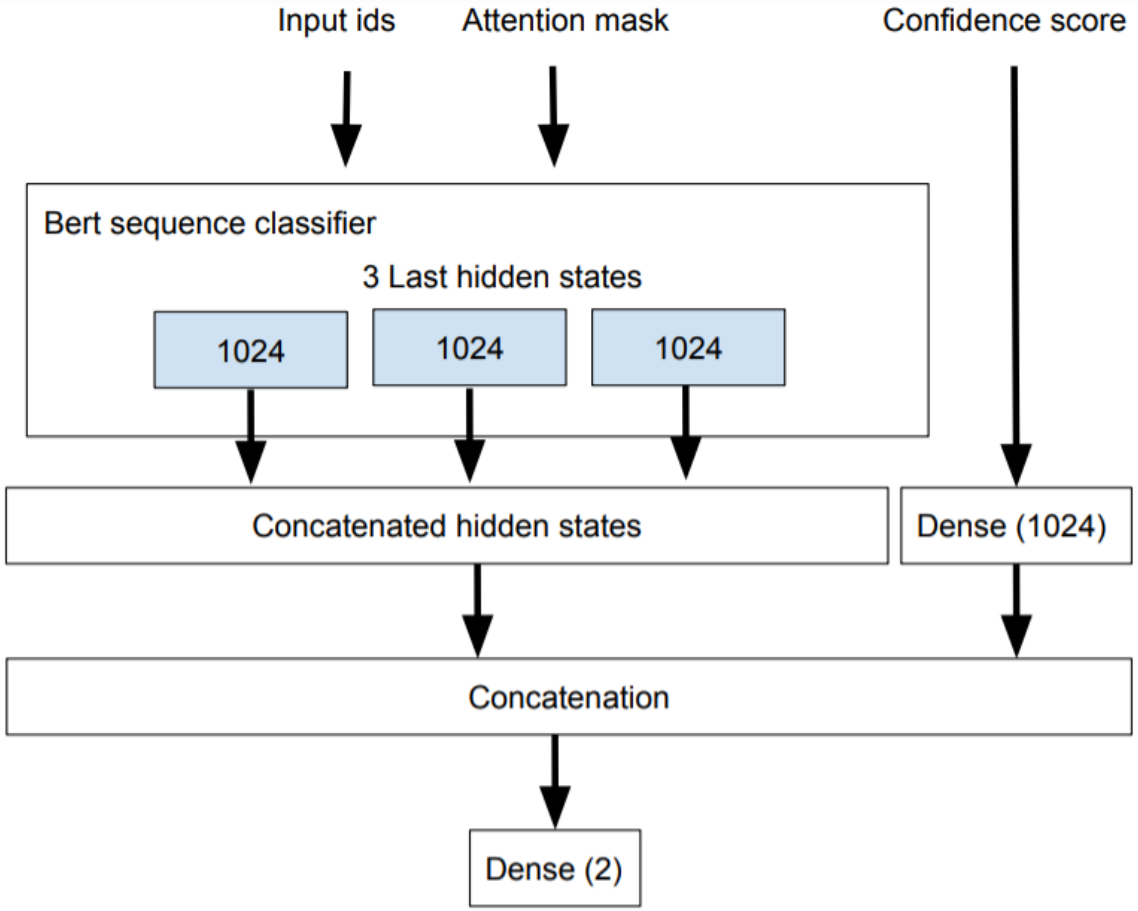


Model 3

The University Of Sheffield.

ASR Hypotheses and Confidence Scores

A range of ASR hypotheses can be derived from the decoder lattices using different parameters:
The language weights were between **6** and **10** with a word insertion penalty of **0**, **0.5** and **1** producing **30** different hypotheses for each recording.



The University Of Sheffield.

Experiment results

Table 4: *Final classification results on the evaluation set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	80.69	81.04	78.39	81.17	78.72	80.50	80.06
Model 2	83.63	83.02	79.82	83.08	79.32	82.25	81.59
Model 3	86.25	86.33	83.57	83.08	82.80	82.65	83.47
Model 4	**98.69**	**96.26**	**95.78**	**98.85**	**97.22**	**97.54**	**97.39**
Model 5	98.46	95.19	94.51	98.66	96.45	96.90	96.69

The acoustic accuracy is 80.06%.

The best linguistic accuracy is 97.39%

Table 5: *Final classification results on the test set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	72.50	77.42	80.56	68.57	76.32	72.73	74.65
Model 2	76.19	**86.21**	**88.89**	71.43	82.05	78.13	80.28
Model 3	76.92	81.25	83.33	74.29	80.00	77.61	78.87
Model 4	85.29	81.08	80.56	85.71	82.86	83.33	83.10
Model 5	**87.88**	81.58	80.56	**88.57**	**84.06**	**84.93**	**84.51**

The acoustic accuracy is 74.65%.

The best linguistic accuracy is 84.51%

The University Of Sheffield.

SP AND H

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

Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-based Alzheimer's Dementia Detection through Spontaneous Speech

Yilin Pan[1*], *Bahman Mirheidari*[1*], *Jennifer M Harris*[2,3], *Jennifer C Thompson*[2,4], *Matthew Jones*[2,4], *Julie S Snowden*[2,4], *Daniel Blackburn*[5], *Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, UK
[2]Division of Neuroscience and Experimental Psychology, University of Manchester, UK
[3]Department of Psychology, University of Exeter, UK
[4]Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, Salford UK
[5]Department of Neuroscience, University of Sheffield, UK
{yilin.pan, b.mirheidari, heidi.christensen}@sheffield.ac.uk

Abstract

Exploring acoustic and linguistic information embedded in spontaneous speech recordings has proven to be efficient for automatic Alzheimer's dementia detection. Acoustic features can be extracted directly from the audio recordings, however, linguistic features, in fully automatic systems, need to be extracted from transcripts generated by an automatic speech recognition (ASR) system. We explore two state-of-the-art ASR paradigms, Wav2vec2.0 (for transcription and feature extraction) and time delay neural networks (TDNN) on the ADReSSo dataset containing recordings of people describing the Cookie Theft (CT) picture. As no manual transcripts are provided, we train an ASR system using our in-house CT data. We further investigate the use of confidence scores and multiple ASR hypotheses to guide and augment the input for the BERT-based classification. In total, five models are proposed for exploring how to use the audio recordings only for acoustic and linguistic information extraction. The test results on best acoustic-only and best linguistic-only are 74.65% and 84.51% respectively (representing a 15% and 9% relative increase to published baseline results).
Index Terms: Automatic speech recognition, Alzheimer's dementia, computational paralinguistics

1. Introduction

Research on patients' speech has revealed that linguistic and acoustic abilities are affected even at the early stages of Alzheimer's dementia (AD) [1, 2]. Automatic methods for detecting such impoverishment has focused on extracting acoustic and linguistic information and learning distinguishable patterns for people with and without dementia. For embedding the *linguistic* information, multiple approaches e.g., word2vec [3], hierarchical neural network systems [4], and BERT [5, 6] has been proven to be effective in previous research. For extracting *acoustic* features, both the traditional pipeline systems based on conventional acoustic features [7, 8] and the more recent end-to-end systems for learning task-specific acoustic feature extraction [8, 9] have been explored.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) challenge is organized for advancing research into automatic AD detection [10]. It contains audio recordings of people describing the Cookie Theft picture. The challenge has not provided manual transcripts of the recordings, in order to reflect more real applications in which it is costly and time-consuming to provide human transcripts. However, automatic approaches to cognitive assessment rely on a combination of linguistic and acoustic information to detect symptoms more comprehensively [11]. To learn linguistic information, the most widespread approach is to first transcribe the audio into text with an automatic speech recognition (ASR) system. Not having manual transcripts of the recordings makes training a good ASR system challenging. To overcome some of these challenges two types of ASR paradigms are explored: i) the traditional ASR systems, based on a pipeline comprising of acoustic, language and lexical models, and ii) end-to-end systems, which directly map a sequence of input acoustic waveforms into a sequence of graphemes or words with an integral structure. For the traditional ASR system, time delay neural network (TDNN) has been used in our previous research and got an outstanding result on dementia-related speech recognition [12–14].

In this paper, BERT is used for extracting linguistic information. However, to mitigate the problems of using ASR-generated transcripts (with erroneous words and lacking punctuation), we explore the use of ASR lattice information. An ASR lattice provides time alignment, recognised words and confidence scores for different hypotheses. Usually, only the most likely hypothesis is selected for the subsequent AD detection [4, 12]. Here, we explore both the use of multiple hypotheses (can be seen as a form of data augmentation) as well as using the hypothesis with the highest word confidence scores (*HighConfHyp*). In addition, we incorporate the confidence scores into the BERT sequence classifier.

Wav2vec2.0 (denoted as *w2v* in the following), as a self-supervised end-to-end ASR system, can achieve similar performance as the traditional ASRs (supervised systems) but with less transcribed audio data. For the w2v structure, information embedded in the transformer layers has been explored in previous research [15, 16]. In our paper, w2v is used both for audio transcription and embedded acoustic feature extraction.

The contribution of the paper can be summarized as follows: (1). Using the acoustic features and automatic transcripts extracted by the w2v results in a superior performance compared to the baseline features reported in [10]. (2). Using mul-

[*] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1519

tiple ASR hypotheses and confidence scores as an input to the BERT system improves the performance compared to using just the best hypothesis as in previously proposed approached [12]. (3). Exploring the performance of feature fusion on the acoustic and linguistic information improves results on the provided test set. In the remainder of the paper, Section 2 presents the related work. Section 3 introduces the acoustic and linguistic features used. The experimental setup and the results are discussed in Section 4 and Section 5. Conclusions are given in Section 6.

2. Related work

In the ADReSS Challenge, the predecessor to the ADReSSo challenge, the organisers provided both the acoustic recordings and the corresponding manual transcripts. The best performances were achieved by [6] and [5] using acoustic-only and linguistic-only systems, respectively. Among the 34 participating teams, 7 out of 13 Interspeech-2020 papers used BERT for linguistic-based modelling, thus demonstrating the efficiency of BERT when manual transcripts are available. However, for fully automatic systems, the transcription task is handled by an ASR which introduces word-level errors. It has been found that fine-tuning using such noisy text data, can negatively impact the performance of BERT [17]. Our paper explores how to increase the performance of BERT-based classification when working with ASR transcripts.

Disfluencies and unclear pronunciation can decrease the performance of ASR systems, whilst at the same time be beneficial if the related information is used to inform the classification [5, 12]. In [5], the pause and disfluency annotation was used for punctuation generation providing important linguistic information in addition to the manual transcripts. In [12], ASR-transcribed words with low confidence scores were removed from the generated transcripts to successfully guide the linguistic information extraction. The time alignment information from the ASR output is used for designing the rhythm features for assisting the extracted acoustic features.

Table 1: *Analysis of HighConfHyp on HC and AD.*

parameters (mean&var)	**AD**	**HC**
word duration (s)	$.129 \pm (.013)$	$.120 \pm (.013)$
pause duration (s)	$.267 \pm (.589)$	$.126 \pm (.154)$
confidence score	$.867 \pm (.036)$	$.886 \pm (.033)$
#words/transcript	$65 \pm (1910)$	$81 \pm (3207)$

The lack of manual transcripts in the challenge means that it is difficult to train an ASR system tailored to the specifics of ADReSSo. It also makes it more challenging to evaluate the outputs of any transcripts produced for the ADReSSo data. To get around this, confidence scores are used as a proxy measure for accuracy, essentially replacing the information provided by monitoring WERs for different ASR systems during development. To assess the meaningfulness of using this approach, we analysed the transcripts produced by our system [12] when run on the ADReSSo dataset. Table 1 shows that the mean and variance of pause duration in the AD group was longer than the HC (healthy control) group. Likewise, fewer words were recognised in the AD group compared to the HC group. This might be because people living with AD tend to speak less and with more disfluencies. Moreover, a significant number of the words pronounced by people living with AD are typically not clear enough to be recognised correctly by the ASR as demonstrated in our previous work [12].

3. Acoustic and linguistic features

In this section, two ASR paradigms are introduced for transcribing the audio recordings for linguistic information extraction using BERT. The w2v, is also used for acoustic information extraction. The feature fusion is implemented on the ASR transcripts and the extracted acoustic features.

3.1. Automatic speech recognition

To transcribe the audio recordings into texts, we trained a conventional Kaldi-based ASR which produces decoding lattices allowing us to construct different hypotheses (extracted text from different paths in the lattice). For each hypothesis it is possible to calculate the confidence scores of the words, reflecting how confident the ASR is in recognizing each word.

Table 2: *Datasets used for training the ASR. Len.:the total length in hours/mins, Utts.:number of utterances, Spks.:number of speakers, and Avg.Utts.:Average utterance length in seconds.*

Dataset (No)	**Len.**	**Utts.**	**Spks.**	**Avg. Utts.**
DR INTERVIEWS (295)	64.3h	39.2k	736	5.9s
IVA (168)	26.7h	8.3k	219	11.5s
HALLAM (54)	26.14h	10.5k	139	4.8s
SHEFMAN CT (238)	3.9h	0.2k	238	11.5s
LIBRISPEECH (281241)	961.1h	281.2k	5466	12.3s
AMI (682)	95.5h	133.9k	171	2.6s

Since the manual transcripts of the training set of the challenge were not available and using any part of Dementia Bank not permitted, training a high performance ASR tuned to recognise spontaneous speech, ideally of people describing the Cookie Theft picture, was challenging. LIBRISPEECH is a well-known dataset containing almost 1000 hours of audio recordings of people reading books. It was used to build a base TDNN ASR following Kaldi's LIBRISPEECH recipe [18]. Since the dataset is read speech, a transfer learning technique was applied to adapt to spontaneous speech data using a number datasets: (AMI [19], DR INTERVIEWS, IVA, and HALLAM. Table 2 shows information such as length, number of utterances and speakers of the datasets. DR INTERVIEWS , and HALLAM are two datasets collected locally at the Sheffield Royal Hallamshire hospital. They contain audio interviews between neurologists and people with seizure/non-epileptic attacks, and dementia or other memory issues, respectively. In addition, the IVA dataset (also in-house) contains conversation between patients and an Intelligent Virtual Agent. Moreover, 238 Cookie Theft description from Sheffield and Manchester universities (SHEFMAN CT) were used for a second round of adaptation. Of these, a small subset of 20 out of 238 samples was held out for ASR testing and the rest were added to the other datasets for transfer learning. Following [20] (using both the structure and weights of the base ASR and then running one epoch of the DNN model to adapt to the new datasets) the acoustic model of our ASR was constructed. To train the language model, the four-grams with Turing smoothing was applied on the training set. Decoding on a held out set of the SHEFMAN CT data (10%) resulted in a WER of 8.23%.

3.2. Acoustic feature extraction

W2v as an end-to-end ASR paradigm was used for both the audio transcription and the acoustic feature extraction. W2v encodes raw wave $\mathcal{X}$ with multiple Convolutional Neural Networks (CNNs) into latent representations $Z \in z_1, ..., z_T$ for T time-steps. Before passing the inputs to the Transformer to get the textualized representation, first Z was fed to a quantization module [21] for masking [22–24]. The model was first trained on the unlabeled data and then fine-tuned on a small labelled dataset with a Connectionist Temporal Classification (CTC) loss [25, 26]. Research has shown that the representations underlying pre-trained w2v can capture the speaker features and language features [15] embedded in the acoustic recordings. The contextualized representations of the input raw wave were built into the 24 layers transformer architecture. For extracting the acoustic features from the w2v structure, the transformer layers' outputs were extracted as the acoustic representations of the input waveform segments. For each hidden layer output, the extracted hidden feature matrix $[N * feat\_dim]$ was averaged across the length N of the feature. At the same time, the transcribed words for the waveform were used for the following linguistic-based information modelling.

The pre-trained model[1] (named as pre-trained w2v for convenience) is adapted on the IVA dataset (see Table 2) for a better performance on the ADReSSo data. The IVA recordings were split into 60/20/20 parts training, evaluation and testing. After data adaptation, the WER decreased from 31.9% to 18.6% on this IVA test set. The ASR transcripts are used as the input of BERT for linguistic information modelling.

3.3. Combining linguistic and acoustic features

The acoustic features described in Section 3.2 and the linguistic features extracted from the last transformer layer in BERT by concatenating with a fully connected layer are combined as follows. The acoustic features $\mathcal{V} \in [N * feat\_dim]$ were averaged over feature numbers into a vector $v_1 \in feat\_dim$. After dimension reduction with a fully connected layer, the feature v_1 was concatenated with the BERT last-second layer output feature v_2 for fine-tuning.

3.4. Using ASR hypotheses and confidence scores

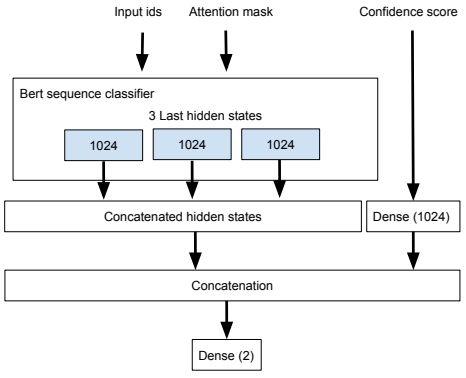


Figure 1: *The last three states of the BERT sequence classifier are concatenated with the confidence score input, before the last classification output layer.*

A range of ASR hypotheses can be derived from the decoder lattices using different parameters: the language weights were between 6 and 10 with a word insertion penalty of 0, 0.5 and 1 producing 30 different hypotheses for each recording. The average confidence scores of the words were calculated for each of the hypotheses. Two models (with different maximum word length of the input sequence: 105 and 100) were built using the uncased BERT large sequence classifier and hypotheses with the confidence scores. Figure 1 shows the structure of the models which was simply a concatenation of the last three hidden states of the BERT classifier with the input confidence score layer. The combination of these two layers was passed finally to the output layer to classify between AD and HC transcripts. Using a variety of outputs from the ASR (which we know are erroneous) alongside the corresponding confidence scores, could help the network to be trained more robustly on the words produced by the ASR.

4. Experimental setup

Table 3 shows the list of the five models submitted to the challenge. Model 1 and Model 2 correspond to the acoustic features and ASR transcripts as output by the w2v model described in 3.2, while Model 3 corresponds to the feature fusion proposed in Section 3.3. For these three models, 10-fold cross-validation (CV) was applied[2]. Then the results (Section 5) were averaged across the 10 folds, and the test labels were estimated by the majority voting on the predict labels from the 10 folds trained models. However, for Model 4 and Model 5 corresponded to Section 3.4, since it was a time-consuming process to run multiple fold-based evaluations, instead a single evaluation set was constructed by holding out 20% of the training set.

4.1. Model setup

BERT-for-Sequence-Classification[3] [27] was used for modelling the linguistic information. Two configurations of BERT models were used with a transformer layer inside the models of 12 layers (BERT$_{base}$) and 24 layers (BERT$_{large}$). Two dense layers were added to the BERT$_{base}$ for feature fusion with 256 dimensions. To fine-tune the BERT$_{base}$ and BERT$_{large}$ using the ASR-derived transcripts of the ADReSSo training set, the parameters were set as in Table 3. For fine-tuning the w2v, 168 IVA recordings, shown in Table 2, were used.

4.2. Model selection

We used our evaluation set to select the five models submitted to the challenge. As part of this process, we also evaluated a number of proposed models based on our and other's previous work. Using the approach in [5], the performance of the HighConfHyp on BERT$_{base}$ using 10 fold averaged accuracy was 72.77% on our evaluation set. Also, we replicated the experiments in [5] on the BERT3p model (the best performance with BERT in the paper) based on the HighConfHyp. The time alignment information was used for generating the punctuation insertion instead, and the 10-fold CV averaged accuracy on the evaluation set was 75.06%. Inspired by the experiments in [12], the transcribed words in the HighConfHyp are selected with the confidence scores by replacing the word with the confidence score lower than 0.87 (selected based on the average confidence score of the two classes in Table 1) by <*unk*>. The 10-fold CV

[1] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

[2] nine folds of training set for training and one fold for evaluation

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

averaged accuracy on the evaluation set is 76.65%. The evaluation result on the pre-trained w2v transcripts is 78.01% under the same parameters as in Model 2.

Table 3: *AD detection models, TB=Tree Bagger, CS=confidence scores, model parameter: e=epochs, mwl=max word length, ne=number of estimators, bs=batch size]*

Alias	Information Input	Classifier	Parameters
Model 1	w2v acoustic feat.	TB	ne=10
Model 2	fine-tuned w2v ASR	$\text{BERT}_{base.}$	mwl=512; e=8;bs=4
Model 3	w2v outputs fusion	BERT_{base}	mwl=512; e=8;bs=4
Model 4	ASR hypotheses+CS	BERT_{large}	mwl=105; e=1; bs=64
Model 5	ASR hypotheses+CS	BERT_{large}	mwl=100; e=1; bs=64

5. Results

5.1. Acoustic feature comparison

For classifying the extracted acoustic features, four linear classifiers were selected, namely decision trees (DT), nearest neighbour (KNN), TB and support vector machines (SVM). The acoustic features are normalized before classifying. The acoustic features' evaluation results corresponding to the 24 transformer layers extracted from the pre-trained w2v model and fine-tuned model are shown in Figure 2 respectively. As shown, the highest accuracy (80.08%) was achieved by the TB classifier with the 16th hidden layer features extracted from the pre-trained model; this was selected as model 1.

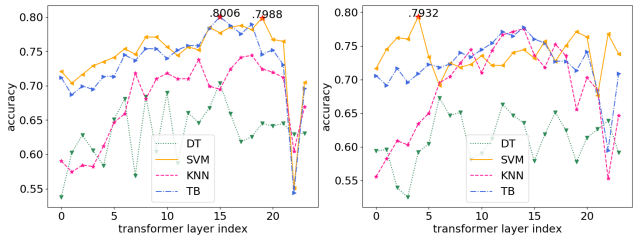


Figure 2: *Comparing classifiers using acoustic features extracted from the wav2vec2-large-960h-lv60 pre-trained model (left) and fine-tuned model (right).*

5.2. Evaluation results

The baseline results on the leave-one-subject-out (LOSO) evaluation (CV) set reported by the authors of the challenge [10] were 78.92% and 72.89% on the acoustic and linguistic-based models respectively. The results corresponding to our five models are listed in Table 4. All models outperform the CV baseline[4]. In particular, the feature fusion result on the evaluation set (83.47%) from model 3 is better than the acoustic-only result (80.06% for model 1) and linguistic-only (81.59% for model 2), though not as good as the multiple ASR hypotheses and confidence scores corresponded models (model 4 and model 5), which achieved accuracies of 97.39% and 96.69% respectively.

[4]A direct comparison on evaluation data is difficult, as we have not been able to evaluate using LOSO because of time constraints and have instead evaluated using 10-fold CV (models1, 2 and 3) or on 20% held out data (models 4 and 5).

Table 4: *Final classification results on the evaluation set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	80.69	81.04	78.39	81.17	78.72	80.50	80.06
Model 2	83.63	83.02	79.82	83.08	79.32	82.25	81.59
Model 3	86.25	86.33	83.57	83.08	82.80	82.65	83.47
Model 4	**98.69**	**96.26**	**95.78**	**98.85**	**97.22**	**97.54**	**97.39**
Model 5	98.46	95.19	94.51	98.66	96.45	96.90	96.69

Table 5: *Final classification results on the test set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	72.50	77.42	80.56	68.57	76.32	72.73	74.65
Model 2	76.19	**86.21**	**88.89**	71.43	82.05	78.13	80.28
Model 3	76.92	81.25	83.33	74.29	80.00	77.61	78.87
Model 4	85.29	81.08	80.56	85.71	82.86	83.33	83.10
Model 5	**87.88**	81.58	80.56	**88.57**	**84.06**	**84.93**	**84.51**

5.3. Test results

The baseline results on the test set were 64.79% and 77.46% for the acoustic and linguistic systems, respectively [10]. The final classification results of the five models we proposed are shown in Table 5. The acoustic-only and best linguistic-only results achieved 74.65% and 84.51% accuracy respectively (models 1 a& 5), which all outperforms the baseline models. Interestingly, the feature fusion based model (model 3) performed better than model 1 and model 2 on the evaluation set, but not as well as the linguistic-only models on the test set. This might indicate a mismatch between the evaluation set and the test set.

6. Conclusion

In our paper, two ASR paradigms were adopted for linguistic and acoustic feature extraction. For modelling the linguistic information, multiple ASR hypotheses and confidence scores were passed to the pre-trained BERT for model tuning on the ASR transcripts from the ADReSSo training set. The acoustic features were extracted from the transformer outputs of the pre-trained w2v model. The BERT model based combination of the acoustic and linguistic information improved the performance of the classifier on the evaluation set, but not on the test set. In the future, the combination between the acoustic information and the multiple ASR hypotheses is expected to be explored to improve the test set performance.

