frontiers
in Computer Science

BRIEF RESEARCH REPORT
published: 09 February 2021
doi: 10.3389/fcomp.2021.624659

Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech

***Zehra Shah*[1]*, *Jeffrey Sawalha*[2], *Mashrura Tasnim*[1], *Shi-ang Qi*[1], *Eleni Stroulia*[1] and *Russell Greiner*[1,2,3]**

[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada, [2]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada, [3]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

Alzheimer's dementia (AD) is a chronic neurodegenerative illness that manifests in a gradual decline of cognitive function. Early identification of AD is essential for managing the ensuing cognitive deficits, which may lead to a better prognostic outcome. Speech data can serve as a window into cognitive functioning and can be used to screen for early signs of AD. This paper describes methods for learning models using speech samples from the DementiaBank database, for identifying which subjects have Alzheimer's dementia. We consider two machine learning tasks: 1) binary classification to distinguish patients from healthy controls, and 2) regression to estimate each subject's Mini-Mental State Examination (MMSE) score. To develop models that can use acoustic and/or language features, we explore a variety of dimension reduction techniques, training algorithms, and fusion strategies. Our best performing classification model, using language features with dimension reduction and regularized logistic regression, achieves an accuracy of 85.4% on a held-out test set. On the regression task, a linear regression model trained on a reduced set of language features achieves a root mean square error (RMSE) of 5.62 on the test set. These results demonstrate the promise of using machine learning for detecting cognitive decline from speech in AD patients.

Keywords: speech and audio classification