7. Acknowledgements

This work is supported by the European Union's H2020 Marie Skłodowska-Curie programme (TAPAS; Grant Agreement No. 766287), the Rosetrees Trust and the Stoneygate Trus (COMPASS, Grant Agreement No. M934 and the Fast ASsessment and Treatment in Healthcare funded by Engineering and Physical Science Research Council (EPSRC) (Reference. EP/N027000/1).

8. References

[1] 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.

[2] 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.

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

[4] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting ad." in *INTERSPEECH 2019*. ISCA-International Speech Communication Association, 2019, pp. 4105–4109.

[5] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," *INTERSPEECH 2020*, pp. 2162–2166, 2020.

[6] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech," *INTERSPEECH 2020*, pp. 1–5, 2020.

[7] 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.

[8] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[9] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," *INTERSPEECH 2020*, pp. 4806–4810, 2020.

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

[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] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 4961–4965.

[13] B. Mirheidari, D. Blackburn, R. O'Malley, A. Venneri, T. Walker, M. Reuber, and H. Christensen, "Improving cognitive impairment classification by generative neural network-based feature augmentation," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2527–2531.

[14] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2732–2736.

[15] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[17] A. Srivastava, P. Makhija, and A. Gupta, "Noisy text data: Achilles' heel of BERT," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020, pp. 16–21.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88. Citeseer, 2005, p. 100.

[20] V. Manohar, D. Povey, and S. Khudanpur, "Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.

[21] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[22] 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.

[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," *arXiv preprint arXiv:1907.11692*, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[25] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios

Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velazquez, Piotr Żelasko, Jesús Villalba, and Najim Dehak
Johns Hopkins University

Introduction

- Prevalence of Alzheimers (AD) and related dementias in the USA for populations older than 65 is 11.5%
- Language impairment is commonly observed sign in AD affected patients among other signs
  - Inability to find correct words – patients likely describe the word meaning but can not recall the word itself
  - Frequent use of verbal filters such as /um:/ or /eh:/
  - Stuttering, repetition of ideas and questions happen in advanced stages
  - Frequent difficulties in forming simple sentences
- Can we detect AD and assess its severity from speech and its content alone? It could assist doctors in diagnosis
- Can we exploit advances in speech and language technologies to detect Alzheimer better?
  - Exploration of efficacy of transfer learning from speech recognition, speaker recognition, audio event detection models
  - Evaluation of several ASR transcriptions using linguistic models

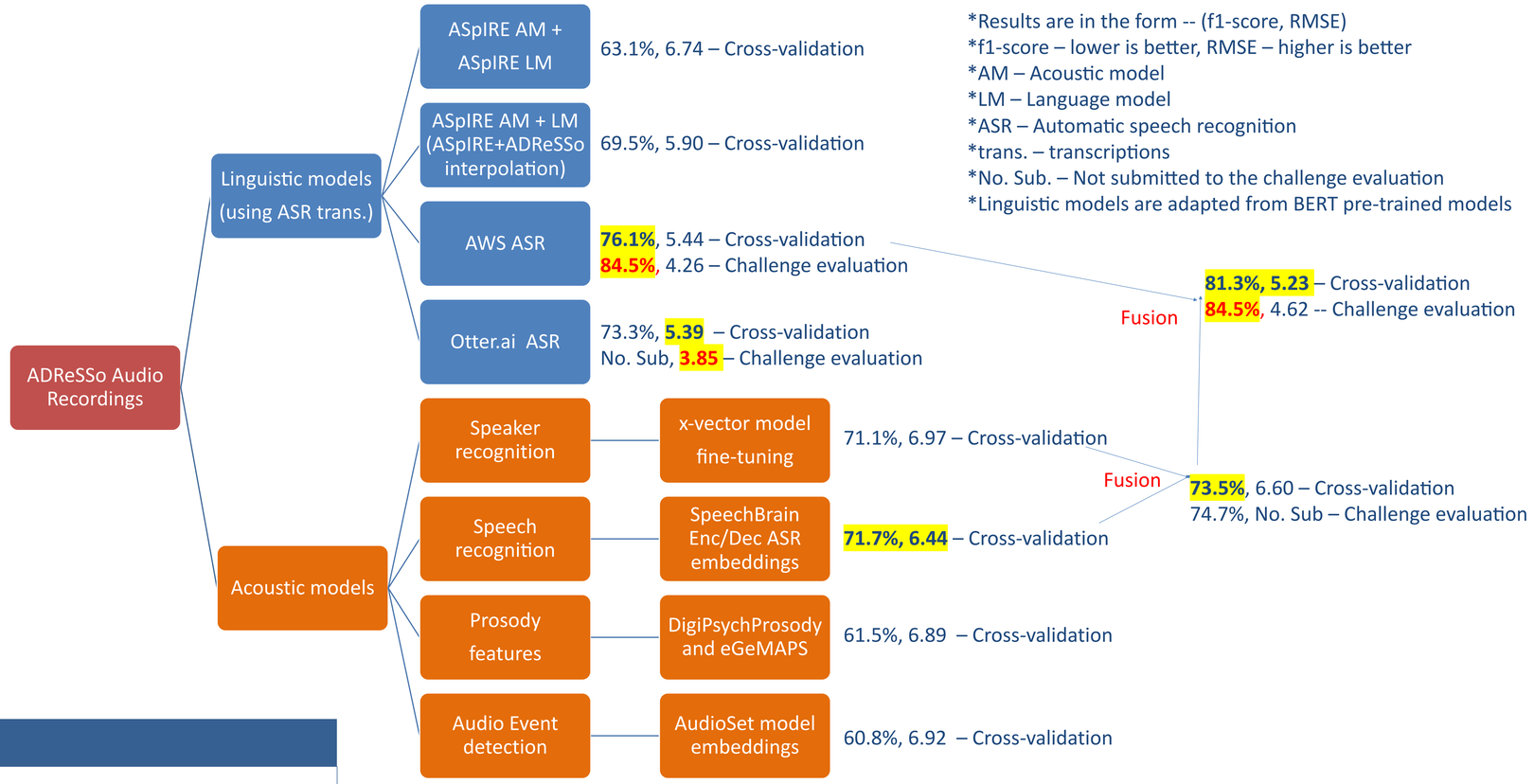
The ADReSSo challenge

- AD detection task: classification of subjects with and without AD using a short speech session
- MMSE prediction task: prediction of mini-mental status evaluation (MMSE) from the same short speech sessions used for AD detection
- Training dataset: 87 recordings from subjects with AD and 79 from control subjects
- Evaluation dataset: 71 recordings (labels are not known to the participants)

Experimental setup

- Detection models specifics:
  - Logistic regression is used for classification on embeddings
  - Fine-tuning – Replace the last layer of the pre-trained model with task-specific layer and optimize cross-entropy loss function
  - Choosing best epoch based on best f-score on development set
- MMSE prediction model specifics:
  - Gradient boosting regressor is used for MMSE prediction on embeddings
  - Fine-tuning – Replace the last layer of the pre-trained model with task-specific layer and optimize RMSE loss function
  - Choosing best epoch based on best RMSE on development set
- Adam optimizer with default parameters is used for training
- Score-level fusion -- we extract output scores of all the recordings from the considered models and their concatenated vector acts as input to fusion model (logistic regression)
- 10-fold Cross-validation to estimate models reliably
  - For the evaluation data, we average predictions from all fold models and feed it to fusion model
- Metrics:
  - F1-score (harmonic average of precision and recall) for detection task
  - Root mean square error (RMSE) for MMSE prediction task

Methods and results



Conclusions and future work

- Linguistic models:
  - Transcriptions obtained with language model interpolation performed 6.4% f-score better
  - Commercial ASRs outperformed ASpIRE models
  - Improvement in transcriptions translates to better results in AD detection and MMSE prediction
- Acoustic models:
  - Adaptation of x-vector model and SpeechBrain ASR embeddings performed similarly, and their fusion provided 1.8% improvement
    - Fine-tuning of x-vector models with augmentation did not help (results not shown above)
  - Prosody features and embeddings extracted from event detection models were producing almost random predictions
- Linguistic + Acoustic Fusion:
  - Improvements in detection score (5.2%), MMSE score (0.21) in cross-validation but did not improve on challenge test set
- Linguistic models performed best on challenge test set for both detection and MMSE prediction tasks
- In future, we plan to explore multi-modal techniques to exploit temporal cues in acoustic and linguistic modalities

Contact

Raghavendra Reddy Pappagari
Center for Language and Speech Processing,
Johns Hopkins University, USA
Email: rpappag1@jhu.edu

References

1. Raghavendra Pappagari, J. Cho, L. Moro-Vel´azquez, and N. Dehak, “Using state of the art speaker recognition and natural language processing technologies to detect alzheimer’s disease and assess its severity,” Proc. Interspeech 2020, pp. 2177–2181, 2020.
2. Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, “x-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” in *IEEE International Conference on Acous- tics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

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

Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios

Raghavendra Pappagari[1], *Jaejin Cho*[1], *Sonal Joshi*[1], *Laureano Moro-Velazquez*[1], *Piotr Żelasko*[1,2], *Jesús Villalba*[1,2], *Najim Dehak*[1,2]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{rpappag1, jcho52, sjoshi12, laureano, pzelasko, jvillal7, ndehak3}@jhu.edu

Abstract

In this study, we analyze the use of speech and speaker recognition technologies and natural language processing to detect Alzheimer disease (AD) and estimate mini-mental status evaluation (MMSE) scores. We used speech recordings from Interspeech 2021 ADReSS$_o$ challenge dataset. Our work focuses on adapting state-of-the-art speaker recognition and language models individually and later collectively to examine their complementary behavior for the tasks. We used speech embedding techniques such as x-vectors and prosody features to characterize the speech signals. We also employed automatic speech recognition (ASR) with interpolated language models to obtain transcriptions used to fine-tune the BERT models that classify and assess the speakers. Our results indicate that the fusion of scores obtained from the multiple acoustic and linguistic models provides the best detection results, suggesting that they contain complementary information. A separate analysis of the models indicates that linguistic models outperform acoustic models in detection and prediction tasks. However, acoustic models can provide better results than linguistic models under certain circumstances due to the errors in ASR transcriptions, which indicates that the performance of linguistic models relies on the performance of ASRs. Our best models provide 84.51% accuracy in automatic detection of AD and 3.85 RMSE in MMSE prediction.

Index Terms: Alzheimer Disease, Automatic Speech Recognition, Mini-Mental Status Evaluation

1. Introduction

The most common signs of Alzheimer disease (AD)[1], are memory decline, disorientation, confusion, and behavior changes. This leads to loss of independence, having a clear impact on patients, their families, and the society [2]. The prevalence of AD and related dementias in the USA for populations older than 65 years old is 11.5%, with an increasing incidence due to the improvement in life expectancy in the coming decades, which would double the associated burden by 2060 [3].

While two of the most typical signs of AD are memory and cognitive decline, language impairment is also common, as it is linked to cognitive and memory-related problems and neurodegenerative processes. In this respect, speech technologies can deliver new precision medicine tools that will provide an objective quantitative analysis and reliable proof, analysis, comparison, and circulation for a faster diagnosis.

The literature suggests some common signs in the speech of AD patients related to articulatory aspects such as apraxia of speech [4] or others linked to communication and word retrieval deficits such as progressive, logopenic, or anomic aphasia [5, 6], or anomia [4]. In this respect, pause and silence-related features allow characterizing the loss of verbal fluency, which is associated with AD [7]. These problems of verbal fluency are caused, in part, by the difficulties that patients have in recalling words, finding the appropriate vocabulary, or finishing sentences. The use of verbal fillers such as /um:/, or /eh:/ or the description of a word instead of the use of that word, are also common. In more advanced stages, stuttering, repetition of ideas and questions, and difficulties forming simple sentences become frequent [7].

In the Interspeech 2020 ADReSS challenge [8], numerous teams proposed different approaches to detect AD and automatically predict mini-mental status evaluation (MMSE)[2] in a dataset containing speech and manual transcriptions from 78 AD patients and 78 sex and age-matched controls. Whereas some of the participant teams focused on using either speech or linguistic approaches, the results from several teams indicate that approaches containing a combination of different linguistic aspects and, in some cases, acoustic aspects lead to better results [9, 10, 11, 12, 13], providing detection accuracy over 75% in the evaluation subset. Some authors employed term frequency–inverse document frequency (TF-IDF) features such as grammatical dependency and universal dependency features [9, 14], with different classifiers such as XGBoost or logistic regression. Other authors used a transformer-based pre-trained language model (LM) based on bidirectional encoder representations from transformers (BERT) [10, 15, 16, 14, 17, 13, 18], and other neural network approaches such as bi-directional Hierarchical Attention Networks [11] or Transformer XL [13]. All of the linguistic approaches used the manual transcriptions provided by the challenge organizers, and none of them analyzed the use of any automatic speech recognition (ASR) system to obtain transcriptions using audio in the detection or regression tasks.

Approaches using acoustic modeling involved the use of x-vectors [10, 17], i-vectors [17], bag of audio words [11], spectral and cepstral features with different classifier backends [9, 16, 11], acoustic features obtained with OpenSMILE [8, 19, 20, 13], and VGGish [13, 18] with heterogeneous results that, in general, did not outperform linguistic approaches.

In this study, we propose the use of several acoustic and linguistic models to detect and assess AD for the Interspeech 2021 ADReSS$_o$ challenge [21]. The main difference from the 2020

[1]The possessive form has been deliberately removed in this article, following the World Health Organization and the US National Institutes of Health recommendations [1].

[2]MMSE ranges between 0 and 30 and is used to assess the dementia status of patients, being values higher than 24 considered as normal cognition.

http://dx.doi.org/10.21437/Interspeech.2021-1850

challenge is that the manual transcriptions of the audio recordings were not included in the dataset. Thus, we employed ASR systems with adapted LMs, and commercial ASR platforms to obtain transcriptions from the audio and used linguistic models using the obtained text. At the same time, we employed acoustic models using several acoustic representations derived from state-of-the-art speaker and speech recognition technologies as well as prosodic features to detect and assess AD automatically.

The code of the experiments is being shared by the authors of this paper[3].

2. The ADReSS$_o$ challenge

In this paper, we addressed two of the three tasks proposed by the challenge organizers:

- *AD detection task* - automatic differentiation between participants with and without AD using a short speech session.
- *MMSE prediction task* - automatic prediction of the participant's MMSE using the same dataset of the AD detection task.

2.1. The Dementia Bank-ADReSS$_o$ 2021 dataset

The ADReSS$_o$ challenge dataset, described in [21], contains the *diagnosis dataset* with speech from speakers with and without AD. The recordings include a picture description, employed for AD detection and MMSE regression tasks. In most cases, these recordings consist of a participant's interaction with one investigator under several recording conditions with different types of background noise. The dataset is divided into *training* and *evaluation* subsets. The *training* subset contains 87 recordings from speakers with AD and 79 from control subjects and the *evaluation* subset, a total of 71 recordings.

3. Methods

In this study, we analyzed multiple acoustic and linguistic modeling approaches to carry out the automatic AD detection and MMSE prediction tasks proposed in Section 2. Then, we performed a score-level fusion of these approaches to obtain new predictions, as indicated in the following sections. There were two types of experiments:

- Cross-validation: performed by training and testing with the *training subset*, using a 10-fold scheme where class distributions were consistent over the folds.
- Evaluation: obtained by testing the models trained on the *training subset* with the *evaluation subset*. For each separate approach, we propagated the *evaluation subset* through an ensemble classifier that averages the scores from the 10 cross-validation models.

3.1. Acoustic modeling

We used several types of acoustic modeling to characterize the speech from the dataset and represent the speakers' articulatory, prosodic and phonatory traits. On the one hand, we used an end-to-end classifier by fine-tuning an x-vector model [22]. On the other hand, we extracted different types of embeddings and acoustic features from available libraries as input to logistic regression and XGBoost classifiers.

[3]https://github.com/sonal-ssj/ADReSSo_2021_JHU

3.1.1. x-vectors

An x-vector model is a deep neural network that generates one single vector (embedding) per recording, characterizing the full signal. 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 [23, 24] and can be used to characterize the influence of neurological diseases, such as Parkinson disease on speech [25]. The x-vector architecture considered in this study is the same as the employed in [26], and contained three main parts: an encoder network to extract frame-level representation from Mel-frequency cepstral coefficient (MFCC), a global temporal pooling layer to produce the embedding (x-vector), and a feed-forward classification network to produce speaker class posteriors. For the encoder, we used a ResNet-34 [27] structure consisting of a sequence of 2D convolutional layers with residual connections between them. The pooling network comprises a multi-head attention layer and operates on the ResNet output. Different heads are designed to capture different speech aspects of the input signal. We concatenated the attention heads output and pass it through a fully connected layer whose output is passed through an utterance-level classifier to obtain model decision.

We pre-trained this model for speaker recognition using VoxCeleb1, VoxCeleb2, NIST SRE4-10, and Switchboard datasets similarly as in [10]. Then, we replaced the last part of the model, the fully connected layer, to detect AD using softmax in the output or provide MMSE values using linear activation in the output depending on the task, and retrained the whole model. Additionally, we trained a second model in the same way but with noise and music augmentation (x-vectors augm), as indicated in [26] to obtain more robust representations. Both types of models (x-vectors and x-vectors augm) were trained considering two different frame-lengths: 25 and 250 ms. We note that as x-vector models were pre-trained for speaker classification, our models could perform well on AD detection using attributes related to speaker classification instead of using AD characteristics. By evaluating on unseen subjects, however, we made sure that our model's performance is reflective of its ability to capture AD characteristics.

3.1.2. Embeddings and prosody features

Encoder-decoder ASR embeddings. We computed embeddings using the encoder of an encoder-decoder ASR system included in the SpeechBrain library [28]. We used an acoustic model trained on LibriSpeech [29], that consists of an encoder with convolutional recurrent deep neural networks (CRDNN) architecture followed by a bidirectional LSTM and a fully connected layer to obtain the acoustic representation that we call as "encoder-decoder ASR embeddings" (SB Enc/Dec).

Prosody features. Previous studies have found that temporal features of AD patients differ from those of controls as the patients tend to have more silent pauses than controls [7]. We used DigiPsychProsody [4] to compute prosody features. These included total speech time, total pause time, percentage pause time, speech pause time, mean pause duration, and pause variability. These features are computed using 3 different intensities of the WebRTC[5] Voice Activity Detector. We obtained these features: (1) for the entire conversation recordings (2) per speaker – investigator and patient.– In this last case, we used the

[4]https://github.com/NeuroLexDiagnostics/DigiPsych_Prosody
[5]https://github.com/wiseman/py-webrtcvad

segmentation files given by the organizers to separate speech for each speaker. When we used segmentation, we concatenated the prosody features obtained from investigator and patient.

VGGish features. VGGish is a feature embedding front-end for audio classification models that has provided good results in AD detection [13, 18]. We used a pre-trained model [6] trained using the AudioSet dataset [30].

eGeMAPS Features. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features are a selected standardized set of statistical features that characterize affective physiological changes in voice production. We extracted these features for the entire recording as we expect them to capture overall speaker characteristics.

Embeddings and prosody features classification. The feature vectors obtained with SpeechBrain, the prosody features extractor, VGGish and eGeMAPS were employed to carry out the different challenge tasks in combination with logistic regression and XGBoost classifiers. All possible combinations of representations and classifiers were evaluated in cross-validation.

3.2. Linguistic modeling

3.2.1. Automatic speech recognition

Since the challenge data does not include human-annotated transcripts, we used ASR models to transcribe the recordings. As the recordings contain conversational speech and has noticeable noise and reverberation – with the microphones often being located far from speakers – we used a pre-trained ASpIRE recipe model[7] in Kaldi [31]. ASpIRE was a far-field speech recognition challenge held by IARPA, and the model in the recipe is trained on the English portion of the Fisher corpus [32], which is conversational speech, with data augmentation through room impulse response convolution and background noise [33]. This model is denoted as ASR 1 in our experiments.

Then, to improve the transcription quality, we interpolated the ASpIRE LM with an LM trained on automatic transcripts of the target domain (*training subset* of ADReSS$_o$ 2021) obtained with ASR 1. This interpolated model will provide more likelihood to frequent words of the target domain, leading to a lower word error rate. This system is denoted as ASR 2.

Lastly, to obtain other automatic transcripts from ASR systems trained on more varied acoustic data and possibly cover multiple linguistic domains, we employed two commercially available ASR systems: Amazon Web Services (AWS) and the Otter.ai ASR models denoted as ASR 3 and ASR 4, respectively. Since challenge data does not have manual transcriptions, we could not compare how different ASRs perform in regard to word error rate (WER) on the data.

3.2.2. BERT language model

We modeled the linguistic-phonological manifestations of AD using a pre-trained LM, BERT [34], on the automatic transcriptions of the speech recordings. BERT has provided state-of-the-art performances in multiple applications such as question answering, natural language inference, sentence, and word prediction, sentiment prediction, among many others [35]. After fine-tuning BERT, the transformer-based model can be used to model language context, flow, and complexity in the tasks of interest of this study [10]. In similar lines, previous works have shown promising results in other tasks such as depression detection [36] and sentiment analysis [37].

The BERT architecture consists of self-attention layers and feed-forward layers, similar to transformer encoder layers. The inputs of the model were tokens from the automatic transcript using WordPiece tokenizer [38]. The input token sequence was processed through the multiple encoder layers until the penultimate layer to obtain embeddings for each token. Then, the sequence of token embeddings was pooled to pass through a last linear layer to obtain the final prediction. In our case, we used a pre-trained BERT model and adapted it to our tasks (AD detection and MMSE prediction) in the following manner:

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

For each iteration of the cross-validation experiments, 8 folds from the *training subset* were employed for BERT fine-tuning, 1 fold for early stopping, and the remaining fold for testing. We fine-tuned the models for up to 5 epochs.

3.3. Model fusion

We explored model fusion by using the output scores of the models as the inputs of a logistic regression classifier to obtain a final prediction (detection or MMSE prediction). We first obtained fused models by combining the acoustic models. Then, we combined the best acoustic model and the best linguistic model, As for linguistic modeling, we used just BERT with different ASR transcriptions. Finally, we combined the best fused acoustic model with the best linguistic model, i. e., we first fused several acoustic models, and the resulting scores were fused with those from BERT.

4. Results and discussion

Acoustic models. Table 1 contains the results of the different acoustic models for the detection and MMSE prediction using logistic regression as a classifier.[8] Detection results are reported using accuracy (%) and MMSE prediction results, using root mean square error (RMSE).

Results in the first block of Table 1 indicate that all of the acoustic features provide some differentiation between classes. SB Enc/Dec embeddings with logistic regression and x-vectors model with 250 ms frame-length provide the best cross-validation results for AD detection, whereas SB Enc/Dec embeddings and eGeMAPS provide the best RMSE values for MMSE prediction. Prosody features related to pause times and pause vs. speech ratios characterize the loss of verbal fluency, which is associated with AD [7] and, whereas these provide only 61% accuracy, the results suggest that these help to automatically differentiate between speakers with and without AD.

Linguistic models. The second block in Table 1 includes the accuracy and RMSE results of four BERT models fine-tuned with automatic transcriptions obtained with four ASR systems. The best detection result is obtained from the one fine-tuned with the automatic transcriptions from ASR 3, and best MMSE prediction with ASR 4, both of which are based on commercially available ASR systems. These results suggest that transcription errors lead to worse prediction and detection results

[6] https://github.com/tensorflow/models/tree/master/research/audioset, https://github.com/harritaylor/torchvggish

[7] https://kaldi-asr.org/models/m1

[8] We also used other classifiers like XGBoost, however, since logistic regression outperformed other classifiers, due to space constraints, we have included results for only logistic regression.

Table 1: *Best AD detection accuracy (%) and MMSE prediction RMSE using acoustic (top), linguistic (middle), and fusion (bottom) modeling during cross-validation.*

Model	Detection accuracy(%)	MMSE RMSE
Acoustic		
x-vectors	69.3	7.18
x-vectors (250 ms)	71.1	6.97
x-vectors augm	58.4	6.92
x-vectors augm (250 ms)	63.9	7.01
VGGish	60.8	6.92
SB Enc/Dec	**71.7**	**6.44**
Prosody (20 ms)	60.8	6.94
Prosody (30 ms)	59.0	6.89
Prosody per speaker (20 ms)	61.5	6.96
Prosody per speaker (30 ms)	61.5	7.05
eGeMAPS	63.3	6.76
Linguistic		
BERT (ASR 1)	63.1	6.74
BERT (ASR 2)	69.5	5.90
BERT (ASR 3)	**76.1**	5.44
BERT (ASR 4)	73.3	**5.39**
Fusion		
All acoustic models	69.9	6.74
x-vector (250 ms), SB Enc/Dec	72.3	6.55
x-vector, x-vector (250 ms), SB Enc/Dec	73.5	6.60
x-vector, x-vector (250 ms), SB Enc/Dec, Prosody (20 ms)	72.9	6.76
BERT (ASR 3), {x-vector, x-vector (250 ms), SB Enc/Dec }	**81.3**	**5.23**

using linguistic approaches. Whereas in our previous work [10], linguistic models trained with manual transcriptions always outperformed acoustic models, in this study, some acoustic models from Table 1 outperform those linguistic models that possibly have high word error rates in their transcriptions. On the other hand, there is a remarkable improvement in the results of the linguistic model with ASR 2 transcription compared to with ASR 1 transcription. This indicates that interpolating the LM in ASR with the automatic transcription from the in-domain recordings can improve the final linguistic modeling. Therefore, new linguistic-based diagnostic tools will benefit from LM interpolation when the speech tasks (cognitive tests) are known.

Score-level fusion. Results of score fusion are included in the third block of Table 1. All of these values are reported using logistic regression as the fusion back-end, since it led to the best results. The fusion of the scores of all acoustic models led to worse results than the fusion of only two or three acoustic models. For score fusion with only acoustic models, the score fusion of SB Enc/Dec with several modalities of x-vector models showed the best result with 73.5% accuracy in cross-validation. However, the fusion of the acoustic models does not reduce the RMSE in the prediction task. The fusion of the linguistic model trained with the automatic transcriptions from ASR 3 and three acoustic models provides the best cross-validation results of the study, 81.3% and lowest RMSE in prediction, 5.23. This coincides with the findings of past work [10] that suggests that acoustic and linguistic approaches can have complementary information for detection and assessment of AD.

Evaluation results. Following the challenge rules, we submitted the scores from five different models on the detection task and other five on the MMSE task for challenge evaluation. Table 2 includes the results of that evaluation in terms of precision, recall, F1-score, and accuracy for detection, and RMSE in the prediction task. The trend of the results is similar to those in the cross-validation. Notably, the linguistic approach employing the ASR 3 transcription and its fusion with acoustic approaches provide the best results in terms of accuracy. However, in the evaluation result, the linguistic approach employing the ASR 4 transcriptions provides the best MMSE prediction. In general, the evaluation results are better than those obtained using cross-validation. One possible reason can be the use of ensemble models, built with the 10 cross-validation models for each of the submitted approach for evaluation, prevents overfitting.

Table 2: *ADReSS$_o$ challenge evaluation results for the detection and prediction tasks. Best results are marked in bold. Ac. fusion refers to the fusion of scores from acoustic models x-vector, x-vector (250 ms) and SB Enc/Dec. Global fusion refers to the fusion of Ac. fusion scores with BERT (ASR 3) scores. Following the challenge rules, 5 models were submitted for evaluation of detection, and other 5 models for prediction tasks. * indicates that the system was not submitted for evaluation*

		Detection			*Prediction*
Model	Class	Rec/Prec	F1	Accu (%)	RMSE
Baseline [21]	CC	0.78/0.80	0.78	78.87	5.28
	AD	0.80/0.78	0.78		
SB Enc/Dec	CC	0.72/0.72	0.72	71.80	5.74
	AD	0.71/0.71	0.71		
Ac. fusion	CC	0.75/0.75	0.75	74.70	*
	AD	0.74/0.74	0.74		
BERT (ASR 2)	CC	0.92/0.77	0.84	81.70	4.67
	AD	0.71/0.89	0.79		
BERT (ASR 3)	CC	0.94/0.79	0.86	**84.51**	4.26
	AD	0.74/0.92	0.83		
BERT (ASR 4)	CC	*	*	*	**3.85**
	AD	*	*		
Global fusion	CC	0.94/0.79	0.86	**84.51**	4.62
	AD	0.74/0.92	0.83		

5. Conclusions and future work

In this study, we have analyzed the use of acoustic and linguistic approaches for the automatic detection of AD and MMSE prediction in a low resource scenario proposed by the ADReSS$_o$ challenge organizers. The acoustic approaches consisted of speaker and speech recognition embeddings and prosodic features, whereas the linguistic models were built with BERTs trained on different ASR transcripts. Our findings suggest that acoustic and linguistic approaches contain complementary information for automatic detection and assessment of AD. The x-vector model and encoder-decoder automatic speech recognition embeddings provided the best results among acoustic models, and the BERT fine-tuned with automatic transcriptions from a commercial ASR system yielded the best results for the linguistic approach. Also, the use of the interpolated LM to adapt the ASR to the target domain produced an absolute improvement of 6.4% accuracy and 0.84 in the detection and MMSE prediction tasks, respectively.

In future work, we will evaluate the use of several iterations of LM interpolation to adapt and refine the ASR to the target domain. We will also explore multi-modal approaches in which the classifier uses aligned linguistic and acoustic information in order to extract more precise cues and exploit the bi-modal complementarity.

6. References

[1] K. Ayesu, B. Nguyen, S. Harris, and S. Carlan, "The case for consistent use of medical eponyms by eliminating possessive forms," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, p. 127, 2018.

[2] 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.

[3] K. A. Matthews, W. Xu, A. H. Gaglioti, J. B. Holt, J. B. Croft, D. Mack, and L. C. McGuire, "Racial and ethnic estimates of alzheimer's disease and related dementias in the united states (2015–2060) in adults aged 65 years," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 17–24, 2019.

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

[5] 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.

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

[7] 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.

[8] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *Proc. Interspeech 2020*, pp. 2172–2176, 2020.

[9] M. Martinc and S. Pollak, "Tackling the adress challenge: a multimodal approach to the automated recognition of alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.

[10] R. Pappagari, J. Cho, L. Moro-Velázquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[11] N. Cummins *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[12] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, 2020, pp. 2187–2191.

[13] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multimodal features from pre-trained networks for alzheimer's dementia recognition," *Proc. Interspeech 2020*, pp. 2217–2221, 2020.

[14] T. Searle, Z. Ibrahim, and R. Dobson, "Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech," *Proc. Interspeech 2020*, pp. 2192–2196, 2020.

[15] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[16] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *Proc. Interspeech 2020*, pp. 2167–2171, 2020.

[17] A. Pompili, T. Rolland, and A. Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *Proc. Interspeech 2020*, pp. 2202–2206, 2020.

[18] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *Proc. Interspeech 2020*, pp. 2222–2226, 2020.

[19] E. Edwards *et al.*, "Multiscale system for alzheimer's dementia recognition through spontaneous speech," *Proc. Interspeech 2020*, pp. 2197–2201, 2020.

[20] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity," *Proc. Interspeech 2020*, pp. 2212–2216, 2020.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," *ICASSP 2020*, p. (accepted), 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] M. Ravanelli *et al.*, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[30] J. F. Gemmeke *et al.*, "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.

[31] 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.

[32] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[33] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms." in *ASRU*, 2015, pp. 539–546.

[34] 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.

[35] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4593–4601.

[36] 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.

[37] 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.

[38] M. Johnson *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.

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

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

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

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

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

Abstract

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

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

1. Introduction

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

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

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

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

Our contributions are:

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

⋆equal contribution

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

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

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

2. Data

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

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

3. Methods

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

3.1. Acoustic Features

3.1.1. X-vectors

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

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

3.1.2. Prosody

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

3.1.3. Voice Activity Detection Features

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

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

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

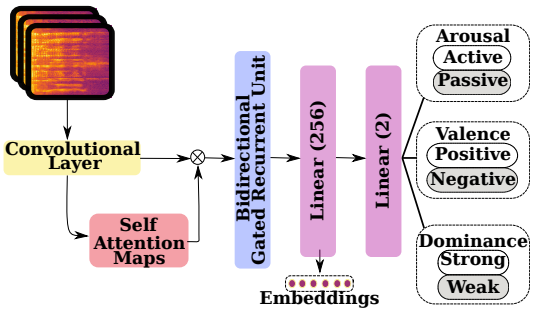


Figure 1: *General architecture of the three pre-trained models based on the PAD emotional model. It consists of four parts: (1) a CNN of 8 filters with a kernel size of (1,3), max pooling (1,2), a batch normalization layer, and a leaky ReLU activation, (2) a self-attention map layer, (3) a Bidirectional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

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

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

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

3.2. Linguistic Features

3.2.1. Word-Embeddings

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

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

3.2.2. Perplexity

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

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

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

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

3.3. Optimization, Classification, and Regression

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

4. Experiments and Results

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

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

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

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

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

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

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

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

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

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

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

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

5. Discussion and Conclusions

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

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

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

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

6. Acknowledgements

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

7. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Alzheimer's Disease Detection from Spontaneous Speech through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models

Yu Qiao[1], *Xuefeng Yin*[1], *Daniel Wiechmann*[2], *Elma Kerz*[1]

[1]RWTH Aachen University, Germany
[2]University of Amsterdam, Netherlands

yu.qiao@rwth-aachen.de, xuefeng.yin@rwth-aachen.de, d.wiechmann@uva.nl, elma.kerz@ifaar.rwth-aachen.de

Abstract

In this paper, we combined linguistic complexity and (dis)fluency features with pretrained language models for the task of Alzheimer's disease detection of the 2021 ADReSSo (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge. An accuracy of 83.1% was achieved on the test set, which amounts to an improvement of 4.23% over the baseline model. Our best-performing model that integrated component models using a stacking ensemble technique performed equally well on cross-validation and test data, indicating that it is robust against overfitting.
Index Terms: Alzheimer's disease, disfluency, pretrained language models, automated Alzheimer's disease detection, linguistic complexity

1. Introduction

Alzheimer's disease (AD) is a gradual and progressive neurodegenerative disease caused by neuronal cell death [1]. The number of people diagnosed with AD is rapidly increasing[1]. The high prevalence of the disease and the high costs associated with traditional approaches to detection make research on automatic detection of AD critical [2]. A growing body of research has demonstrated that quantifiable indicators of cognitive decline associated with AD are detectable in spontaneous speech (see [3] for a recent review). These indicators encompass acoustic features, such as vocalisation features (i.e. speech-silence patterns) [4], paralinguistic features, such as fluency features [5] and speech pause distributions [6], as well as syntactic and lexical features extracted from speech transcripts [7].

This area of research has benefited from recent advances in natural language processing and machine learning, as well as an increasing number of interdisciplinary research collaborations. A prime example of this is the ADReSS(o) (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, aimed at generating systematic evidence for the use of such indicators in automated AD detection systems and towards their clinical implementation. This challenge has made significant contributions to research on AD detection by enabling the research community to test their existing methods, develop novel approaches and to benchmark their AD detection systems on a shared dataset. The ADReSSo Challenge at INTERSPEECH 2021 [8] is geared towards automatic recognition of AD from spontaneous speech and involved three subtasks. Here in this paper, we focus on the AD classification subtask, for which research teams were asked to build a model to predict the label (AD or non-AD) for a short speech session. Participating teams could use the speech signal directly and extract acoustic features or automatically convert the speech to text (ASR) and extract linguistic features from this ASR-generated transcript.

[1]https://www.alz.org/alzheimers-dementia/facts-figures

1.1. Related work

In this section, we provide a concise review of research on automatic AD detection through speech, with particular attention to previous studies conducted as part of the 2020 ADReSS Challenge. The AD classification approaches in this challenge relied on a wide range of acoustic, paralinguistic, and linguistic features or their combination. Classification accuracy scores of the proposed models ranged between 68% and 89.6%. While some approaches either focused on acoustic or linguistic features, the best performing contributions in the 2020 challenge embraced a multi-modal approach combining several types of features (e.g. [9][10][11]). Furthermore, building on earlier work reporting on the effectiveness of the use of word embeddings in AD detection ([12][13]), several approaches successfully employed pretrained language models (e.g. [9][10][11]). Another important issue addressed in several studies concerned how to deal with variance in the predictive performance of pretrained models resulting from fine-tuning for downstream tasks with a small data set. In response to this issue, the authors of the best performing model [9] introduced an ensemble method to increase the robustness of their approach. In response to this issue, the best performing paper of the 2020 challenge [9] introduced an ensemble approach to increase the robustness of their models. Finally, it is important to note that some of the high-performing models in last year's challenge – including the best model described in [9] – used rich manual transcription that included pause and disfluency annotation. Such transcripts were not provided in the 2021 challenge, making it more demanding compared to last year's challenge.

1.2. Modeling approach

The modeling approach presented in this paper builds on key insights reported in the studies reviewed above and extends on these (1) by integrating linguistic indicators of linguistic complexity and sophistication, features of (dis)fluency and transformer-based pretrained language models and (2) by utilizing ensembling methods to combine the information from these feature groups and to reduce the variance in model predictions. Specifically, we perform experiments with classification based on three ensembling techniques: Ensembling by bagging via majority vote, ensembling by bagging using feature fusion, and ensembling by stacking.

http://dx.doi.org/10.21437/Interspeech.2021-1415

2. Data and analysis

2.1. Data

The Alzheimer's Disease Detection dataset provided by the organizers of the ADReSSo Challenge 2021 consists of speech recordings of picture descriptions from the Boston Diagnostic Aphasia Exam produced by 87 individuals with an AD diagnosis and 79 cognitively normal subjects (control group). The recordings were acoustically enhanced (noise reduction through spectral subtraction) and normalised. The data were also balanced with respect to age and gender. The organizers also provided segmentations of the recordings into vocalisation sequences with speaker identifiers. No transcripts were provided.

2.2. Speech Recognition

We used AppTek's Automatic Speech Recognition technology via a cloud API service[2] for automatically transcribing the audio files. The transcripts were converted from XML into raw text formats with full stops being added at the end of each utterance based on the provided segmentations. These files served as the input for the automated text analysis (see Section 2.4).

2.3. (Dis)fluency

To model the speakers' articulatory (in particular (dis)fluency-related) characteristics, we derived several features from the ASR system that fall into four classes. (1) *Silent pauses* - The ASR output contained the start- and end-times as well as confidence scored for each recognized word. Durations of pauses were calculated from forced alignment and binned by duration into short pauses ($< 2sec$) and long pauses ($> 2sec$). In addition, we calculated the total pause duration per sentence (in seconds). (2) *Speed of articulation* - We enriched the output of the ASR with syllable counts from the Carnegie Mellon University Pronouncing Dictionary[3]. Based on this information we assessed the mean syllable duration as well as syllables per minute for each utterance in the speech data. (3) *Filled pauses* - Next to the number and total duration of silent pauses, we derived frequency counts per sentence for two filled pause type, *uh* and *um*, that had been shown to discriminate between AD patients and controls in previous studies [9]. (4) *Pronunciation* - As the known symptoms of AD patients include mispronunciation [14], we calculated average word level confidence scores as a proxy of pronunciation quality, which have been employed for the speech pattern detection in the context of detection of Alzheimer's Disease [15]. All measures were calculated at utterance level. An overview of these measures with descriptive statistics for both groups is presented in Table 1.

2.4. Automated Text Analysis (ATA)

The speech transcripts were automatically analyzed using CoCoGen (short for: Complexity Contour Generator), a computational tool that implements a sliding window technique to calculate within-text distributions of scores for a given language feature (for current applications of the tool in the context of text classification, see [16, 17, 18]). In this paper, we employed a total of 293 features derived from interdisciplinary, integrated approaches to language [19] that fall into four categories: (1) measures of syntactic complexity, (2) measures of lexical richness, (3) register-based n-gram frequency measures, and (4)

[2]https://www.apptek.com/

[3]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Table 1: *Descriptive statistics of (dis)fluency measures*

	AD patients		Control	
(Dis)Fluency measure	M	SD	M	SD
Speed of articulation				
Mean syllable duration	0.28	0.05	0.26	0.03
Syllables per minute	205	45.7	224	35.7
Silent pauses				
Pause time per sentence (in sec)	0.92	0.89	0.63	0.49
N long pauses ($> 2sec$)	1.28	2.22	0.473	0.71
N short pauses ($< 2sec$)	13.2	9.28	15.4	11.7
Filled pauses				
N *uh*	0.29	0.88	0.24	0.54
N *um*	0.07	0.37	0.31	0.74
Pronunciation				
Mean ASR confidence	0.83	0.09	0.86	0.08

information-theoretic measures. In contrast to the standard approach implemented in other software for automated text analysis that relies on aggregate scores representing the average value of a feature in a text, the sliding-window approach employed in CoCoGen tracks the distribution of the feature scores within a text. A sliding window can be conceived of as a window of size ws, which is defined by the number of sentences it contains. The window is moved across a text sentence-by-sentence, computing one value per window for a given indicator. In the present study, the ws was set to 1. The series of measurements generated by CoCoGen captures the progression of language performance within a text for a given indicator and is referred here to as a 'complexity contour' (see Figure 1 for illustration). CoCoGen uses the Stanford CoreNLP suite [20] for performing tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic parsing (Probabilistic Context Free Grammar Parser [21]).

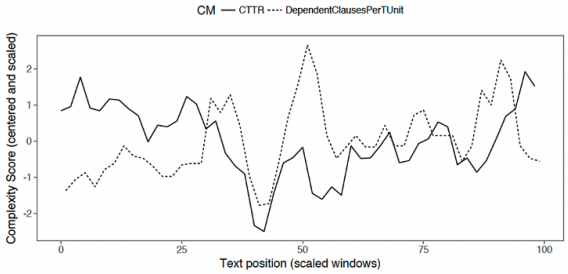


Figure 1: *Schematic representation of 'complexity contours' for two out of 293 complexity measures (CM) investigated: CTTR (Corrected Type Token Ratio) and Dependent Clauses per TU-nit). Centering/scaling was applied here only for purposes of illustration.*

2.5. Pretrained Language Models

Since their inception, transformer-based pretrained language models such as BERT [22] and ERNIE [23] have achieved state-of-the-art performance in various classification tasks.The results of previous research demonstrate that the language characteristics of AD too can be captured by pretrained language models fine-tuned to the task of AD classification (see above). In this paper, pretrained BERT and ERNIE models were fine-tuned for the AD classification task and combined with classifiers trained on complexity and (dis)fluency features (see Section 3). Each of the 161 speakers in the training data is considered as a data point. The input of the model consists of all the text sequences of each speaker obtained by the ASR system, and the output is the class of the corresponding speaker, 0 for Control and 1 for AD.

3. Experimental Setup

In this section we describe the component models used in our approach and how they were combined. To assess the performance of each model, 5-fold cross validation was used.

3.1. CNN Complexity + (Dis)Fluency Models

In order to make optimal use of the complexity and (dis)fluency features, which are sequential in nature, we built convolutional neural network (CNN) models. Originally proposed in computer vision, CNNs have been successfully adapted to various NLP tasks [24] and sentence classification tasks [25][26][27]. The CNN model has the advantage over models that rely on aggregated features, e.g. mean feature values, in that it is capable of capturing patterns in a feature sequence. We followed the approach proposed by [26], but replaced the word embedding with the concatenation of complexity and (dis)fluency features. Due to the small size of the dataset, we set the size of filters to be $2 \times d$, $3 \times d$, $4 \times d$ where d is the input feature dimension. Eight filters were used for each of the three filter types.

3.2. Fine-tuned BERT and ERNIE Models

The Huggingface Transformers library [28] was adopted for fine-tuning pretrained language models. Bert-for-Sequence-Classification was used and initialized with 'bert-base-uncased' and 'nghuyong/ernie-2.0-en' as our pretrained BERT and ERNIE model, respectively. In both cases, the base model was used rather than the large one, as preliminary experiments revealed no reliable differences in terms of classification accuracy between the two models on our dataset. Both models consists of 12 Transformer layers with hidden size 768 and 12 attention heads. The following hyperparameters were used for fine-tuning: the learning rate was set to 2×10^{-5} with 50 warmup steps and l_2 regularization set to 0.1. The maximum sequence length for both models was set to 256. For both models, default tokenizers were used.

3.3. Use of Ensembling Methods

Previous research on predicting AD using pretrained language models has demonstrated that their predictions based on fine-tuning for downstream tasks with a small dataset tend to be brittle and subject to high variance. To reduce this variance, we used an adapted version of the ensembling approach proposed in [9]: Each of the models described above was trained 50 times ($N = 50$). During the prediction phase, each model instance independently generated a prediction. The final classification decision was then determined by hard-voting, i.e. each model contributed its class prediction as a vote and the class that receives the majority of the votes was returned by the ensemble model. Besides using ensemble methods so as to reduce the variance in the prediction of a model, we also employed them to integrate information from different models. To this end, we performed experiments with two types of ensemble based methods, which are referred to here as *ensembling by bagging* and *ensembling by stacking*. Bagging involves fitting several independent models and pooling their predictions in order to obtain a model with a lower variance, while stacking involves combining the models by training a meta-model to output a prediction based on the different models predictions (see below). In each of the combined models, we used the same hyperparameter settings as stated above.

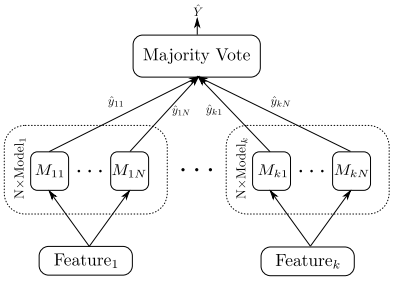


Figure 2: *Structure diagram of Model A. During training, we train each of the k models N times. During inference, jth instance of model i gives prediction $\hat{y}_{ij}$ independently. The final output of the ensembled model $\hat{Y}$ is the label, which the majority of the $k \times N$ model instances agree upon.*

3.3.1. Model A: Ensembling by bagging via majority vote

Ensembling by bagging via majority vote has been shown to be a simple yet effective method to increase the performance of classification models [29][30]. The first classification model (Model A) employed majority voting among 50 CNNs that used complexity and (dis)fluency features and 50 ERNIE models (see Figure 2). That is, as specified above, in this approach, each model was first trained/fine tuned 50 times, meaning that the final classification was based on 100 model instances. The classification in the Model A approach was then determined by counting the votes for each class (AD and controls (CN)) and choosing the more frequent class as the predicted one.

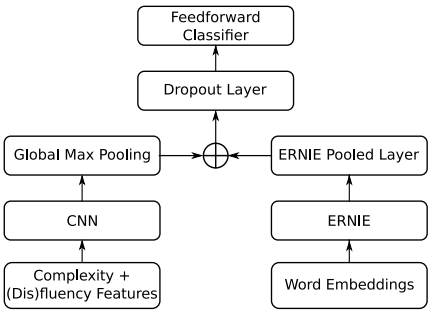


Figure 3: *Structure diagram of Model B.*

3.3.2. Model B: Ensembling by bagging using feature fusion

The second model (Model B) combined a CNN and a ERNIE model, which has previously been shown to perform better than either model alone [31]. Following the approach of [31], we built a model in which complexity and (dis)fluency information was first concatenated at the feature-level and subsequently fed into a CNN (see Figure 3). The hidden vector coming from CNN is then concatenated with the pooled output vector for the [CLS][4] token of Ernie model. The concatenated vector will serve as the input of a feed forward classifier on top of CNN and Ernie. To train this model, we first fine-tune ERNIE model. Then we freeze the parameters of the ERNIE model and jointly train the CNN model and feedforward classifier.

3.3.3. Model C: Ensembling by stacking

The final model, Model C, used in our experiments employed a stacking approach to ensemble all models [32], which has been

[4][CLS], stands for classification, is a special token added in front of every input samples of BERT/ERNIE model to represent sample-level classification [22].

Table 2: *Mean accuracy (with standard deviations), precision, recall and F1 scores over a 5 fold cross-validation*

		Precision		Recall		F1	
Model	Acc	CN	AD	CN	AD	CN	AD
CNN Comp	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
CNN[Comp+DisFl]	0.80 (0.06)	0.79 (0.06)	0.81 (0.08)	0.78 (0.07)	0.83 (0.06)	0.78 (0.05)	0.82 (0.07)
Bert-Base	0.79 (0.06)	0.77 (0.09)	0.84 (0.08)	0.81 (0.11)	0.78 (0.12)	0.78 (0.06)	0.80 (0.07)
Ernie-Base	0.80 (0.04)	0.80 (0.08)	0.81 (0.04)	0.77 (0.07)	0.83 (0.09)	0.78 (0.04)	0.82 (0.05)
Model A: CNN[Comp+DisFl]+[Ernie] (*sep mod, bagging*)	0.76 (0.07)	0.61 (0.13)	0.88 (0.05)	0.79 (0.08)	0.74 (0.08)	0.68 (0.10)	0.80 (0.06)
Model B: CNN[Comp+DisFl]+[Ernie] (*fusion, bagging*)	0.83 (0.06)	0.75 (0.11)	0.89 (0.04)	0.83 (0.09)	0.82 (0.06)	0.78 (0.09)	0.85 (0.04)
Model C: LR[Comp]+LR[DisFl]+[Ernie]+[Bert] (*stacking*)	**0.83** (0.07)	0.82 (0.10)	0.85 (0.09)	0.83 (0.10)	0.84 (0.09)	0.82 (0.08)	0.84 (0.07)

Table 3: *Performance of the three ensemble models on test set*

		Precision		Recall		F1	
Model	Acc	CN	AD	CN	AD	CN	AD
Model A: CNN[Comp+DisFl]+Ernie(*sep mod, bagging*)	0.79	0.77	0.81	0.83	0.74	0.80	0.78
Model B: CNN[Comp+DisFl]+Ernie (*fusion, bagging*)	0.75	0.73	0.77	0.81	0.69	0.76	0.72
Model C: LR[Comp]+LR[DisFl]+Ernie+Bert (*stacking*)	**0.83**	0.82	0.85	0.86	0.80	0.84	0.82

shown to effectively increase the accuracy of the ensembled individual models. Specifically, we employed model stacking to combine two logistic regression models (LR) using complexity and (dis)fluency features respectively, and the two pretrained language models, i.e. BERT and ERNIE. The training procedure consists of two stages (see Figure 4). First, in stage one, each of the four models is trained/fine-tuned independently using 5-fold cross-validation (CV). For each sample in the test fold, we obtain one prediction vector from each of the four models (Models 1 to 4). These predictions vectors are then concatenated and constitute the input data in a subsequent stage (stage 2). The final predictions of Model C are derived from another logistic regression model trained on the concatenated prediction vectors from stage 1. To perform inference on the test set, we take the predictions from all model instances trained in stage 1 and average them by model, which will served as input of stage 2 after concatenation. All hyperparameters for the training/fine-tuning of each of the ensembled models were selected as above.

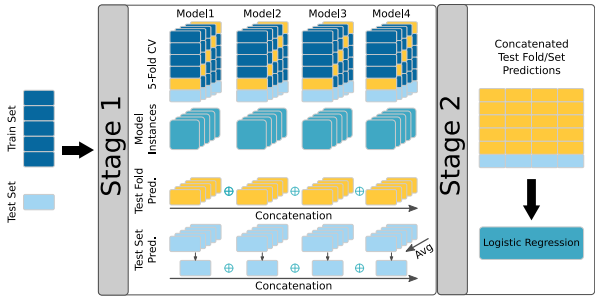


Figure 4: *Schematic representation of ensembling by stacking.*

4. Evaluation

In this section, we present our results on the AD detection task. The evaluation metrics for detection (accuracy, precision, recall, and F1 score) on the cross-validation (CV) set are presented in Table 2. The results on the evaluation set are shown in Table 3. As indicated by boldface numbers, the best performing model in both cross-validation (mean accuracy = 83.16%) and testing (accuracy = 83.10%) was Model C, i.e. the model that combined complexity and (dis)fluency features with both pretrained language models using stacking. Model B, which combined a CNN trained on utterance-level complexity and (dis)fluency features with the best performing fine-tuned pretrained language model (ERNIE) using late fusion and ensembling by bagging, fell close behind reaching 82.7% accuracy in CV. Model A, which combined the same features using majority voting with separate classifiers, performed below the accuracy levels of its component models, reaching 75.69% accuracy in CV. On the test set, the accuracy score of 83.1% of the best performing model, Model C, constitutes an improvement by 4.23% over the baseline model, which was based on fusion of linguistic and acoustic features [8]. Surprisingly, the relative performances of Model A and Model B were reversed on the test set, with Model A matching the performance of the baseline exactly (accuracy = 78.87%) and Model B falling just short of that (accuracy = 74.65%). The considerable discrepancies between the CV and test set classification accuracy for these models suggest that they suffer from overfitting. In contrast, Model C, which employed the stacking technique, performed equally well on CV and test data, indicating that it is robust against overfitting.

5. Discussion and Conclusion

The work presented here combined linguistic complexity and (dis)fluency features with pretrained language models for the task of Alzheimer's disease detection. An accuracy of 83.1% was achieved on the test set, which amounts to an improvement of 4.23% over the baseline model, which was based on fusion of linguistic and acoustic features. Our best performing model combined component models using a stacking ensemble technique. A key finding of this study is that incorporating information on linguistic complexity and (dis)fluency improved the performance of fine-tuned pretrained language models in AD classification by 3%, suggesting that different component models encode complementary information regarding the characteristic language patterns of AD. Another important aspect of our results is that the ensemble model trained on 'complexity contours', i.e. utterance-level measurements of human-interpretable complexity and fluency features, was able to match the performance of both fine-tuned pretrained BERT-like language models: Using 5-fold cross-validation with ensembling of 50 models in each fold, we obtained robust performance scores ($\approx$ 80%) for both types of models. This finding has important implications in light of increasing calls for moving away from black-box models towards white-box (interpretable) models for critical industries such as healthcare, finances and news industry [33, 34].

6. References

[1] M. P. Mattson, "Pathways towards and away from alzheimer's disease," *Nature*, vol. 430, no. 7000, pp. 631–639, 2004.

[2] J. Zeisel, K. Bennett, R. Fleming *et al.*, "World alzheimer report 2020: Design, dignity, dementia: Dementia-related design and the built environment," 2020.

[3] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, no. Preprint, pp. 1–27, 2020.

[4] 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.

[5] E. L. Campbell, R. Y. Mesía, L. Docío-Fernández, and C. García-Mateo, "Paralinguistic and linguistic fluency features for alzheimer's disease detection," *Computer Speech & Language*, vol. 68, p. 101198, 2021.

[6] P. Pastoriza-Dominguez, I. G. Torre, F. Dieguez-Vide, I. Gomez-Ruiz, S. Gelado, J. Bello-Lopez, A. Avila-Rivera, J. Matias-Guiu, V. Pytel, and A. Hernandez-Fernandez, "Speech pause distribution as an early marker for alzheimer's disease," *medRxiv*, pp. 2020–12, 2021.

[7] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

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

[9] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[10] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[11] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[12] 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.

[13] 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.

[14] J. B. Orange, R. B. Lubinski, and D. J. Higginbotham, "Conversational repair by individuals with dementia of the alzheimer's type," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 4, pp. 881–895, 1996.

[15] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," *Proc. Interspeech 2020*, pp. 4961–4965, 2020.

[16] E. Kerz, Y. Qiao, D. Wiechmann, and M. Ströbel, "Becoming linguistically mature: Modeling english and german children's writing development across school grades," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 65–74.

[17] Y. Qiao, D. Wiechmann, and E. Kerz, "A language-based approach to fake news detection through interpretable features and brnn," in *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, 2020, pp. 14–31.

[18] M. Ströbel, E. Kerz, and D. Wiechmann, "The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning," *Language Learning*, vol. 70, no. 3, pp. 732–767, 2020.

[19] M. H. Christiansen and N. Chater, "Towards an integrated science of language," *Nature Human Behaviour*, vol. 1, no. 8, Jul. 2017.

[20] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[21] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

[22] 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.

[23] 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.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[25] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[26] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[27] M. Ma, L. Huang, B. Xiang, and B. Zhou, "Dependency-based convolutional neural networks for sentence embedding," *arXiv preprint arXiv:1507.01839*, 2015.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[29] C. Costello, R. Lin, V. Mruthyunjaya, B. Bolla, and C. Jankowski, "Multi-layer ensembling techniques for multilingual intent classification," 2018.

[30] N. C. Oza and K. Tumer, "Classifier ensembles: Select real-world applications," *Information fusion*, vol. 9, no. 1, pp. 4–20, 2008.

[31] I. Alghanmi, L. Espinosa-Anke, and S. Schockaert, "Combining bert with static word embeddings for categorizing social media," 2020.

[32] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[33] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[34] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154 096–154 113, 2019.

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

Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs

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, UK
[2]Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
{m.rohanian, j.hough, m.purver}@qmul.ac.uk

Abstract

We present two multimodal fusion-based deep learning models that consume ASR transcribed speech and acoustic data simultaneously to classify whether a speaker in a structured diagnostic task has Alzheimer's Disease and to what degree, evaluating the ADReSSo challenge 2021 data. Our best model, a BiLSTM with highway layers using words, word probabilities, disfluency features, pause information, and a variety of acoustic features, achieves an accuracy of 84% and RSME error prediction of 4.26 on MMSE cognitive scores. While predicting cognitive decline is more challenging, our models show improvement using the multimodal approach and word probabilities, disfluency, and pause information over word-only models. We show considerable gains for AD classification using multimodal fusion and gating, which can effectively deal with noisy inputs from acoustic features and ASR hypotheses.

Index Terms: Cognitive Decline Detection, Alzheimer's dementia, disfluency, lexical predictibility

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease that affects memory, language, cognitive skills, and the ability to perform simple everyday tasks.

Throughout the course of AD, patients have been observed suffering a loss of lexical-semantic skills, including suffering anomia, reduced word comprehension, object naming problems, semantic paraphasia, and a reduction in vocabulary and verbal fluency [1, 2]. Speech in patients with AD is mostly characterised by a low speech rate and frequent hesitations at the phonetic and phonological level; however, the syntactic ability is better preserved than lexical-semantic ability in AD patients at the early stages of the disease[3].

The presence of cognitive dysfunction must be confirmed by neuropsychological tests such as the mini-mental state assessment (MMSE) performed in medical clinics before an AD diagnosis can be made. The existence of typical neurological and neuropsychological characteristics and a clinical examination of the patient's history are used to make a diagnosis.

Detecting early diagnostic biomarkers that are non-invasive and cost-effective is of great value for clinical assessments. Several previous studies have investigated AD diagnosis via acoustic, lexical, syntactic, and semantic aspects of speech and language. More interactional aspects of language, like disfluencies, and purely non-verbal features, such as intra- and inter-speaker silence, can be key features of AD conversations. If useful for diagnosis, these features can have many advantages: they are easy to extract and are relatively language, subject, and task agnostic.

In terms of speech features, the number of pauses, pause proportion, phonation time, phonation–to–time ratio, speech rate, articulation rate, and noise–to–harmonic ratio were all found to be related to the severity of Alzheimer's disease [4]. Weiner et al. [5] used a Linear Discriminant Analysis (LDA) classifier with a set of acoustic features including the mean of silent segments, silence durations, and silence-to-speech ratio to differentiate subjects with AD from the control group, achieving an 85.7% AD binary classification. Ambrosini et al. [6] used selected acoustic features (pitch, voice breaks, shimmer, speech rate, syllable duration) to detect mild cognitive impairment from a spontaneous speech task.

Lexical features from spontaneous speech have been shown to be informative in terms of features that assist AD detection. For example, Jarrold et al. [7] merged acoustic features with the frequency occurrence of 14 distinct parts of speech features. Abel et al. [8] modeled patient speech errors (naming and repetition disorders) to aid AD diagnosis.

Modeling multimodal input for AD detection has also been studied. Gosztolya et al. [9] looked at how two SVM models with different sets of acoustic and linguistic features could be combined. Their research demonstrated how audio and lexical features could provide additional knowledge about an individual with AD.

Among other similar tasks within cognitive state prediction like depression, research has been done on integrating temporal information from two or more modalities using multimodal fusion [10]. The different predictive capacities of each modality and their different levels of noise are a major challenge for these models. A gating mechanism is effective in controlling the level of contribution of each modality to the final prediction in a variety of multimodal tasks, including in AD classification and regression [11].

This paper constitutes an entry into the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) challenge 2021 [12], which involves an AD classification and MMSE score regression tasks, in addition to a cognitive decline (disease progression) inference task using only the audio data from formal diagnosis interviews with patients as input. In the first two tasks, participants are required to rate the severity of Alzheimer's disease in various subjects, with the target severity determined by their MMSE scores. In the third task, participants should identify those patients who exhibit cognitive decline within two years.

In this paper, we were particularly interested in the benefit of fusing ASR results (rather than transcripts) with acoustic data

http://dx.doi.org/10.21437/Interspeech.2021-1633

and whether self-repair disfluencies and unfilled pauses in individuals' speech and language model probabilities (a measure of lexical predictability) from automatic speech recognition (ASR) results would help predict the severity of the patient's cognitive impairment.

Inspired by [11], to detect AD, we used audio and text features to model the sessions in a Bidirectional Long-Short Term Memory (BiLSTM) neural network. We used the Bidirectional Encoder Representations from Transformers (BERT) model to classify AD from speech recognition results in a separate experiment. Our findings suggest that AD can be identified using pure sequential modelling of the speech recognition results from the interview sessions with limited details of the structure of the description tasks. Disfluency markers, unfilled pauses, and language model probabilities were also found to have predictive power for detecting Alzheimer's disease.

2. Data and features

Two distinct datasets were used for the ADReSSo Challenge:

1. a set of speech recordings of picture descriptions produced by both patients with an AD diagnosis and subjects without AD (controls), who were asked to describe the Boston Diagnostic Aphasia Exam's Cookie Theft picture [12].
2. a set of speech recordings of Alzheimer's patients performing a category (semantic) fluency task [13] at their baseline visit for prediction of cognitive decline over two years.

Dataset 1 for AD classification and severity detection includes 237 audio recordings, and the state of the subjects is assessed based on 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 a severe impairment.

Dataset 2 for the disease prognostics task (prediction of cognitive decline) was created from a longitudinal cohort study involving AD patients. The period for assessing disease progression spanned the baseline and the year-2 data collection visits of the patients to the clinic. The task involves classifying patients into 'decline' or 'no-decline' categories, given speech collected at baseline as part of a verbal fluency test.

Various features were extracted automatically from both datasets for the 3 ADReSSo tasks as described below.

2.1. Acoustic 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 [14]. 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 the training set.

2.2. Linguistic Features

For automatically transcribing the audio files, we used the free trial version of IBM's Watson Speech-To-Text service.[1] The service offers ASR on the audio data which has considerable noise and may be affected by non-standard North American dialect of the patients - the average Word Error Rate (WER) on 10 transcripts we randomly selected from the training data is 32.8%. The Watson service, crucially for our task, does not filter out hesitation markers or disfluencies [15]. It also outputs word timings that we use as features in our system.

For our models which did not use BERT, 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.

2.3. Disfluencies

Disfluencies are usually seen as indicative of communication problems caused by production or self-monitoring issues [17]. 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 [18, 19].

We automatically annotate self-repairs and edit terms using [20]'s multi-task learning model in a left-to-right, word-by-word manner to predict disfluency tags. Here each word is either tagged as one of $\{$*repair onset*, *edit term*, *fluent word*$\}$ by the disfluency detector- we concatenate the disfluency tags with the word vectors to create the input for the text-based LSTM classifier described below.

2.4. Unfilled Pauses

Durations of pauses were calculated from the word timings provided by the ASR hypotheses, using the latency between the end of the previous word to the beginning of the patient's current word as the pause length, with the value for the first word being 0. We further categorized pauses into either *short pause* (SP) and *long pause* (LP). An SP is a silence that occurs inside a single speaker turn, which in the range $[0.5, 1.5)$ seconds; an LP is a longer pause within a single speaker turn defined as a speech pause of 1.5 seconds or greater. Pauses in the interviewer's speech were excluded.

2.5. Language Model Probabilities

People with speech disorders or cognitive impairment express themselves in different ways when compared to control groups [21]. Language model probabilities, which can be interpreted to estimate the predictability of a sequence of words, can be used to assess a participant's language structure, including vocabulary and syntactic constructions. The present work uses a Multi-task Learning (MTL) LSTM language model [20] based on the Switchboard corpus [22], a sizable multispeaker corpus of conversational speech and text. The language model uses

[1] https://www.ibm.com/uk-en/cloud/watson-speech-to-text

standard Switchboard training data for disfluency detection (all conversation numbers starting sw2*,sw3 * in the Penn Treebank III release: 100k utterances, 650k words) and is trained in combination with other tasks, including disfluency detection as described in [20]. This corpus can be viewed as an approximation of control, non-AD disorder spoken dialogue. The model is then tested on the ASR transcript of each session, and the probability of each word is calculated. Finally, we concatenate the probability of the current word given the history $p(w_t|w_0...w_{t-1})$ with the word vectors to create the lexical input for our model.

3. Proposed Approach

We experiment with different deep-learning architectures for predicting AD in both classification and regression and for cognitive decline prediction:

1 unimodal LSTM models utilising using either acoustic or lexical features.

2 multimodal LSTM model using lexical and acoustic information, including disfluency and pause tagging.

3 unimodal BERT based classifier using lexical features.

4 multimodal BERT model with gating using lexical and acoustic information.

3.1. Sequence modeling

Our approach is to model the speech of individuals as a sequence to predict whether they have AD or not, and if so, to what degree, using either LSTMs or BERT models.

LSTM The potential of neural networks lies in the power to derive representations of features by non-linear input data transformations, providing greater capacity than traditional models. As we were interested in modelling the 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, then used late fusion to combine their probabilities.

BERT Pre-trained BERT models are fine-turned for the AD classification task. Each of the training instances is considered a data point. The input to the model consists of a sequence of words from the transcript for every speaker. Following [23] we used Bert-for-Sequence-Classification2 for fine-tuning. The standard default tokenizer was used, and two special tokens, [CLS] and [SEP], were added to the beginning and the end of each input. Specifically for regression, the last layer is the shape (hidden size, 1), and we use MSE loss instead of cross-entropy.

3.2. Multimodal Model with Gating

Since learned representation for the text can be undermined by corresponding audio representation and ASR results can be unreliable, we need to minimise the effects of noise and overlaps during multimodal fusion. For audio and textual input for the BiLSTM models, we use two branches of the LSTM, one for each of the modalities, with their outputs combined into final *feed-forward highway layers* [24], with gating units that learn by weighing text and audio inputs at each time step to regulate information flow through the network.

The concatenated output is passed through N highway layers (where the best value *N* was determined from optimizing on held-out 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.

For the BERT-based multimodal models with gating, the output from the BERT-based textual classifier is combined with the acoustic data into the final feed-forward highway layers.

Table 1: *Result of the AD classification and regression experiments with our models against baseline models on test set*

Models	Features	Accuracy	RMSE
Baseline ([12])			
LDA	Linguistic	0.76	-
DT	Linguistic	0.75	6.24
SVM	Acoustic+Linguistic	0.79	-
SVR	Acoustic+Linguistic	-	5.29
GP	Linguistic	-	5.95
Our Models			
LSTM	Words	0.76	-
LSTM	Words+Words Probabilities	0.77	4.75
LSTM	Words+Disf+Pause	0.81	4.43
BERT	Words	0.80	4.49
BERT w/ Gating	Words+Acoustic	-	4.38
LSTM w/ Gating	Words+Acoustic+Disf+Pse+WP	**0.84**	**4.26**

4. Experiments

4.1. Implementation and Metrics

We set up our model to learn the most helpful information from modalities for predicting AD. All experiments are carried out without being conditioned on the identity of the speaker.

For the LSTM models, the sizes of layers and the learning rates are calculated by grid search on validation test. 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 1 with four 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. We use a block of 3 stacked highway layers. The LSTM models were trained using ADAM [25] with a learning rate of 0.0001. We used Binary Cross-Entropy to model binary outcomes for the loss function and Mean Square Error (MSE) to model regression outcomes.

For the BERT models, following [23] we use the "bert-large-uncased" model, with the hyperparameters: learning rate = 2e-5, batch size = 4, epochs = 8, max input length of 256.

For binary classification of AD and non-AD, we report binary accuracy scores. For the MMSE prediction task, we report the Root Mean Square Error (RMSE) for the prediction error score. For the cognitive decline task, we report the mean of F1 classification scores.

The code used in the experiments is publicly available in an online repository.[2]

4.2. Baseline Models

We compare the performance of our models to the ADReSSo Challenge baselines [12] with an ensemble of audio and linguistic features provided with the dataset. The best baselines we include here include decision trees (DT), linear discriminant analysis (LDA), support vector machines (SVM), support vector regression (SVR), and Gaussian process regression (GP).

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

Table 2: *Result of the AD classification and regression experiments with our models in cross validation*

Models	Features	Accuracy	RMSE
LSTM	Acoustic	0.68	6.03
LSTM	Words	0.74	5.31
LSTM	Words+Words Probabilities	0.78	4.78
LSTM	Words+Disfluency+Pause	0.78	5.02
BERT	Words	0.80	4.94
BERT	Words+Acoustic	0.78	**4.72**
LSTM w/ Gating	Words+Acoustic	0.79	4.88
LSTM w/ Gating	Words+Acoustic+Disf+Pse+WP	**0.81**	4.75

5. Results

AD classification and regression tasks In Table 1, we present our proposed models' performance against that of the baselines models on AD classification and regression tasks on the provided test set and in Table 2 in a cross-validation setting. For AD detection, our proposed LSTM model with gating and additional features (disfluency, unfilled pause, and language model probabilities) achieves an accuracy of **0.84** and RMSE of **4.26**, outperforming all the baselines. Overall, the results support our hypothesis that a model with a gating structure can more effectively reduce individual modalities' errors and noise, including that from errorful ASR results. Furthermore, our proposed LSTM model with gating and additional features (disfluency, unfilled pauses, and language model probabilities) outperforms the BERT fine-tuned models in unimodal and multimodal situations (ACC 0.84 vs. 0.80; RMSE 4.26 vs. 4.49 and 4.38). It should also be noted that the BERT model is very large in comparison to the LSTM models. BERT has approximately 21 times the number of parameters as our second largest model (105 million vs. 4.9 million). Therefore, compared to the BERT model, our LSTM models need fewer resources for development.

Effect of disfluency and unfilled pause features We found that disfluencies and unfilled pauses help as features in AD detection. Adding disfluency and pause features to the lexical features lead to improvement on the test set (ACC 0.81 vs. 0.76) and in CV (ACC 0.78 vs. 0.74; RMSE 5.02 vs. 5.31). Our LSTM model with disfluencies and unfilled pauses outperforms the BERT model in both class-action and regression tasks on the test set (ACC 0.81 vs. 0.80; RMSE 4.43 vs. 4.49).

Effect of language model probabilities Language model probabilities (as an indicator of grammatical integrity) are useful as features in the diagnosis of AD. Adding language model probabilities to the lexical features improves the test set (ACC 0.77 vs. 0.76) and in CV (ACC 0.78 vs. 0.74; RMSE 4.78 vs. 5.31).

Effect of multimodality On both the test set and in CV, the multimodal LSTM with gating model outperforms the single modality AD detection models in classification and regression tasks. In CV, integrating textual and audio modalities with gating improves performance over single modality models (ACC 0.79 vs. 0.74; RMSE 4.88 vs. 5.31). Even though each LSTM branch has different steps and timestep inputs in multimodal models, adding audio features improves performance. The multimodal model with BERT outperforms the single modality BERT in the regression task on both the test set and in CV (RMSE 4.72 and 4.38 vs. 4.94 and 4.49). However, integrating BERT and audio model with gating decreases performance over BERT for classification in CV (ACC 0.78 vs. 0.80). Text features are more informative than audio features as using text modality only predicts AD better than using unimodal audio modality sequentially in CV (ACC 0.74 vs. 0.68; RMSE 5.31 vs. 6.03).

Table 3: *Result of Task3: cognitive decline progression results (mean of F1Score) for leave-one-subject-out CV and Test set*

Models	Features	CV	Test
Baseline ([12])			
LDA	Linguistic	0.55	0.54
DT	Linguistic	**0.76**	**0.67**
SVM	Linguistic	0.45	0.40
Our Models			
LSTM	Words	0.59	0.55
LSTM	Words+Disfluency+Pause	0.55	0.50
BERT	Words	0.63	0.54
LSTM w/ Gating	Words+Acoustic+Disf+Pse+WP	0.66	0.62

Cognitive decline (disease progression) inference task In Table 3, we present our results for disease progression task. As can be seen, our models do not reach the best baseline of the Decision-Tree based classifier. However, as with AD classification, the multimodal LSTM with Gating model outperforms all other competitors and is close to the DT classifier in performance on the test data (ACC **0.62** vs. 0.67). Overall, this task seems to have a considerably greater variation in performance across baseline classifiers and feature sets than the other two tasks. The lower performance of the LSTM model using words with disfluency and pause information model compared to using words alone (ACC 0.55 vs. 0.59) suggests these extra features are not as useful compared to the lexical information alone. This suggests the ASR quality is more critical, and the comparison of the IBM Watson system used here against the results obtained by the Google Cloud-based Speech Recogniser used by [12] would be a future step to take.

6. Conclusions

We have presented two multimodal fusion-based deep learning models which consume ASR transcribed speech and acoustic data simultaneously to classify whether a speaker in a structured diagnostic task has Alzheimer's Disease and to what degree. Our best model, a BiLSTM with highway layers using words, word probabilities, disfluency features, pause information, and a variety of acoustic features, achieves an accuracy of 84%. While predicting cognitive decline is more challenging, our models show improvements using the multimodal approach and word probabilities, disfluency, and pause information over word-only models. In addition, we show there are considerable gains for AD classification using multimodal fusion and gating, which can effectively deal with noisy inputs from acoustic features and ASR hypotheses.

7. 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.

8. References

[1] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurological Sciences*, vol. 26, no. 4, pp. 243–254, 2005.

[2] 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.

[3] G. Kavé and Y. Levy, "Morphology in picture descriptions provided by persons with Alzheimer's disease," *Journal of Speech, Language, and Hearing Research*, 2003.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] M. Rohanian, J. Hough, M. Purver *et al.*, "Detecting depression with word-level multimodal fusion," *Proc. Interspeech 2019*, pp. 1443–1447, 2019.

[11] M. Rohanian, J. Hough, and M. Purver, "Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 2187–2191. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2721

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

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

[14] 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.

[15] T. Baumann, C. Kennington, J. Hough, and D. Schlangen, "Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there," in *Dialogues with social robots*. Springer, 2017, pp. 421–432.

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

[18] 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.

[19] S. Nasreen, M. Rohanian, M. Purver, and J. Hough, "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features," *Frontiers in Computer Science*, vol. 3, p. 49, 2021.

[20] M. Rohanian and J. Hough, "Re-framing incremental deep language models for dialogue processing with multi-task learning," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 497–507. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.43

[21] K. Gabani, M. Sherman, T. Solorio, Y. Liu, L. Bedore, and E. Pena, "A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children," in *Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics*, 2009, pp. 46–55.

[22] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[23] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

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

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

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

Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech

Zafi Sherhan Syed[1], *Muhammad Shehram Shah Syed*[2], *Margaret Lech*[2], *Elena Pirogova*[2]
[1]Mehran University, Pakistan
[2]RMIT University, Australia
zafisherhan.shah@faculty.muet.edu.pk
muhammad.shehram.shah.syed|margaret.lech|elena.pirogova@rmit.edu.au

Abstract

This paper addresses the Interspeech Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge 2021. The objective of our study is to propose the approach to a three task automated screening that will aid in distinguishing between healthy individuals and subjects with dementia. The first task is to differentiate between speech recordings from individuals with dementia. The second task requires participants to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech. The third task requires participants to leverage speech recordings to identify whether individuals have suffered from cognitive decline. Here, we propose a system based on functionals of deep textual embeddings with special preprocessing steps integrating the effect of silence segments. We report that the developed system outperforms the challenge baseline for all three tasks. For Task 1, we achieve an accuracy of 84.51% compared to the baseline of 77.46%, for Task 2, we achieve a root-mean-square-error (RMSE) of 4.35 compared to the baseline of 5.28, and for Task 3, we achieve an average-f1score of 73.80% compared to the baseline of 66.67%. These results are a testament of the effectiveness of our proposed system.

Index Terms: alzheimer's dementia, computational paralinguistics, social signal processing

1. Introduction

Alzheimer's disease is a chronic neurodegenerative disorder that detrimentally impacts cognitive and physical well-being of a person. According to the World Health Organization (WHO) [1], dementia currently affects more than 50 million people worldwide, with millions of new patients being diagnosed every year.

The ever-growing use of artificial intelligence (AI) in healthcare-related applications has facilitated development of innovative and advanced medical diagnostic approaches to various types of disorders [2, 3, 4, 5]. The main advantage of such techniques is that they can be successfully employed for objective diagnosis of disorders. The limited human interference assists in reducing human errors and bias. Considerable effort has been directed towards the development of diagnostic methods which can be used to identify individuals with Alzheimer's dementia [6].

The Interspeech Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge 2021 [7] aims to provide a common platform to researchers to not only propose methods for automated screening of Alzheimer's dementia but also encourages researchers to compete and evaluate their work against their peers. The challenge this year may be considered as an extension of last year's ADReSS 2020 challenge [8] with an important difference. Whereas last year, the dataset contained manually transcribed and CLAN [9] annotated transcripts, this year's challenge expects participants to work with automatically generated speech transcripts.

In last year's ADReSS challenge, our developed system performed very well, achieving an accuracy of 85.42% compared to the challenge baseline of 77.00% and an RMSE score of 4.30 compared to the baseline of 5.20 for the test partition. In essence, first we observed that features derived from textual modality offer much better performance than those from the audio modality. Secondly, we had demonstrated the prowess of a simple but very effective method for representing speech transcripts of subjects as feature vectors. To that end, we had first computed deep textual embeddings (DTE) from transformer based models and applied functionals of descriptive statistics to pool their values into a feature vector.

This paper describes our proposed system [1] for tackling the ADReSSo challenge 2021. Here, we demonstrate the efficacy of a system based on functionals of deep textual embeddings with special preprocessing steps integrating the effect of silence segments. We also show the potential benefits of class-imbalance aware multi-model fusion.

2. Dataset

The ADReSSo challenge consists of two distinct datasets. The first dataset is called 'Diagnosis' and is used for Task 1 and Task 2 of the challenge. In Task 1, the objective is to differentiate between speech recordings from individuals with dementia amongst a set of recordings from healthy individuals. In Task 2, the objective is to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech. The second dataset 'Progression' is used for Task 3 of the challenge. Here, the objective is to identify, based on characteristics of their speech, whether subjects have suffered from a cognitive decline over years. For further details regarding the dataset, we refer the reader to the ADReSSo challenge baseline paper [7].

3. Methodology

A block diagram representation of our proposed system for the ADReSSo challenge is provided in Figure 1 where it can be seen that the system starts with automated speech recognition (ASR) for speech-to-text conversion. Next, we experiment with various preprocessing methods (detailed in Section 3.2). This is followed by a process of generating feature representations for transcripts using DTE as well as handcrafted features. We used handcrafted features to compare the performance of deep textual embeddings against domain-knowledge features. The

[1]Our previous work [10] provides the necessary context to our current work

http://dx.doi.org/10.21437/Interspeech.2021-1572

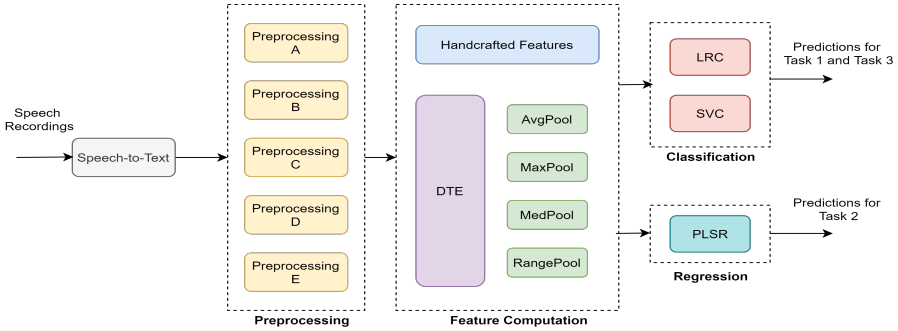


Figure 1: *Block diagram for our proposed automated screening system*

final step is to train classification and regression methods on the training partition and make predictions for the test partition.

3.1. Generating Speech Transcripts

An important aspect of the ADReSSo challenge 2021 is for participants to work without manually annotated transcripts. The dataset contains time-stamps that could be used to identify speech segments from the subject and the interlocutor. However, through preliminary experiments we discovered that these time-stamps are not always aligned with speech recordings. Although it was possible to use speaker diarization to identify segments of speech that belong only to a particular subject, we decided to use ASR to generate transcripts for the entire speech recording without diarization. We assume that (a) speech from the subject will dominate the recording and the contribution from the interlocutor will be relatively small, and (b) the speech and language from the interlocutor will also reflect the cognitive state of the subject. For example, the interlocutor will use simplified language to communicate with an individual that suffers from language impairments due to dementia.

For ASR, we experimented with wav2vec2 model from Huggingface toolkit [11], Silero [12] toolkit, and Microsoft Stream [2]. Initial results showed that the latter provided the most promising results in terms of word error rate. We plan to compare the performance of various automated ASR approaches for the task at hand in due course.

3.2. Preprocessing

We have experimented with five types of text preprocessing methods to investigate whether a particular method leads to an improvement in classification or regression tasks of the challenge.

- Preprocessing-A: Here, we resolved word contractions, removed punctuation and extra whitespaces from transcripts, and converted all text in lower-case. Thus, the entire transcript was set up as a single sentence.
- Preprocessing-B: We removed extra whitespaces and represented text in lower-case only.
- Preprocessing-C: Here, we decided to add special keywords into the speech transcript depending on the duration of a silence segment between two successive utterances. We were inspired to explore this method given the success reported by [13] using a similar technique. However, unlike Yuan et al., we test this method on pretrained embeddings only. Therefore in Preprocessing-C, we started with the setup of Preprocessing-A and if the silence duration was determined to be between 2 and 4 seconds, we added the text 'uhm'. If the silence duration was between 4 and 6 seconds, we added the word 'uhm uhm'. Finally, if the silence duration exceeded 6 seconds, we added the text 'long silence'.
- Preprocessing-D: In this method, we followed the procedure of adding special keywords as in Preprocessing-C after removing extra whitespaces and also converted text in lower-case.
- Preprocessing-E: Here, the preprocessing was performed as in Preprocessing-D except that the replacement text for silence segments was a period symbol ('. ') instead of 'uhm'. For example, if the silence duration was between 2 and 4 seconds, we added a '. ' as text. If the silence duration was between 4 and 6 seconds, we added '. . ' , and in case of the silence duration exceeding 6 seconds, we added the text 'long silence'.

3.3. Feature Computation

Once speech transcripts were generated and preprocessed, the next step was to compute textual features. As mentioned previously, we computed both, handcrafted features and deep textual embeddings. For handcrafted features, we computed a set of textual features inspired by the work of Fraser et al. [14]. These features can be categorized as (a) syntactic, (b) readability, and (c) lexical diversity. As the name suggests, syntactic features provide information about the syntax of written communication. In our work, we used SpaCy toolkit [3] to compute normalized histogram counts of parts-of-speech and dependency tags for the transcript of each subject. The second type of handcrafted features used in this work measured the readability of subjects' transcripts. We suggest that there are differences in the readability of speech transcripts between healthy subjects and those with dementia. Hence, we used the Readability toolkit [4] to compute eight features that quantify the readability of speech transcripts. Finally, given that Alzheimer's dementia disorder affects memory, we posit that subjects with dementia will use a repetitive and less diverse vocabulary compared to healthy subjects. To quantify the diversity of their vocabulary, we computed ten features based on text-to-token-ratio using the Lexical Diversity toolkit [5].

[2]https://www.microsoft.com/en-us/microsoft-365/microsoft-stream

[3]https://spacy.io

[4]https://pypi.org/project/readability

[5]https://pypi.org/project/lexical-diversity

Table 1: *Summary of Results of Task 1 for the Training partition*

Preproc.-ID	Feature Name	Acc.	avg-f1score	Sens.	Spec.
E	facebook_bart_base__AvgPool	84.34	84.34	79.31	89.87
B	distilroberta_base__MaxPool	83.73	83.72	82.76	84.81
B	facebook_bart_base__Median	82.53	82.52	80.46	84.81
B	bert_base_multilingual_uncased__AvgPool	81.93	81.93	78.16	86.08
D	bert_base_multilingual_uncased__AvgPool	81.93	81.92	80.46	83.54
B	bert_large_uncased__MaxPool	81.93	81.89	82.76	81.01
A	facebook_bart_base__Median	81.93	81.86	83.91	79.75

Table 2: *Summary of Results of Task 2 for the Training partition*

Preproc.-ID	Feature Name	RMSE	MAE	Pearson's r
B	bert_base_multilingual_uncased__MaxPool	4.64	3.65	0.75
D	facebook_bart_base__AvgPool	4.86	3.75	0.72
A	distilbert_base_uncased__RangePool	4.87	3.84	0.72
B	distilbert_base_uncased__RangePool	4.90	3.92	0.71
D	bert_base_multilingual_uncased__RangePool	4.91	3.86	0.71

Table 3: *Summary of Results of Task 3 for the Training partition*

Preproc.-ID	Feature Name	Acc.	avg-f1score	Sens.	Spec.
C	facebook_bart_base__AvgPool	83.56	73.49	53.33	91.38
C	bert_large_uncased__MaxPool	83.56	70.08	40.00	94.83
D	facebook_bart_base__AvgPool	83.56	70.08	40.00	94.83
C	bert_base_multilingual_uncased__RangePool	84.93	69.41	33.33	98.28
C	facebook_bart_base__Median	80.82	69.07	46.67	89.66

In addition to the above, we investigated the efficacy of deep textual embeddings, such as Bidirectional Encoder Representations from Transformers (BERT) [15] and its derivatives. These models use multi-headed self-attention [16] based encoder and decoder layers which enable them to learn sophisticated latent representations from text [15, 17, 18]. Jawahar et al. [18] have shown that transformer-based models can capture structural and linguistic properties of the English language as classical tree-like structures.

We surmise that such models can represent linguistic characteristics of speech and as such be useful for differentiating between speech transcripts of healthy subjects and those with dementia. To this end, we experiment with embeddings generated using nine pre-trained transformer-based models which include: $bert\_base\_uncased$, $bert\_large\_uncased$, $distilbert\_base\_uncased$ [19], $roberta\_base$, $roberta\_large$ [20], $distilroberta\_base$,$bert\_base\_multilingual\_uncased$, $allenai\_biomed\_roberta\_base$ [21], and $facebook\_bart\_base$ [22]. We used the Huggingface library [11] in order to compute these embeddings.

It should be mentioned here that these embeddings are computed for each input token (for example, a word), and not the entire transcript as a single entity. Therefore, to generate a single feature vector for the entire transcript, we used functionals of descriptive statistics for pooling. For example, average pooling (AvgPool), maximum value pooling (MaxPool), percentile-based range pooling (RangePool), and median value pooling (MedianPool) are used in this work. The resultant feature vector is passed down to the machine learning pipeline as shown in Figure 1.

3.4. Classification and Regression

As mentioned earlier, the ADReSSo challenge 2021 consists of two classification tasks (Task 1 and Task 3) and one regression task (Task 2). We used a logistic regression classifier (LRC) and support vector machine classifier (SVC) with a linear kernel for Tasks 1 and 3, whereas for Task 2, we used partial least squares regressor (PLSR). We have previously used success using these tools [4, 10, 23]. The regularization parameter 'C' for LRC and SVC was optimized using leave-one-subject-out cross-validation (LOSO-CV) over a logarithmically spaced grid between 10^{-7} and 10^{3} whilst using the avg-f1score as the metric of classification performance. It should be mentioned that although the official performance metric for Task 1 is accuracy, we decided to use avg-f1score to optimize the regularization parameter. This was done due to the class imbalance in the training partition for the dataset provided for Task 1. Meanwhile, the 'number of components' hyper-parameter for PLSR was optimized using LOSO-CV over a grid between 2 and 40 to minimize the RMSE score. We used the scikit-learn toolkit [24] for training models for classification and regression.

4. Experiments and Results

4.1. Predictions for the training partition

In Table 1, we report the results of the top-5 models for the training partition of Task 1. Here, one can note that the best performing model uses Preprocessing-E with DTE features com-

puted from the $BART\_basemodel$. It achieves an accuracy of 84.34% with a specificity of 89.87% but a relatively poor sensitivity of 79.31%. Meanwhile, the second-placed model achieved an accuracy of 83.73% on the training partition but does not use any special processing to integrate silence information into speech transcripts. It is interesting to note that this model offers an improved sensitivity at the cost of decreased specificity when compared with the best-performing model.

In Table 2, we summarize the results for training partition of Task 2. The objective here is to predict the MMSE score assigned to subjects from the Diagnosis dataset. We report that the best performing model achieves an RMSE of 4.64 compared to the baseline of 6.42, which is a significant reduction. It is also interesting to note that there does not appear to be any advantage of integrating silence information into the speech transcript for Task 2, however, this requires further investigation.

Finally, the results for Task 3 have been summarized in Table 3 where our best performing model achieves an avg-f1score of 73.49% compared to the challenge baseline of 66.67%. Most importantly, all of the top-5 models use special preprocessing steps.

4.2. Predictions for the test partition

For Task 1, we first used confidence-based fusion to combine predictions from the top-3 performing models. This step achieved an accuracy of 81.69% for the test partition. Next, we used label-fusion of top-5 models and this increased the classification accuracy to 83.10%. Finally, we attempted label-fusion of five models selected on the basis of their specificity and sensitivity scores for the training partition. Two of these models are $facebook\_bart\_base\_\_AvgPool$ and $bert\_base\_multilingual\_uncased\_\_AvgPool$ which had provided high specificity for the training partition, whereas the remaining three, i.e. $distilroberta\_base\_\_MaxPool$, $bert\_large\_uncased\_\_MaxPool$, and $facebook\_bart\_base\_\_Median$ had provided high sensitivity. We assume that the fusion of models with high specificity only (as is the case with top-5 models) will bias predictions towards a particular class and therefore lead to poorer results overall. The resultant predictions for this fusion method achieve the best results for Task 1 where we achieved a classification accuracy of 84.51% compared to the challenge baseline of 77.46%, which is a significant improvement.

In Table 5, we summarize the results for the test partition of the dataset for Task 2. In our first attempt, we used predictions for the test partition as generated by the best-performing model for the training partition. This yielded an RMSE of 4.93 for the test partition. Our second and third attempts used averaging and median-based fusion with predictions from the top-3 models for the training partition. With these methods, we achieved an RMSE score of 4.71 and 4.54, respectively. Finally, we attempted averaging and median-based fusion for fourth and fifth attempts and achieved RMSE scores of 4.45 and 4.35, respectively. It is interesting to note that all of our attempts at predicting MMSE score achieved better performance than the challenge baseline of 5.28.

Finally, in Table 6, we summarize the results for Task 3, where the objective was to identify whether the subject has suffered from a cognitive decline over two years. Here, our best result for the test partition is an avg-f1score of 73.80%. This was achieved via predictions generated by the model which yielded the best performance for the training partition. We also experimented with label-fusion of predictions for the top-3 models for the training partition, however, this led to a decrease in avg-f1score. It should be mentioned here that we did not experiment with class-imbalance aware fusion, as used for Task 1, although in hindsight it may have been a better option.

Table 4: *Summary of Results of Task 1 for the Test partition*

Predictions (source)	**Acc.**	**avg-f1score**
baseline	77.46	–
Conf. Fusion of Top-3 models	81.69	81.64
Label Fusion of Top-5 models	83.10	82.94
Label Fusion selected models	*84.51*	84.45

Table 5: *Summary of Results of Task 2 for the Test partition*

Predictions (source)	**RMSE**
baseline	5.28
Single best model	4.93
Average-value fusion from Top-3 models	4.71
Median-value fusion from Top-3 models	4.54
Average-value fusion from Top-5 models	4.45
Median-value fusion from Top-5 models	*4.35*

Table 6: *Summary of Results of Task 3 for the Test partition*

Predictions (source)	**Acc.**	**avg-f1score**
baseline	–	66.67
Single best model	78.13	*73.80*
Label Fusion of Top-3	68.75	54.29

5. Conclusions

Alzheimer's dementia is a disease that greatly reduces the quality of life of those who suffer from it. Early detection of this disorder may assist to enhance the quality of their day-to-day lives. The purpose of the ADReSSo challenge was to develop an automated screening tool for dementia recognition. In this paper, we proposed a system based on functionals of deep textual embeddings and benchmarked its performance against the official baselines set by the ADReSSo challenge organizers. We also demonstrated that one can enrich speech transcripts with silence segments in speech recordings to yield improved performance. Overall, our proposed solution for the ADReSSo challenge offers a significant improvement for all three tasks as follows. For Task 1, we achieved an accuracy of 84.51% compared to the baseline of 77.46%, for Task 2, we achieved an RMSE of 4.35 compared to the baseline of 5.28, and for Task 3, we achieved an avg-f1score of 73.80% compared to the baseline of 66.67%. These results are a testament to the efficacy of our proposed system.

6. Acknowledgements

We would like to thank Abbas Syed for useful, constructive discussions, and assistance with running some of the experiments.

7. References

[1] World Health Organisation, "Dementia: Key Facts," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[3] Z. S. Shah, K. Sidorov, and D. Marshall, "Psychomotor Cues for Depression Screening," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2017, pp. 1–5.

[4] 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.

[5] Z. S. Syed, S. A. Memon, and A. L. Memon, "Deep Acoustic Embeddings for Identifying Parkinsonian Speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 726–734, 2020.

[6] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review." *Journal of Alzheimer's disease : JAD*, vol. 78, no. 4, pp. 1547–1574, 2020.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," *medRxiv*, pp. 1–5, 2021.

[8] ——, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH*, 2020, pp. 2172–2176.

[9] B. Macwhinney, *The CHILDES project: Tools for analyzing talk*, 3rd ed. Psychology Press, 1992.

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

[11] 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.

[12] S. Team, "Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks," 2021. [Online]. Available: https://github.com/snakers4/silero-models

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

[14] 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.

[15] 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.

[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. 1–11.

[17] Y. Goldberg, "Assessing BERT's syntactic abilities," *arXiv:1901.05287*, pp. 1–4, 2019.

[18] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3651–3657.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, pp. 1–5, 2019.

[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:1907.11692*, pp. 1–13, 2019.

[21] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *arXiv preprint:2004.10964*, pp. 1–19, 2020.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "{BART}: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[23] M. S. S. Syed, E. Pirogova, and M. Lech, "Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach," *Electronics*, vol. 10, no. 11, pp. 1–13, 2021.

[24] 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.

Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition

Conference Paper · August 2021

DOI: 10.21437/Interspeech.2021-1288

6 authors, including:

Esaú Villatoro-Tello
Metropolitan Autonomous University
74 PUBLICATIONS **332** CITATIONS
SEE PROFILE

S Pavankumar Dubagunta
Idiap Research Institute
10 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Julian Fritsch
Idiap Research Institute
7 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Gabriela Ramirez-de-la-Rosa
Metropolitan Autonomous University
40 PUBLICATIONS **112** CITATIONS
SEE PROFILE

Late Fusion of the Available Lexicon and Raw Waveform-based Acoustic Modeling for Depression and Dementia Recognition

Esaú Villatoro-Tello[1,2], *S. Pavankumar Dubagunta*[2,3], *Julian Fritsch*[2,3], *Gabriela Ramírez-de-la-Rosa*[1], *Petr Motlicek*[2], *Mathew Magimai.-Doss*[2]

[1] Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico City, Mexico
[2] Idiap Research Institute, Martigny, Switzerland
[3] École polytechnique fédérale de Lausanne (EPFL), Switzerland

evillatoro@cua.uam.mx, pavankumar.dubagunta@idiap.ch, julian.fritsch@idiap.ch, gramirez@cua.uam.mx, petr.motlicek@idiap.ch, mathew.magimaidoss@idiap.ch

Abstract

Mental disorders, e.g. depression and dementia, are categorized as priority conditions according to the World Health Organization (WHO). When diagnosing, psychologists employ structured questionnaires/interviews, and different cognitive tests. Although accurate, there is an increasing necessity of developing digital mental health support technologies to alleviate the burden faced by professionals. In this paper, we propose a multi-modal approach for modeling the communication process employed by patients being part of a clinical interview or a cognitive test. The language-based modality, inspired by the Lexical Availability (LA) theory from psycho-linguistics, identifies the most *accessible* vocabulary of the interviewed subject and use it as features in a classification process. The acoustic-based modality is processed by a Convolutional Neural Network (CNN) trained on signals of speech that predominantly contained voice source characteristics. In the end, a late fusion technique, based on majority voting, assigns the final classification. Results show the complementarity of both modalities, reaching an overall Macro-F1 of 84% and 90% for Depression and Alzheimer's dementia respectively.

Index Terms: Depression Detection, Alzheimer's Disease, Mental Lexicon, Raw Speech, Multi-modal Approach.

1. Introduction

Mental disorders represent a major public health concern, with considerable associated socio-economic costs, and are recognized as a major cause of disability affecting a great number of people. According to the World Health Organization (WHO), depression and dementia are among the main types of mental disorders and are categorized as priority conditions [1, 2]. Although the severity of suffering a mental illness is well known by psychologists, there is an acknowledged necessity for digital solutions for addressing the burden of mental health diagnosis and treatment. It is recognized that won't be possible to treat people by professionals alone, and even if possible, some people might require to use alternative modalities to receive mental health support [3]. Such situation has become more evident with the current COVID-19 pandemic. Interested readers are referred to [4, 5, 6] to know efforts towards this direction.

Accordingly, the research community has been interested in making first steps towards computer-supported detection of mental disorders during face-to-face interviews/tests [7, 8, 9]. The underlying hypothesis of most of previous work relies on the notion of the language as a powerful indicator about our personality, social, or emotional status, and mental health [10, 11]. In dementia, for instance, previous research indicates that assessing the language production represents a useful strategy in detecting early markers of dementia [12]. Thus, designed tests for evaluating the language production in elderly patients such as word association tasks, description of objects in pictures, elicitation exercises, etc., aim at measuring the expository speech, oral expression, as well as comprehension. Similarly, for depression, previous research suggest that using excessive self-focused language, and negative emotions represent important markers for screening depressed users [10, 13, 14] and, recent studies have documented how depressed users suffer some kind of impairment in their speech motor control [15], such as prosodic abnormalities, articulatory and phonetic errors.

Although multi-modal approaches have been explored before [7, 16, 17], the key novelty of our work is to leverage the psycholinguistics theory for approximating the *mental lexicon*[1] of analyzed subjects for processing the language-based modality. The acoustic-based modality aims at modeling the patients' speech in an end-to-end fashion from raw waveform-based CNNs. In conjunction, both modalities allow modeling the language production process employed by subjects with a mental disease during a clinical interview/test. We performed experiments in two well-known clinical datasets, using individual modalities, and in a multi-modal fashion, where a voter makes the final decision through a majority voting mechanism.

2. Methodology

The proposed language-based modality aims at modeling the vocabulary production of subjects suffering from a mental disorder through the Lexical Availability (LA) theory [19]. The LA test is associated with the category fluency tests and the free word association tasks, which taps directly into the semantic information of the *mental lexicon* [20]. Hence, our main hypothesis establishes that it could be possible to approximate the *available lexicon* for a group of people suffering from a mental disease. Contrary to the traditional LA elicitation test, we aim to demonstrate that it is possible to approximate the available lexicon by analyzing subjects' responses in a semi-structured communication process (e.g, a clinical interview/test). To the best of our knowledge, this is the first time the LA theory is adapted to: *i)* obtain the available lexicon from utterances produced during a clinical test and use the extracted features in a traditional

[1] The *mental lexicon* of a community reveals the type, size, and richness of their vocabulary as well as provides evidence of the community member's understanding of a particular culture, or the structure of their context and the existing regularities present [18].

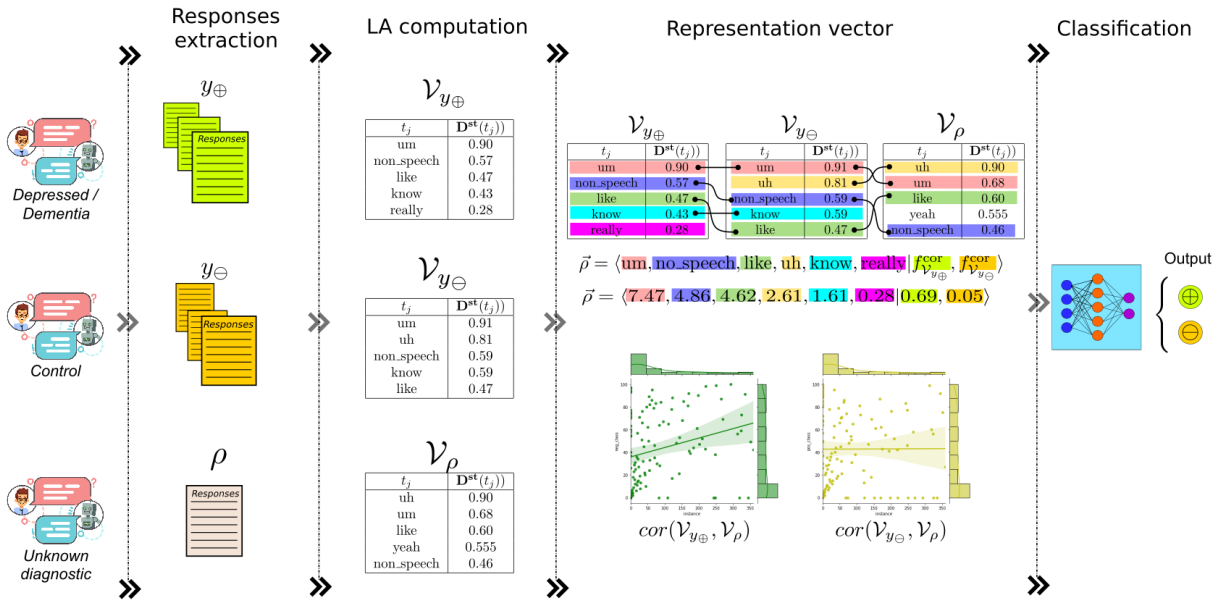


Figure 1: *General overview of the proposed language-based modality.*

classification pipeline; *ii)* fuse its predictions in a multi-modal fashion with a raw waveform-based acoustic approach.

Figure 1 shows the main components of our LA method. First, we identify the available lexicon from each population (i.e., subjects with a mental disorder and control subjects) and then use it to generate a non-sparse text representation to train a classification model to distinguish between *mentally ill* (*D*) and *control* (*C*) subjects. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_h, y_h)\}$ be a training set of h-pairs of documents[2] d_i and class labels $y_i \in \mathcal{Y} = \{y_\oplus, y_\ominus\}$. The first step consists of obtaining the available lexicon ($\mathcal{V}$) for each category, i.e., $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$ for the documents belonging to *D* and *C* categories respectively. The resultant available lexicon for each category y_i is a list of n-pairs of the form $\mathcal{V}_{y_i} = \{(t_1, \mathbf{D^{st}}(t_1)), \ldots, (t_n, \mathbf{D^{st}}(t_n))\}$, where each term t_j is accompanied by its lexical availability score $\mathbf{D^{st}}(t_j)$. Details on how to compute the availability score are depicted in §2.1.

Then, for generating the representation of subject ρ, we define two sets of features: the *availability degree* (f^{avail}), and the *correlation degree* attributes (f^{cor}). Thus, we first compute the available lexicon of subject ρ, referred as $\mathcal{V}_\rho$, and we calculate its *availability* features (f^{avail}) by means of a fusion strategy among the top k terms from $\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus}$, and $\mathcal{V}_\rho$ (see §2.2). For obtaining the *correlation* features (see §2.3) we compare the data distributions between ρ and the two classes ($y_\oplus$ and $y_\ominus$), resulting in a representation vector with the following form:

$$\vec{\boldsymbol{\rho}} = \langle f^{\text{avail}}_{t_1}, \ldots, f^{\text{avail}}_{t_j}, \ldots, f^{\text{avail}}_{t_k} | f^{\text{cor}}_{\mathcal{V}_{y_\oplus}}, f^{\text{cor}}_{\mathcal{V}_{y_\ominus}} \rangle \quad (1)$$

Once we have this representation, we can follow the traditional machine learning pipeline for training a classifier.

2.1. Lexical availability computation

Traditionally, the LA test produces a single word list, i.e., the available lexicon (with its corresponding availability scores), for each analyzed community. To compute the availability scores of this available lexicon, we have to analyze the responses of each individual in that population (see Fig. 1, columns 1-3); to that end, we use the formulation proposed by [21], defined as follows:

$$\mathbf{D^{st}}_{w,k,m}(t_j) = \sum_{i=1}^{n} w^{\left(\frac{i-1}{k-1}\right)^m} \times \frac{f_{ji}}{I} \quad (2)$$

where t_j represents the lexical term for which we want to know its availability score; i is the position indicator where t_j is mentioned in the considered individual responses; n is the maximum position reached by term t_j in all the considered responses; I serves as a normalization factor and is defined as $I = max\_freq$, which depicts the highest frequency found in the vocabulary of the population being analyzed; f_{ji} is the number of participants who produced term t_j at position i in their respective responses; k indicates the position value where the score will be equal to w; w is the desired weight (normally a value close to 0) for position k, and m is a parameter that modulates the weight decay across terms in the final mental lexicon.

Eq. 2 represents a standardized LA metric that allows direct comparisons among studies independently from the size of the produced vocabulary lists of different communities [21]. Accordingly, the $\mathbf{D^{st}}$ equation will assign higher scores (close to 1) to the most available words produced by the analyzed subjects. Conversely, it assigns progressively lower scores to less accessible words until reaching value w in position k, at a weight decay intensity defined by the parameter m. Intuitively, the smaller the value of m, the faster the weight decay across words in consecutive positions. For all our experiments, we defined $w = 0.0001$ and $m = 0.8$.

2.2. Availability features

We defined the *availability* features (f^{avail}) as the single (most representative) LA score for each term $t_j \in (\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus})$. Thus, to obtain the $f^{\text{avail}}_{t_j}$ score of term t_j we apply the CombMNZ [22] data-fusion strategy. Data-fusion strategies aim at integrating many possible answers (scores) for an object into a single

[2] We'll refer as documents to the transcribed text obtained from the subjects' utterances.

best representative score. Therefore, to compute the representative score of t_j we first obtain the available lexicon $\mathcal{V}_\rho$ of the instance ρ applying Eq. 2. Then, for obtaining the $f_{t_j}^{\text{avail}}$ we fuse the scores of word t_j from the list $\mathcal{V}_\rho$ with the available lexicons $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$. For this process, we do as follows:

$$f_{t_j}^{\text{avail}} = \text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\}) \quad (3)$$

where t_j is the word for which we want a fused score, k indicates the maximum position where t_j will be searched in the input lists, and the $\mathcal{V}$'s are the set of lists to be considered for the fusion process. Notice that k has the same interpretation of that in Eq. 2; intuitively, it indicates the number of words (features) to be considered for building the representation vector.

Thus, assuming $N = \text{len}(\{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\})$, D^c as the score of t_j in list c, and $|D^c > 0|$ as the number of non-zero scores given to t_j by any list c, the final score for each unique term t_j is computed as follows:

$$\text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\}) = \sum_c^N D^c \times |D^c > 0| \quad (4)$$

Broadly speaking, the $f_{t_j}^{\text{avail}}$ of term t_j represent a weight value indicating to what category it adjust the best.

2.3. Correlation degree features

The *correlation degree* features aim at measuring the relationship between the two sets of paired words, particularly we compute $cor(\mathcal{V}_{y_\ominus}, \mathcal{V}_\rho)$, and $cor(\mathcal{V}_{y_\oplus}, \mathcal{V}_\rho)$. The correlation ($cor$) value will be an indicator of the association between the available lexicon form subject ρ and the corresponding $\mathcal{V}_{y_\ominus}$ and $\mathcal{V}_{y_\oplus}$ categories. For the experiments performed in this paper, every $f_{t_j}^{\text{cor}}$ feature is formed by two values, the Spearman's correlation coefficient and its corresponding p-value.

2.4. Acoustic based method

The acoustic based method directly models raw waveforms to predict the class-conditional probabilities using a CNN-based architecture. As described in [23], the architecture consists of four 1-D convolutional layers, followed a hidden layer and an output-layer. In order to guide the learning procedure, depending on the task, different approaches were previously proposed: We distinguish between sub-segmental and segmental filtering (see [23] Table 1); raw waveforms can be filtered to extract voice-source related characteristics to guide the learning procedure. Specifically, for the depression detection task, the primary method (denoted as 1stAcoustic) uses zero frequency filtering to get a signal that characterizes the glottal excitation. The secondary method (denoted as 2ndAcoustic) consists of modeling speech at a frame level using linear prediction and subtracting it from the original speech to get the linear prediction residual, which contains voice source related characteristics, while both use an input length of 250ms. However, for Alzheimer's detection, both systems use 4 second length inputs of zero frequency filtered signals, where the primary method (denoted as 1stAcoustic) applies a sub-segmental filtering stage, the secondary method (denoted as 2ndAcoustic) a segmental filtering stage.

2.5. Late fusion

Once both the language-based and acoustic-based modalities are trained independently, the late fusion approach consists of a voter that takes as inputs the predictions made by the language-based and acoustic-based approaches. The final decision is made by means of a majority voting mechanism, where if tied, the output will be always labeled as *C* (i.e., control).

Mod.	Approach	DAIC-WOZ Class.	DAIC-WOZ F1-score *O*	*D*	*C*	ADReSS Class.	ADReSS F1-score *O*	*D*	*C*
Textual	BoW	MLP	0.65	0.48	0.83	SVC	0.84	0.83	0.86
	LIWC	MLP	0.53	0.34	0.72	LR	0.70	0.70	0.70
	BERT	SVC	0.70	0.53	0.86	MLP	0.73	0.74	0.72
	LA-A$_{100}$	PER	0.58	0.40	0.77	SVC	0.77	0.74	0.79
	LA-A$_{500}$	MLP	**0.71**	**0.58**	0.84	LR	0.84	0.83	0.86
	LA-A$_{1000}$	MLP	0.71	0.56	**0.87**	LR	0.84	0.83	0.86
	LA-AC$_{100}$	PER	0.57	0.41	0.73	MLP	0.77	0.75	0.80
	LA-AC$_{500}$	MLP	0.68	0.53	0.83	LR	0.86	0.85	0.87
	LA-AC$_{1000}$	MLP	0.66	0.45	0.86	LR	**0.87**	**0.86**	**0.88**
Acoustic	1stAcoustic	-	**0.58**	**0.41**	**0.76**	-	**0.76**	**0.69**	**0.90**
	2ndAcoustic	-	0.52	0.32	0.71	-	0.76	0.71	0.88

Table 1: *Performance under a 10-CFV strategy on train sets.*

3. Experimental Setup

For the experiments, we use the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset [24] and the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [9]. The DAIC-WOZ dataset contain semi-structured clinical interviews, performed by an (human controlled) animated virtual interviewer, designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic disorder. This dataset was used during the AVEC 2016 challenge [7], and contains audio-visual interviews of 189 participants: 107 for training, 35 for development, and 47 for test. The ADReSS data, introduced for the Interspeech 2020 ADReSS challenge [9], consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [25]. It contains speech and transcripts information from 156 participants: 108 for training, and 48 for test. In the DAIC-WOZ dataset approximately $\approx$30% of the subjects are labeled as depressed (*D*), while the ADReSS data is perfectly balanced. It is worth mentioning that the labeling of each datset was done by expert mental healthcare providers, interested reader is referred to [24, 9].

We evaluate the performance of three well-known text-based methods. First, a traditional Bag-of-Words (BoW) using the top 1000 most frequent words under a Term Frequency Inverse Document Frequency *tf-idf* weighting scheme. Secondly, we use the Linguistic Inquiry and Word Count (LIWC) [26] categories for representing the documents. LIWC psychological categories capture the semantic content of the language produced [27], e.g., allow to detect positive vs. negative emotions, words referencing family/friends/society, pronouns which can capture inclusive language vs. exclusive language, and words referencing how the person is feeling.

As third baseline, we evaluate the impact of recent transformer-based models [28] as a language representation strategy. For our experiments we test an English pre-trained BERT model. As known, the [CLS] token acts an "aggregate representation" of the input tokens, and is considered as a sentence representation for many classification tasks [29]. Accordingly, for generating the representation of each document, we split the document into smaller chunks (max length of 512 tokens), obtain the [CLS] encoding of each chunk, and we apply

Mod.	Approach	DAIC-WOZ Class.	DAIC-WOZ F1-score *O*	DAIC-WOZ F1-score *D*	DAIC-WOZ F1-score *C*	ADReSS Class.	ADReSS F1-score *O*	ADReSS F1-score *D*	ADReSS F1-score *C*
Textual	BoW	MLP	0.53	0.32	0.75	LR	0.85	0.84	0.86
	LIWC	MLP	0.49	0.29	0.69	SVC	0.62	0.57	0.67
	BERT	MLP	0.51	0.30	0.72	SVC	0.81	0.80	0.82
	LA-A_{100}	DT	0.63	0.54	0.73	SVC	0.73	0.70	0.76
	LA-A_{500}	MLP	0.54	0.36	0.71	MLP	0.85	0.86	0.85
	LA-A_{1000}	DT	0.58	0.40	0.76	LR	0.85	0.85	0.86
	LA-AC_{100}	SVC	**0.70**	**0.64**	0.76	MLP	0.75	0.71	0.79
	LA-AC_{500}	MLP	0.51	0.25	**0.79**	LR	**0.87**	**0.88**	**0.86**
	LA-AC_{1000}	PER	0.60	0.48	0.71	LR	0.81	0.82	0.80
Acoustic	1stAcoustic	-	**0.69**	**0.65**	**0.73**	-	**0.79**	**0.82**	**0.75**
	2ndAcoustic	-	0.55	0.53	0.57	-	0.68	0.72	0.65

Table 2: *Obtained performance over the dev and test partitions for DAIC-WOZ and ADReSS datasets respectively.*

Dataset	Fused approaches	F1-score *O*	F1-score *D*	F1-score *C*
DAIC-WOZ	[LA-AC_{100}, LA-A_{100}, 1stAcoustic, 2ndAcoustic]†	**0.84**	**0.80**	**0.89**
	[LA-AC_{100}, LA-A_{100}, BoW, 1stAcoustic, 2ndAcoustic]†	0.82	0.77	0.86
	[LA-AC_{100}, LA-A_{100}, BERT, 1stAcoustic, 2ndAcoustic]	0.79	0.74	0.84
	Al Hanai,T., *et al.* (2018) [16]	*0.77*	-	-
ADReSS	[LA-AC_{500}, LA-A_{500}, 1stAcoustic, 2ndAcoustic]	**0.90**	**0.90**	**0.89**
	[LA-AC_{500}, LA-A_{500}, BoW, 1stAcoustic, 2ndAcoustic]	0.85	0.87	0.84
	[LA-AC_{100}, LA-A_{100}, BERT, 1stAcoustic, 2ndAcoustic]	**0.90**	**0.90**	**0.89**
	Mahajan, P. & Baths, V., (2021) [17]	-	*0.70*	*0.75*

Table 3: *Obtained performance of the late fusion approach. The reported performance in [7] for depression was F1=0.58, while for ADReSS, in [9] the best reached score was F1=0.75.*

a *mean* pooling to obtain the final representation.

Except for the BERT setup, we applied the following normalization steps; all the common contractions, e.g., *we'll*, *can't*, etc., are converted to its formal writing, i.e., *we will*, *can not*, etc. All disfluencies are preserved, non-speech phenomena are labeled as `<non-speech>`, punctuation marks are removed, and number occurrences are labeled as `<number>`, and, all the text is lower cased.

4. Results and discussion

As previous research [7, 9, 23, 30, 31], performance is reported in terms of the F score ($F1$) for both control (*C*) and depression/dementia (*D*) classes, and the Macro-F for the overall problem *(O)*. We acknowledge the limitations regarding the small size of the corpora, however, this is a common shortcoming of all studies that use clinical datasets. Thus, to achieve stable and robust results, we applied two validation strategies: i) the average performance over a stratified 10 cross-fold-validation using *train* partition (10-CFV), and, ii) the performance over the *dev* partition for the DAIC-WOZ[3] dataset and on the *test* partition for the ADReSS dataset.

For the proposed Lexical Availability method, we performed a series of experiments using: *i)* only the *availability degree* features (LA-A), and *ii)* the combination of availability and correlation (LA-AC) as in Eq. 1. Table 1 summarizes our results for the experiments using a 10-CFV strategy; Table 2 shows the performance of the experiments performed on the *dev* and *test* partitions, and Table 3 shows the results of the fused predictions. Given our space restrictions, we only report results from the best learning algorithm (Class. column).[4] For the experiments using the LA-A/LA-AC methods, the number in the sub-index indicates the value of the k parameter.

Clearly, from Tables 1 and 2 we conclude that our LA method outperforms all the proposed textual-based baselines, including very recent transformer-based models (i.e., BERT). Also, observe that adding the *correlation* features helps improving the classification, best performance is obtained under the LA-AC configuration for both tasks (see Table 2) with $k = 100$ for DAIC-WOZ, and $k = 500$ for ADReSS. This variation in the value of k is related to the size of the respective datasets. For instance, the DAIC-WOZ corpus, contrary to the ADReSS dataset, contains more samples of the communicative process (i.e., several utterances from interviewed subject) with a smaller variability of lexical units (i.e., small vocabulary), hence paying attention to a reduced set terms is enough for the LA method.

For the multi-modal experiments (Table 3) we took the best configurations based on the performance on the *dev/test* sets (Table 2). We compare our results against two recent multi-modal approaches. For depression, we considered the work of [16], which evaluates the performance of a multi-modal LTSM recurrent network. For dementia, [17] combines the outputs of CNN-LSTM model and a Speech-GRU cell for making the predictions. As can be observed, our late fusion strategy, outperforms very recent approaches by an important margin.[5]

5. Conclusions

We addressed the problem of detecting mental disorders from clinical tests. Inspired by the LA theory, our method approximates the *mental lexicon* through the identification of the *available lexicon* for mentally ill and control subjects, and use it in a classification process to detect depression/dementia. Additionally, based on previous studies that demonstrated the suitability of raw waveform CNNs, we designed a multi-modal approach, where a voter makes the final decision using a majority vote mechanism. A thorough evaluation in two well known clinical datasets (DAIC-WOZ and ADReSS), shows that the LA method fused with the raw waveform-based CNN is able to outperform, by a large margin, very recent deep NN techniques.

6. Acknowledgements

Esaú Villatoro-Tello, was supported partially by Idiap, SNI-CONACyT, and UAM-Cuajimalpa Mexico. S Pavankumar Dubagunta was supported by Innosuisse under the project Conversation Member Match (CMM). Julian Fritsch was funded through the EU's Horizon H2020 MSCA-ITN-ETN project TAPAS grant agreement No. 766287. Gabriela Ramírez-de-la-Rosa would like to thank UAM-Cuajimalpa for their support.

[3] DAIC-WOZ *test* partition is not publicly available.

[4] Classifiers parameters: Logistic Regresor (**LR** - `solver=lbfgs`), Multilayer Perceptron (**MLP** - `activation=relu, alpha=1e-5, solver=lbfgs, max_iter=300`), Support Vector Machines (**SVC** - `kernel=linear`), Decision Trees (**DT** - `criterion=entropy`, and Perceptron (**PER** - `max_iter=50, tol=1e-3`). All classifiers were set with `random_state=42`.

[5] Symbol † indicates statistical significant results (based on the Wilcoxon signed-rank test with a 90% confidence) in comparison to the non-fused results.

7. References

[1] World Health Organization, *Depression and Other Common Mental Disorders Global Health Estimates*. World Health Organization, 2017. [Online]. Available: https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

[2] ——, *Towards a dementia plan: a WHO guide*. World Health Organization, 2018. [Online]. Available: https://www.who.int/mental_health/neurology/dementia/policy_guidance/en/

[3] T. Wykes, J. Lipshitz, and S. M. Schueller, "Towards the design of ethical standards related to digital mental health and all its applications," *Current Treatment Options in Psychiatry*, vol. 6, no. 3, pp. 232–242, 2019.

[4] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e19, 2017.

[5] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, 2018.

[6] C. Welch, A. Lahnala, V. Perez-Rosas, S. Shen, S. Seraj, L. An, K. Resnicow, J. Pennebaker, and R. Mihalcea, "Expressive interviewing: A conversational system for coping with COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020.

[7] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*. ACM, 2016, pp. 3–10.

[8] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. Association for Computing Machinery, 2019, p. 3–12.

[9] 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 INTERSPEECH 2020*, Shanghai, China, 2020.

[10] A. M. Tackman, D. A. Sbarra, A. L. Carey, M. B. Donnellan, A. B. Horn, N. S. Holtzman, T. S. Edwards, J. W. Pennebaker, and M. R. Mehl, "Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis." *Journal of personality and social psychology*, vol. 116, no. 5, p. 817, 2019.

[11] E. Villatoro-Tello, S. Parida, S. Kumar, P. Motlicek, and Q. Zhan, "Idiap & UAM participation at GermEval 2020: Classification and regression of cognitive and motivational style from text," in *Proceedings of the GermEval 2020 Workshop in conjunction with the 5th SwissText & 16th KONVENS Conference*, 2020, pp. 11–16.

[12] 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.

[13] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[14] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes, "Detecting depression in social media using fine-grained emotions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1481–1486.

[15] 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.

[16] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[17] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 20, 2021.

[18] N. Hernández-Muñoz, C. Izura, and C. Tomé, "Cognitive factors of lexical availability in a second language," in *Lexical availability in English and Spanish as a second language*. Springer, 2014, pp. 169–186.

[19] M. Šifrar Kalan, "Lexical availability and l2 vocabulary acquisition," *Journal of Foreign Language Teaching and Applied Linguistics*, vol. 2, no. 2, 2015.

[20] S. De Deyne, S. Verheyen, and G. Storms, "Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations," in *Towards a theoretical framework for analyzing complex linguistic networks*. Springer, 2016, pp. 47–79.

[21] F. J. Callealta Barroso and D. J. Gallego Gallego, "Medidas de disponibilidad léxica: comparabilidad y normalización (measures of lexical availability: comparability and standardization)," *Boletín de filología*, vol. 51, no. 1, pp. 39–92, 2016.

[22] E. A. Fox and J. A. Shaw, "Combination of multiple searches," *NIST special publication SP*, vol. 243, 1994.

[23] S. P. Dubagunta, B. Vlasenko, and M. M. Doss, "Learning voice source related information for depression detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6525–6529.

[24] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[25] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[26] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[27] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[28] 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.

[29] 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[30] A. Rinaldi, J. Fox Tree, and S. Chaturvedi, "Predicting depression in screening interviews from latent categorization of interview prompts," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7–18.

[31] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

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

Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data

Ning Wang[1], *Yupeng Cao*[1], *Shuai Hao*[1], *Zongru Shao*[2 3*], *K.P. Subbalakshmi*[1]

[1]Stevens Institute of Technology, NJ, USA
[2]Center for Advances Systems Understanding, Görlitz, Germany
[3]Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

nwang7@stevens.edu, ycao33@stevens.edu, shao8@stevens.edu, drssth@gmail.com, ksubbala@stevens.edu

Abstract

In this work, we propose a modular multi-modal architecture to automatically detect Alzheimer's disease using the dataset provided in the ADReSSo challenge. Both acoustic and text-based features are used in this architecture. Since the dataset provides only audio samples of controls and patients, we use Google cloud-based speech-to-text API to automatically transcribe the audio files to extract text-based features. Several kinds of audio features are extracted using standard packages. The proposed approach consists of 4 networks: C-attention-acoustic network (for acoustic features only), C-Attention-FT network (for linguistic features only), C-Attention-Embedding network (for language embeddings and acoustic embeddings), and a unified network (uses all of those features). The architecture combines attention networks and a convolutional neural network (C-Attention network) in order to process these features. Experimental results show that the C-Attention-Unified network with Linguistic features and X-Vector embeddings achieves the best accuracy of 80.28% and F1 score of 0.825 on the test dataset.
Index Terms: Alzheimer's disease, Multi-Modal Approach, CNN-Attention network, Acoustic feature, Linguistic feature

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that is the most common form of dementia and continual cognitive impairments [1]. The number of cases is increasing rapidly every year so that AD has become a non-negligible social public health problem. Therefore, early diagnosis of AD is an essential task and has attracted much attention in recent years.

The ADReSSo challenge at INTERSPEECH 2021 defines a shared task through which different approaches to target the automatic detection of AD [2] can be proposed. The ADReSSo challenge provides only audio data of patients extracted from the Pitt Corpus [3].

Our approach uses both the audio features directly extracted from the audio files and linguistic and other language-based features extracted from the transcribed version of the same audio file. Literature suggests that speech impairment is a common and significant sign of AD even at the early stage of dementia [4, 5]. Therefore, some speech characteristics, such as speech vagueness and abnormal pauses, can function as an important bio-marker. These features in patients' speech can provide useful information about the cognitive status and other aspects related to the level of brain health [6]. Further, studies have shown that several lexical or syntactic features and increases in conversational fillers or non-specific nouns are also indicators of AD [7, 8]. Consequently, Natural Language Processing (NLP) methods can be applied to extract linguistic features from text data [9] and used in the detection of AD [10, 11]. Existing AD classification methods can be divided into three categories depending on the types of features used: leveraging raw audio data or acoustic features using linguistic features derived from text or a combination of acoustic features and linguistic features to detect AD. We have taken the third approach here. The main contributions of this work are as follows: 1) a CNN-attention Network (C-Attention Network) for automated detection of Alzheimer's disease. 2) a method to integrate features extracted from both text and audio.

2. Related Work

Automated detection of Alzheimer's disease has a long history of research. In early automated AD detection work, researchers attempt to quantify the impairments by using computational methods [12]. They first construct or extract different features from the different data sources and then apply traditional machine learning methods to detect Alzheimer's disease. These features can be divided into two categories: linguistic features and acoustic features. Linguistic features including part-of-speech (POS) tag frequencies, measures of lexical diversity were extracted and a linear discriminant analysis or other classifiers were used to identify AD patients [12, 13]. Acoustic features such as mel-frequency cepstral coefficients (MFCC) and low-level descriptors (LDD) were used in [14]. Then, the combination of both acoustic and linguistic features based machine learning approaches were proposed to automatically detect AD [15, 16]. These studies have shown that the methods of combining different types of features have better detection accuracy compared to using features separately.

In recent related research work, the INTERSPEECH 2020 ADReSS challenge provides a baseline paper, which summarized many useful acoustic features [17], including emboase [18], ComParE [19], eGeMAOS [20] and MRGG [21] and followed it with machine learning methods such as Linear Discriminant Analysis (LDA), Decision Tree (DT), Support Vector Machine (SVM) and Random Forests (RF) to detect AD. In ADReSS 2020 challenge, the work [22] utilized two acoustic features, IS10-Paralinguistics feature set from ComParE and Bag-of-Acoustic-Words (BoAW), to achieve a good classification accuracy [22]. Cummins et.al proposed an end-to-end convolutional neural network to directly classify AD [23]. Pan et.al considered the problem of audio data quality and they applied ASR techniques to identify high-quality speech segments

* work done while at Stevens Institute of Technology

http://dx.doi.org/10.21437/Interspeech.2021-2024

for more robust feature extraction to improve detection performance [24]. Some researchers obtain latent features from language embeddings and used the attention mechanism to achieve better performance on text data [6, 9, 25]. A multi-modal approach that fused acoustic and linguistic features was proposed in [26]. In that work, the author used dual-LSTM architecture, one for audio feature and another for text feature. A gating mechanism was used to fuse the two for the final classifier.

3. Proposed Approach

In this section, we introduce the acoustic and linguistic feature sets we use and propose a modular multi-modal architecture to classify AD from non-AD controls.

3.1. Acoustic and Linguistic Feature Sets

3.1.1. Acoustic Features

We used open source audio processing toolkits, OpenSMILE [18] and Kaldi [27], to obtain four different acoustic features from the raw audio file, which are Emobase [18], IS10 [19], VGGish [28] and X-Vector [29]. Specifically, Emobase and IS10 are frame-level acoustic features. VGGish and X-Vector are acoustic embeddings. Frame-level features are directly extracted from audio files and these features capture the frequency characteristics and other statistical information. Different from frame-level features, embedding features are not directly derived from the audio data. The embedding features are from the embedding model, where the embedding model will generate a vector to represent the characteristics in audio data. The embedding model is a deep neural network and pre-trained on large audio datasets. We used these pre-trained embedding models to extract features. Here are the specific descriptions for different feature extraction processes:

Emobase: The Emobase feature set has abundant audio features which include mel-frequency cepstral coefficients (MFCC) information, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP), and intensity features.

IS10: The IS10 feature set includes many frame-level features: 16 types of LLDs, PCM loudness, eight log Mel frequency band (0-7), eight line spectral pairs (LSP) frequency (0-7), F0 envelope, voicing probability, jitter local, jitter DPP, and shimmer local and more MFCC features.

VGGish: This is an acoustic embedding model which is pre-trained on YouTube's Audio dataset [28]. The architecture of VGGish is a CNN-based structure and similar to VGG. The VGGish embedding model extracts and transforms the audio features into semantic and meaningful high-level feature vectors with 128 dimensions.

X-Vector: X-vector is a deep neural network-based audio embedder, widely used in the field of speech recognition [29, 30]. We employ x-vector to represent audio features from raw audio files. The neural network that produces the x-vector consists of three components: the frame-level layers to extract representation from MFCC, a statistics pooling layer which receives output from the last frame-level layer and a segment-level layer that follows the statistics pooling layer to generate the x-vector. Specifically, we obtain the x-vector features according to the following steps: 1) First, all raw audio files are normalized and re-sampled to 16,000Hz and 16-bits by using SOX audio processing software; 2) Second, we compute the x-vector for each audio segment by using Kaldi that uses the SER16 pre-trained x-vector model. The SER16 pre-trained model is trained on Switchboard, Mixer 6, and NIST SERs datasets [29, 30]; 3) Third, we convert x-vector to a binary file to make it easier for our proposed model to read.

3.1.2. Features from Transcribed Text

We used Speech-to-Text API [1] provided by Google cloud to automatically transcribe speech recordings. Then based on the transcripts, we extracted linguistic features and sentence embeddings.

Linguistic Features: We used two tools to generate linguistic features: 1) Like [2], we converted transcripts into CHAT format, then ran EVAL and FREQ commands in CLAN [31] to generate a composite profile of 34 measures and Moving Average Type Token Ratio [32]; 2) we generated 22 Part-of-Speech tags using NLTK [15]. After removing all-zero and duplicate features, 50 linguistic features in total were extracted.

Sentence embeddings: We used Universal Sentence Embedding (USE) [33] to represent each sentence in the context.

3.2. Proposed Architectures: C-Attention Networks

We propose a modular multi-modal architecture consisting of three standalone networks. The architectures are shown in Fig 1. The left-hand side leg processes acoustic features, such as Emobase and IS10 features, and is called C-Attention-Acoustic Network. The middle leg processes linguistic features, and is called C-Attention-FT Network. The right-hand side leg processes embedding features, such as USE, VGGish and X-Vector, and is called C-Attention-Embedding Network.

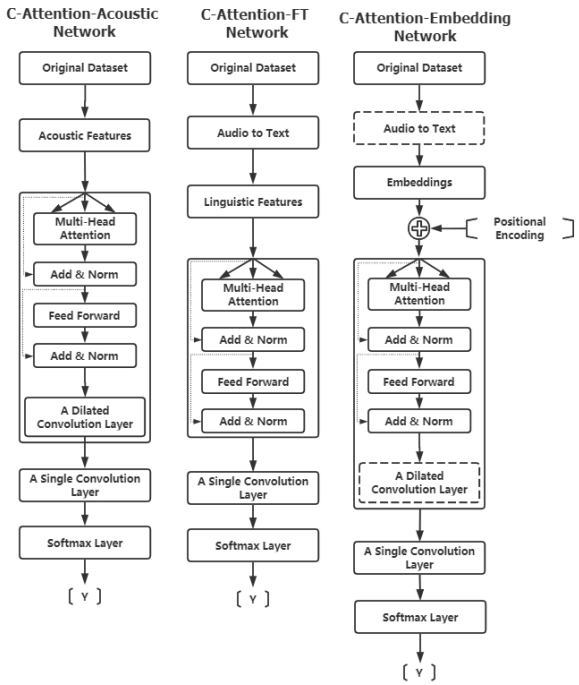


Figure 1: *The proposed architecture of C-Attention-Acoustic Network, C-Attention-FT Network and C-Attention-Embedding Network. The C-Attention-Acoustic Network uses acoustic features, the C-Attention-FT network uses the linguistic features, and the C-Attention-Embedding network uses embeddings of the patient/control's recordings.*

[1] https://cloud.google.com/speech-to-text

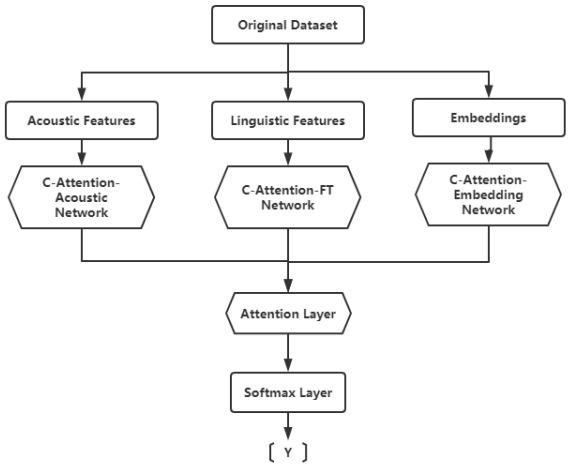


Figure 2: *The Architecture of Unified C-Attention Network for Acoustic Features, Linguistic Features, and Embeddings*

3.2.1. C-Attention Acoustic Model

This architecture (C-Attention Acoustic Network) is depicted on the left-hand side of Figure 1. The C-Attention Acoustic Model comprises of a multi-head-attention (MHA) module together with a dilated convolution layer [34, 35]; followed by a 1-D CNN layer and a softmax layer. We used the same MHA module and the encoder structure of the transformer that was proposed in [36]. Let $R = \{r_1, r_2, ..., r_n\}$ be the set of speech recordings, then r_i is the i^{th} record in the dataset. We extract acoustic feature sets presented in Sec 3.1.1 and generate the acoustic feature vectors, Let $F = \{F_1, F_2, ..., F_n\}$ be the set of acoustic feature vectors, and F_i be the i^{th} vector in the acoustic matrix. The MHA transforms the feature matrix $F = \{F_1, F_2, ..., F_n\}$ to another matrix of n-dimensional vectors $A = \{A_1, A_2, ..., A_n\}$. After each MHA module, we use a dilated convolution layer to further distill the MHA matrix $A = \{A_1, A_2, ..., A_n\}$ to half its original size. This is done to reduce the dimensions of the acoustic features which are too large for the attention mechanism to capture interactions well. This procedure forwards from j^{th} layer into $(j+1)^{\text{th}}$ layer as

$$\mathrm{X_{j+1}} = \mathrm{MaxPool}\left(\mathrm{ELU}\left(\mathrm{Conv1d}\left(\mathrm{X_j}\right)\right)\right) \tag{1}$$

Where the Conv1d(·) performs a 1-D convolutional filters and ELU(·) [37] is the activation function. The MHA and dilated CNN module is followed by a 1-layer CNN and a softmax layer to get the final classification.

3.2.2. C-Attention FT Model

This architecture (C-Attention FT Network) is depicted in the middle of Figure 1. It is proposed to capture the interaction among linguistic features. This architecture is similar to the proposed C-Attention (Sec 3.2.1) except for the removal of dilated CNN layer.

3.2.3. C-Attention Embedding Model

This architecture (C-Attention Embedding Network) is depicted on the right-hand side of Figure 1. We propose this architecture as a means of capturing latent feature information implicit in embeddings. This architecture is similar to the proposed C-Attention (Sec 3.2.1) except for the addition of a positional encoding module. The positional encoding module is used to maintain the relative positions of the embedding features and is the same as that used in the transformer [36] architecture. More specifically, the Audio to Text layer is only applied to text embeddings and the dilated convolution layer is only used on X-Vector embeddings.

3.2.4. C-Attention Unified Model

This architecture (C-Attention Unified Network) is depicted in Figure 2. In this architecture, we use all three types of features: acoustic features, linguistic features, and embedding features. We used another attention layer to fuse the outputs from C-Attention Acoustic Network, the C-Attention-FT network, and the C-Attention-Embedding network followed by a softmax layer. In order to fuse these other models together, we omit the final softmax layers in each of the four modules.

4. Experiments

In this work, we employed four models on acoustic features, linguistic features, and embeddings to detect AD, and evaluated our proposed models on the ADReSSo challenge dataset.

4.1. Dataset

In this work, we employed four models on acoustic features, linguistic features, and embeddings to detect AD, ane dataset is a balanced sub-dataset of the DementiaBank [38] with respect to age and gender. It consists of spontaneous speech recordings of spoken picture descriptions elicited from participants through the Cookie Theft picture description in the Boston Diagnostic Aphasia Exam [39]. The training set consists of 166 speech recordings, including 87 speech recordings from AD patients and 79 speech recordings from healthy controls. The testing set consists of 71 speech recordings without annotations.

4.2. Experiment Setup

We implemented our proposed models using Pytorch and trained them using the 10-fold cross-validation (CV) approach. Three types of features were extracted: acoustic features, linguistic features, and embeddings (including text embeddings and acoustic embeddings). For all models implemented in this paper, each model has 6 multi-head attention layers. Apart from that, in the C-Attention-Acoustic network and C-Attention-Embedding Network, each multi-head attention module is followed by a dilated convolution layer (kernel width=3) and a max-pooling layer with stride 2 which downsizes the feature set into its half slice. We found that due to the variation of feature size, the best configuration of modules is slightly different among various feature sets. For Emobase and IS10 features, 6 multi-head attention modules and 6 dilated CNN modules gave the best performance. However, 6 multi-head attention modules plus two dilated CNN modules is the best setting for X-Vector embeddings. Dilated CNN modules were not used on VGGish embeddings.

4.3. Feature Generation

We have described how to generate each type of feature in Sec 3.1.1. Here we add a few additional explanations on acoustic features and acoustic embeddings used in our experiments.
Acoustic Features: We generated Emobase and IS10 on each

speech recording, no segmentation was applied.

Acoustic Embeddings: 1) **VGGish Embeddings**: We applied 16k-downsampling on single-channel audio signals, and computed the log mel spectrogram. Then each log mel spectrogram was segmented with a non-overlapping 960ms window. Finally, we generated 128-length VGGish embedding on each segmented sample; 2) **X-Vector embeddings**: Similarly, we segmented each speech recording with a non-overlapping 960ms window, then generated 512-length X-Vector embedding on each segmented sample.

4.4. Experiment Results

The performance results are shown in Table 1. We note that on the training dataset, the C-Attention-Unified model with Linguistic and X-Vector features achieved the best performance in respect to the accuracy, precision, and F1 score, the best Recall was achieved by the C-Attention-Unified model with Linguistic, IS10, and X-Vector. Due to time limitation, in C-Attention-Unified network, we were not able to use all feature sets, such as USE. Given C-Attention-Embedding (USE) did not perform better than other approaches but it took longer to train, we truncated this feature set in our unified model.

Table 1: *AD classification accuracy on 10-fold cross-validation (CV)*

Approach	Accuracy	Precision	Recall	F1
C-Attention-Acoustic(Emobase)	0.614	0.632	0.632	0.632
C-Attention-Acoustic(IS10)	0.62	0.615	0.736	0.67
C-Attention-FT(Linguistic)	0.735	0.753	0.736	0.735
C-Attention-Embedding(USE)	0.657	0.679	0.655	0.667
C-Attention-Embedding(VGGish)	0.735	0.759	0.724	0.741
C-Attention-Embedding(X-Vector)	0.753	0.774	0.747	0.76
C-Attention-Unified(Linguistic + USE)	0.711	0.714	0.747	0.73
C-Attention-Unified(Linguistic + VGGish)	0.747	0.771	0.736	0.753
C-Attention-Unified(Linguistic + X-Vector)	**0.772**	**0.787**	0.74	**0.763**
C-Attention-Unified(Linguistic + IS10 + X-Vector)	0.725	0.724	**0.778**	0.75

Our experiment results would indicate that: 1) using both audio embeddings and linguistic features seems to be the best way to approach the problem of detecting AD, rather than choosing only one; 2) On the text side, handcrafted linguistic features perform better than USE representations on AD detection; 3) However, on the audio side, audio embeddings, such as X-Vector and VGGish show better performance on AD detection than frame-level acoustic features, such as Emobase and IS10.

Further analysis of the values in this table would indicate that using only the latent NLP-based features does not perform as well as using only audio embeddings (X-Vector). However, it is worthy to mention that the transcripts used in this work were automatically converted from speech recordings. The automatic conversion might have introduced errors and noises. Within audio embeddings, X-Vector performs better than VGGish.

As part of the ADReSSo challenge, we were provided the test dataset and asked to submit the labels from five attempts of our algorithm on this dataset. Since we had multiple models, we used the following method to decide on which model's result to report. We randomly split the training dataset into 80% for training, 10% for validation, and 10% for testing. We tried multiple random seeds, then used the models which performed best, on an average, on the training dataset. The best performing model was the C-Attention-Unified model with Linguistic features and X-vectors. Hence we used this model to submit the five attempts on the test dataset as required by the challenge rules. The organizers then calculated the accuracy, precision, recall and F scores based on the ground truth labels (which were not revealed to the participants) of the test dataset. Table 2 shows the results returned to us by the organizers, for our model.

Table 2: *Attempts on test dataset*

Attempts	Accuracy	Precision		Recall		F1	
Attempt 1	**0.8028**	0.7500	0.8889	0.9167	0.6857	**0.8250**	0.7742
Attempt 2	0.7746	0.7500	0.8065	0.8333	0.7143	0.7895	0.7576
Attempt 3	0.7887	0.7692	0.8125	0.8333	0.7429	0.8000	0.7761
Attempt 4	0.7746	0.7273	0.8519	0.8889	0.6571	0.8000	0.7419
Attempt 5	0.7606	0.7111	0.846	0.8889	0.6286	0.7901	0.7213

5. Future Work

In this challenge, due to time limitation, we were not able to apply segmentation on acoustic features, nor apply 100ms window size segmentation on either VGGish or X-Vector embeddings. We believe that our models could learn better on the acoustic features if time series segmentation is applied. Further, we will continue to address the other subtasks set out in the challenge, viz.: evaluate the models' performance by calculating the MMSE score and generalize the proposed models to predict the cognitive decline.

6. Conclusions

In this paper, we proposed a modular multimodal approach to detect Alzheimer's disease and this approach includes four architectures using CNN and multi-head attention on the training set of the ADReSSo Challenge. Three types of feature sets were used in this work: acoustic features, linguistic features, and embeddings. One architecture uses only the acoustic features, one architecture uses only the linguistic features, one uses only the embeddings and the unified architecture uses all of those features. Extensive experimental evaluations on the training dataset show that our proposed model can detect AD with an accuracy of 77.2%, F1 of 0.763 using the C-Attention-Unified model with Linguistic and X-Vector features. Using the same model and feature sets, the best accuracy of our models was 80.28% and F1 of 0.825 on the test dataset.

7. 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] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," in *Submitted to INTERSPEECH 2021*, 2021. [Online]. Available: https://edin.ac/31eWsjp

[3] 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.

[4] 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.

[5] 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.

[6] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language pro-

cessing technologies to detect alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[7] P. Garrard, L. M. 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*, vol. 128, no. 2, pp. 250–260, 2005.

[8] V. Berisha, S. Wang, A. LaCross, and J. 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*, vol. 45, no. 3, pp. 959–963, 2015.

[9] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[10] N. Wang, M. Chen, and K. P. Subbalakshmi, "Explainable CNN-attention networks (c-attention network) for automated detection of alzheimer's disease." ACM SIGKDD, 2020. [Online]. Available: https://arxiv.org/pdf/2006.14135.pdf

[11] N. Wang, F. Luo, R. Chandramouli, K. P. Subbalakshmi, and V. Peddagangireddy, "Personalized early stage alzheimer's disease detection: A case study of president reagan's speeches." ACL, 2020.

[12] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[13] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*, vol. 3. IEEE, 2005, pp. 1569–1574.

[14] 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.

[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] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[23] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[24] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," *Proc. Interspeech 2020*, pp. 4961–4965, 2020.

[25] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[26] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, 2020, pp. 2187–2191.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[34] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[35] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *arXiv preprint arXiv:2012.07436*, 2020.

[36] 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.

[37] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[38] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

[39] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

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

WavBERT: Exploiting Semantic and Non-semantic Speech using Wav2vec and BERT for Dementia Detection

Youxiang Zhu[1]*, Abdelrahman Obyat*[1]*, Xiaohui Liang*[1]*, John A. Batsis*[2]*, and Robert M. Roth*[3]

[1]Department of Computer Science, University of Massachusetts Boston, MA, USA
[2] School of Medicine, University of North Carolina, Chapel Hill, NC, USA
[3] Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

{Youxiang.Zhu001, Abdelrahman.Obyat001, Xiaohui.Liang}@umb.edu
John.Batsis@unc.edu, Robert.M.Roth@hitchcock.org

Abstract

In this paper, we exploit semantic and non-semantic information from patient's speech data using Wav2vec and Bidirectional Encoder Representations from Transformers (BERT) for dementia detection. We first propose a basic WavBERT model by extracting semantic information from speech data using Wav2vec, and analyzing the semantic information using BERT for dementia detection. While the basic model discards the non-semantic information, we propose extended WavBERT models that convert the output of Wav2vec to the input to BERT for preserving the non-semantic information in dementia detection. Specifically, we determine the locations and lengths of inter-word pauses using the number of blank tokens from Wav2vec where the threshold for setting the pauses is automatically generated via BERT. We further design a pre-trained embedding conversion network that converts the output embedding of Wav2vec to the input embedding of BERT, enabling the fine-tuning of WavBERT with non-semantic information. Our evaluation results using the ADReSSo dataset showed that the WavBERT models achieved the highest accuracy of 83.1% in the classification task, the lowest Root-Mean-Square Error (RMSE) score of 4.44 in the regression task, and a mean F1 of 70.91% in the progression task. We confirmed the effectiveness of WavBERT models exploiting both semantic and non-semantic speech.

Index Terms: Speech analysis, automatic speech recognition, non-semantic information, dementia detection

1. Introduction

Researchers have exploited spontaneous speech for early detection of Alzheimer's disease, as the collection of speech data is more practical and less costly compared to conventional cognitive assessment methods such as neuropsychological evaluation [1] and Magnetic Resonance Imaging (MRI) [2]. In the 2020 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, researchers studied the spontaneous speech dataset [3] and demonstrated that transcript-based models are more effective in dementia detection than audio-based models [4, 5, 6, 7]. We envisioned that the low performance of the audio-based models is due to the large variance and hard-to-interpret nature of audio signals [8, 9]. In comparison, transcript-based models were built using the manual transcripts from human transcription, which takes advantage of the human transcriber's knowledge, including rules of the description task and information units within the picture. However, human transcription is a costly and impractical process, which prevents the speech-based evaluation from being a fully automatic approach. The 2021 ADReSS speech only (ADReSSo) challenge thus aims at the development of fully automatic models that detect dementia using only speech data [10].

Automatic Speech Recognition (ASR) aims to generate ASR transcripts automatically from speech audio data. ASR transcripts can be used as the inputs to transcript-based models when manual transcripts are not available. However, we have two concerns. First, ASR might generate a transcript with uncertain errors, especially for the speech from patients with cognitive impairment. Such uncertain errors might negatively affect the performance of transcript-based models. Second, the integration of ASR could analyze both semantic and non-semantic information for a more accurate dementia detection, while transcript-based models focus on the analysis of semantic information of the transcripts only. Previous research demonstrated the usefulness of non-semantic information in dementia detection, such as filled and silent pauses [11, 12, 13], paralinguistic features [10, 14], and Mel Frequency Cepstral Coefficient (MFCC) [15, 8, 9]. As such, we seek an effective integration of ASR and transcript-based models for enhanced dementia detection. For the transcript-based models, Bidirectional Encoder Representations from Transformers (BERT) dominates the Natural language processing (NLP) research due to its power of self-supervised training and transfer learning strategy [16]. It consists of two steps: i) pre-training a model with unlabeled data and self-supervised training strategy, and ii) fine-tuning the model with downstream data and tasks. In the ADReSS 2020 challenge, it was proven that transcript-based model BERT outperforms traditional machine-or-deep learning models using handcrafted features in dementia detection [17].

We recently explored transfer learning over audio dataset directly for dementia detection but achieved limited performance [8, 9]. Our conclusion is that the selected pre-trained audio models did not extract good representation from the audio data of the dementia task from the dementia detection perspective. However, transfer learning has been effective in ASR; researchers proposed Wav2vec model [18] and the corresponding ASR achieved state-of-the-art Word Error Rate (WER) on the LibriSpeech dataset [19]. We thus applied the Wav2vec ASR model to generate ASR transcripts from the 2020 ADReSS dataset, compared the ASR transcripts with the manual transcripts, and found the high similarity of the two transcripts.

In this paper, we first propose a basic WavBERT model to generate ASR transcripts via Wav2vec, use the ASR transcripts and dementia-related labels to fine-tune BERT, and derive the dementia detection results of the testing dataset using the fine-tuned BERT. While the basic WavBERT model discards the non-semantic information, we propose two extended WavBERT models that utilize the intermediate results of the

http://dx.doi.org/10.21437/Interspeech.2021-332

ASR, containing the non-semantic information. Specifically, the **first method** is to derive the locations and lengths of inter-word pauses by counting the blank tokens and separate tokens from the intermediate results of the Wav2vec ASR. Furthermore, we do not manually set the thresholds for pauses. Instead, we propose automatic methods that use BERT or training samples to determine the thresholds. The **second method** is to convert the Wav2vec output embedding to the BERT input embedding using a pre-trained embedding conversion network. The module is pre-trained with a large-sized audio dataset and its corresponding ASR transcripts and assists the fine-tuned BERT to detect dementia using both semantic and non-semantic information from the speech data. Our contributions are three-fold:

First, we propose a basic WavBERT model that concatenates Wav2vec ASR with BERT, enabling an automatic process of dementia detection.

Second, we extend the WavBERT model to determine the locations and lengths of pauses using the ASR intermediate results. The thresholds for setting pauses are automatically generated. The extended WavBERT model achieves the highest accuracy of 83.1% in the classification task and the lowest Root-Mean-Square Error (RMSE) score of 4.44 in the regression task.

Third, we extend the WavBERT model by converting the Wav2vec output embedding to the BERT input embedding for preserving non-semantic information. The extended WavBERT achieves the highest accuracy of 70.91% in the progression task.

2. ADReSSo dataset

The ADReSSo challenge consists of three tasks, an AD classification task, a Mini-Mental State Examination (MMSE) regression task, and a cognitive decline progression task [10]. The first two tasks share the same data, including 237 audio files, which were collected using a Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam [20]. The data is balanced with class, age, and gender. The data for the cognitive decline progression task was collected from a category fluency task, including 105 audio files. The first-round data was provided as the baseline, and the second-round data was collected in two years and used for inferring cognitive decline. The data of this task is unbalanced; the non-decline samples are significantly more than the decline samples. In the challenge, 70% of both datasets were used for training and 30% for testing [10].

3. Basic WavBERT

We propose a basic WavBERT model consisting of Wav2vec ASR and BERT, labeled to path 1 in the Figure 1. The basic model converts speech data to ASR transcripts and inputs the ASR transcripts to BERT for dementia detection.

Wav2vec aims to learn the representations from speech data using self-supervised training [18]. As shown on the left of the Figure 1, Wav2vec first inputs speech data into a Convolutional Neural Network (CNN) to obtain the latent representations, which are then inputted into a transformer encoder. The transformer encoder generates context representations in the output embedding and employs a pre-training task following the self-supervised training strategy [18]. After pre-training, Wav2vec uses a fine-tuning process with a character inference component, optimized with a Connectionist Temporal Classification (CTC) loss. The character inference component consists of a 1D convolutional layer and a softmax layer. The convolutional layer convolutes according to the time dimension using both kernel size and stride set to 1. The output of the character inference component can be 26 English letters, single quote ', blank token <s>, and separator token |. Finally, Wav2vec merges consecutively repeated characters, removes blank tokens, and uses separator tokens to separate words. The transcript has no punctuation, contains the semantic information of the speech data, and can be inputted to transcript-based models.

BERT derives the general representation of the language model by employing a pre-training process with large-scale datasets (i.e., BooksCorpus and Wikipedia) [16]. BERT generally includes the four steps depicted on the right of the Figure 1. Given a transcript, BERT first pre-processes the transcript with the WordPiece tokenizer [21], splits words into sub-word-level tokens, and then adds special tokens [CLS] and [SEP]. All tokens are converted to an input embedding, which is further inputted to a transformer encoder [22] to obtain the output embedding. Two pre-training tasks were adopted: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, given that some of the input tokens are masked, the classification objective is to use the embedding of the unmasked tokens to infer the masked tokens. In NSP, given that a single [SEP] token is inserted between two selected sentences, a binary classification objective is to infer whether the first sentence is followed by the second in the transcript.

Inference layers. For the dementia detection task, we use the BERT output embedding of all the tokens except for the [CLS] token as the input of the inference layers. Specifically, we use a 1D convolutional layer with both kernel size and stride set to 1. The number of neurons is equal to the hidden size of the BERT. We use a Global Average Pooling (GAP) layer to calculate an averaged vector according to the time dimension, and use a Fully Connected (FC) layer with softmax or LeakyReLU for the classification or regression tasks, respectively.

4. Extended WavBERT

We extend WavBERT models with ASR pause preservation and embedding conversion, shown in paths 2 and 3 of the Figure 1.

4.1. ASR pause preservation

Force alignment methods are often used to align transcripts and audio data and determine both inter-word and inter-sentence pauses [13]. However, as the ASR produces uncertain errors, force alignment between the ASR transcript and audio data might not be effective. In our model, we exploit the CTC property of Wav2vec for determining the pauses. Wav2vec produces blank tokens and separate tokens as intermediate results. As such, we modify the Wav2vec post-processing as follows. First, we merge the consecutively repeated letters and single quotes. Second, we remove the blank token between English letters and the single quote. In this way, the blank tokens within any word were removed. Last, we combine the remaining blank tokens with the separate tokens, count the number of the blank and separate tokens between words, and use that number of tokens to determine the lengths of pauses between words.

BERT requires the input as transcripts and punctuation marks. Therefore, we convert pauses to the punctuation marks, i.e., periods and commas. Specifically, we design automatic methods to determine the thresholds ϵ_p and ϵ_c, which are used to set sentence-level pauses and in-sentence pauses.

Sentence-level pause. BERT has extensive prior knowledge of sentence-level pauses from large-scale pre-training datasets. Thus, we use BERT to determine the threshold ϵ_p for sentence-level pauses. Specifically, we aim to maximize

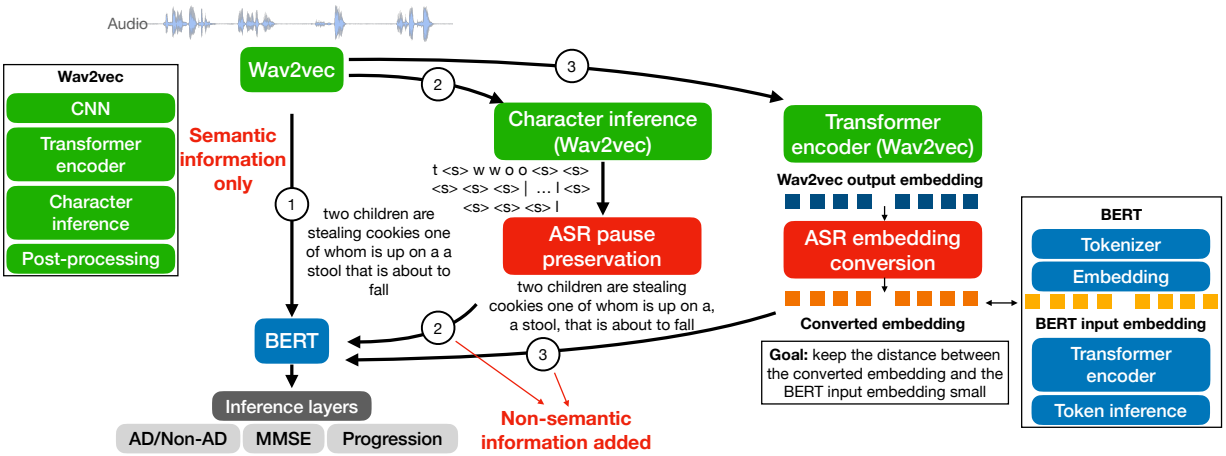


Figure 1: *The basic WavBERT model (Path 1) and extended WavBERT models (Paths 2&3)*

the sample-level cross-entropy difference between the AD and non-AD samples from the training dataset. We consider n_a AD samples $\mathcal{X}_a$ and n_{na} non-AD samples $\mathcal{X}_{na}$. These samples are generated using Wav2vec and do not include any punctuation marks. Using the BERT tokenizer, we obtain k tokens of a transcript as $x_i = \{x_{i,1}, \cdots, x_{i,k}\}$. Given a threshold ϵ_p, we determine the locations of pauses that have lengths $\geq \epsilon_p$, then insert punctuation marks of periods at these locations, and finally obtain t tokens of the transcript as $x_i^{\epsilon_p} = \{x_{i,1}^{\epsilon_p}, x_{i,2}^{\epsilon_p}, \ldots, x_{i,t}^{\epsilon_p}\} \in \mathbb{R}^{t*v}$ $(t \geq k)$. Each token $x_{i,j}^{\epsilon_p} \in \mathbb{R}^v$ is a v-size vector where v is the vocabulary size of BERT. Then we input $x_i^{\epsilon_p}$ to BERT to obtain the corresponding self-supervised token inference after softmax activation $z_i^{\epsilon_p} \in \mathbb{R}^{t*v}$. Then, we remove the tokens of the same indexes of added punctuation marks from $z_i^{\epsilon_p}$ and obtained $\bar{z}_i^{\epsilon_p} \in \mathbb{R}^{k*v}$. We aim to find ϵ_p to maximize the target:

$$\underset{\epsilon_p}{\operatorname{argmax}} | \underset{x_i \in \mathcal{X}_a}{\operatorname{Med}}^{n_a} (\ell_i^{\epsilon_p}) - \underset{x_i \in \mathcal{X}_{na}}{\operatorname{Med}}^{n_{na}} (\ell_i^{\epsilon_p}) | \tag{1}$$

where $\mathrm{Med}()$ is the median function, $\ell_i^{\epsilon_p}$ is the cross-entropy loss of the sample x_i:

$$\ell_i^{\epsilon_p} := \frac{1}{k} \sum_{j=1}^{k} \sum^{v} -log(\bar{z}_{i,j}^{\epsilon_p}) * x_{i,j} \tag{2}$$

In-sentence pause. BERT has no prior knowledge of in-sentence pauses. Thus, we design a statistical method to determine the threshold for locating the in-sentence pauses. Specifically, we measure the lengths of selected pauses in both AD and non-AD samples where we set the maximum length of pauses as ϵ_p. Then, we count the number of pauses with length β as $\pi_{\beta,a}$ and $\pi_{\beta,na}$ for $1 \leq \beta < \epsilon_p$ for AD and non-AD samples, respectively. Finally, we aim to find ϵ_c to maximize the target:

$$\begin{aligned} \underset{\epsilon_c}{\operatorname{argmax}} | \frac{1}{n_a} \sum_{\beta=1}^{\epsilon_c} \pi_{\beta,a} - \frac{1}{n_{na}} \sum_{\beta=1}^{\epsilon_c} \pi_{\beta,na} | + \\ | \frac{1}{n_a} \sum_{\beta=\epsilon_c+1}^{\epsilon_p} \pi_{\beta,a} - \frac{1}{n_{na}} \sum_{\beta=\epsilon_c+1}^{\epsilon_p} \pi_{\beta,na} | \end{aligned} \tag{3}$$

After we determine pauses using ϵ_p and ϵ_c, we insert periods for sentence-level pauses and commas for in-sentence pauses in the ASR transcripts and input the transcripts to BERT.

4.2. ASR embedding conversion

An embedding conversion network converts Wav2vec output embedding to BERT input embedding. The network design faces two challenges: i) the Wav2vec output embedding is at the character-level, while the BERT input embedding is at the sub-word-level; and ii) in order to utilize the pre-training parameters of BERT, we should make the converted embedding close to the BERT input embedding.

For the first challenge, we design a mapping method. We generate an ASR transcript from an audio sample, use the BERT tokenizer to derive sub-word tokens from the ASR transcript, and generate the BERT input embedding for each token. Then, we use the characters of the token to find the corresponding Wav2vec output embedding. Although the characters generated from Wav2vec can be repeated, we can effectively identify any repeated characters using the property of CTC. For example, the Wav2vec output embedding of "t <s> w w o o" corresponds to the BERT input embedding of token "two." Finally, we average the Wav2vec output embedding according to the time dimension to obtain the averaged embedding as e_w and map e_w to the BERT input embedding e_b of the corresponding token. Special tokens such as [CLS] and [SEP] were excluded in this process.

For the second challenge, we design an embedding conversion network with an aim to convert the Wav2vec output embedding e_w to the BERT input embedding e_b. The network consists of two 1D convolutional layers with a layer-norm in between. The 1D convolutional layers use both kernel size and stride set to 1. The number of neurons is equal to the hidden size of BERT. We further design a pre-training process. First, we run the Wav2vec ASR on LibriSpeech [19] to obtain the ASR transcripts and the Wav2vec output embedding, and then we use the BERT tokenizer to obtain the BERT input embedding from the ASR transcripts. We input the Wav2vec output embedding into the embedding conversion network and optimize its outputs with BERT input embedding using l_1 loss. In the training step, after we obtain the converted embedding, we add the embedding of tokens of punctuation marks (from the pause preservation) and special tokens, and finally, input the integrated embedding to BERT.

5. Evaluation

We implemented five models: M_b uses the ASR transcripts as input to BERT, M_{p1} extends M_b by adding sentence-level pauses to ASR transcripts, M_{p2} extends M_b by adding both sentence-level and in-sentence pauses to ASR transcripts, M_e extends M_b with embedding conversion, and M_{e+p2} extends M_b with embedding conversion and sentence-level/in-sentence

Table 1: *Results of classification, regression, and progression tasks over ADReSSo testing dataset. The design of the baseline linguistic model and the definitions of precision, recall, F1, mean F1, accuracy, and RMSE can be found at the baseline paper [10].*

Task	1. Classification (%)						2. Regression	3. Progression (%)					
	Class	Precision	Recall	F1	Mean F1	Accuracy	RMSE	Class	Precision	Recall	F1	Mean F1	Accuracy
Baseline [10]	non-AD	80.00	77.80	78.87	78.87	78.87	5.28	non-decline	83.30	68.20	75.00	66.67	68.75
	AD	77.80	80.00	78.87				decline	50.00	70.00	58.30		
M_b	non-AD	71.79	77.78	74.67	73.16	73.24	4.60	non-decline	64.00	72.73	68.09	39.92	53.13
	AD	75.00	68.57	71.64				decline	14.29	10.00	11.76		
M_{p1}	non-AD	80.00	88.89	84.21	83.02	**83.10**	4.45	non-decline	62.96	77.27	69.39	34.69	53.13
	AD	87.10	77.14	81.82				decline	0	0	0		
M_{p2}	non-AD	77.50	86.11	81.58	80.19	80.28	**4.44**	non-decline	64.29	81.82	72.00	36.00	56.25
	AD	83.87	74.29	78.79				decline	0	0	0		
M_e	non-AD	78.95	83.33	81.08	80.25	80.28	4.46	non-decline	79.17	86.36	82.61	69.08	75.00
	AD	81.82	77.14	79.41				decline	62.50	50.00	55.56		
M_{e+p2}	non-AD	77.78	77.78	77.78	77.46	77.46	4.47	non-decline	81.82	81.82	81.82	**70.91**	75.00
	AD	77.14	77.14	77.14				decline	60.00	60.00	60.00		

pauses. We report the results of the five models in Table 1.

5.1. Implementation and training strategy

We trained the five models with the ADReSSo training dataset and reported the performance of five models over the provided testing dataset. Considering the random states of the models and the limited size of the dataset, we trained each model for 10 rounds and submitted the average results of the 10 rounds. For the classification and progression tasks, we averaged the probabilities from the softmax activation. For the regression task, we averaged the output of MMSE scores. For the classification task, the corresponding models output the non-AD class only if its probability is ≥ 0.5. For the progression task, we implemented classification models for the progression task by treating decline samples as AD samples and non-decline as non-AD. Considering the unbalanced classes of the training dataset, the corresponding models output non-decline class only if its probability is ≥ 0.79, based on the class ratio of the training dataset.

We implemented the five models with PyTorch[1], employing the "bert-base-uncased" and "wav2vec-vox-960h-pl" settings. We filtered out one ASR transcript that has < 20 words in the progression training dataset, which could be caused by either the failure of ASR or inaudible samples. In training, we unfroze all BERT layers and inference layers while freezing the embedding conversion network. We used batch size 8 and learning rate 10^{-6} with the Adam optimizer [23]. We used the cross-entropy loss for the first and third tasks, and we used the mean squared error for the second task. We trained our models with a maximum of 2000 epochs and stopped the training if the loss is smaller than 10^{-6}. Besides, we used a similar setting as above for the pre-training of the embedding conversion network with LibriSpeech [19], but changed the learning rate to 5×10^{-5} and the maximum number of epoch to 100. One-round training using the ADReSSo training dataset took less than 6 hours with one V100 GPU, and one-round pre-training of the embedding conversion network took less than a day with six V100 GPUs.

5.2. Experimental results on testing dataset

The Wav2vec models outperformed the baseline model in all three tasks. Our observations are as follows:

Classification. As shown in Table 1, the basic WavBERT M_b achieved an accuracy of 73.24%. With non-semantic information added into the analysis, M_{p1} and M_e achieved 83.10% and 80.28%, respectively. These accuracy improvements confirmed that the effectiveness of our models, which utilized pause preservation and embedding conversion for non-semantic information. However, the results of M_{p2} and M_{e+p2} exploiting in-sentence pauses were worse compared to the M_{p1} and M_e. We consider that the in-sentence pauses produced a negative impact because the in-sentence pauses were learned from the limited training datasets, which may lead to overfitting. In comparison, the sentence-level pauses were automatically derived using BERT, which provided a positive impact.

Regression. The basic WavBERT M_b produced an RMSE score of 4.60, lower than 5.28 of the baseline model. All extended WavBERT outperformed the basic WavBERT. Specifically, M_{p1} with sentence-level pauses produced an RMSE score of 4.45, M_{p2} with sentence-level/in-sentence pauses further lowered the RMSE score to 4.44, and M_e with embedding conversion produced an RMSE score of 4.47. The performance improvements confirmed that both pause preservation and embedding conversion produced a positive impact. Lastly, M_{e+p2} produced a slightly larger RMSE score, which may have resulted from the limited training dataset and overfitting problem.

Progression. M_b, M_{p1} and M_{p2} resulted in poor performance. By checking the ASR transcripts, we found that the transcripts have a word-misspelling problem for two reasons. First, Wav2vec is a character-level model, and thus the transcripts may have added or missed characters of words. Second, the progression dataset was collected from a category fluency task, significantly different from the training dataset of Wav2vec ASR, thus downgrading the ASR performance. However, M_e and M_{e+p2} with embedding conversion, achieved mean F1 scores 69.08%, 70.91%, outperforming 66.67% of the baseline. We considered that the embedding conversion network effectively mitigated the word-misspelling problem by inputting embedding, not misspelled transcripts, to BERT.

6. Conclusions

We propose WavBERT models by integrating Wav2vec ASR with BERT for an automatic process of dementia detection. While the basic Wav2BERT used ASR transcripts and focused on semantic information, the extended Wav2BERT exploits non-semantic information with a pause preservation module and an embedding conversion network. Our experimental results confirmed that the extended WavBERT models outperformed the baseline linguistic model. Our future goal includes exploring the transformer encoder of BERT for pre-training the embedding conversion network.

7. Acknowledgements

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. 1R01AG067416.

[1]Codes are available at https://github.com/billzyx/WavBERT

8. References

[1] M. D. Lezak, D. B. Howieson, D. W. Loring, J. S. Fischer *et al.*, *Neuropsychological assessment*. Oxford University Press, USA, 2004.

[2] E. E. Bron, M. Smits, W. M. Van Der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. M. Orellana, R. Meijboom *et al.*, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CAD Dementia challenge," *NeuroImage*, vol. 111, pp. 562–579, 2015.

[3] 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.

[4] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[5] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting Multi-Modal Features From Pre-trained Networks for Alzheimer's Dementia Recognition," *arXiv preprint arXiv:2009.04070*, 2020.

[6] E. Edwards, C. Dognin, B. Bollepalli, M. Singh, and V. Analytics, "Multiscale System for Alzheimer's Dementia Recognition through Spontaneous Speech," *Proc. Interspeech 2020*, pp. 2197–2201, 2020.

[7] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[8] Y. Zhu and X. Liang, "Exploiting fully convolutional network and visualization techniques on spontaneous speech for dementia detection," *arXiv preprint arXiv:2008.07052*, 2020.

[9] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring Deep Transfer Learning Techniques for Alzheimer's Dementia Detection," *Frontiers in Computer Science*, vol. 3, p. 22, 2021.

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

[11] K. C. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, 2013, pp. 47–54.

[12] 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.

[13] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for Detection of Alzheimer's Disease," *Frontiers in Computer Science*, vol. 2, p. 57, 2020.

[14] F. Haider, S. De La Fuente, and S. Luz, "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[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] 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.

[17] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection," *arXiv preprint arXiv:2008.01551*, 2020.

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] J. T. Becker, F. Boller, 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.

[21] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

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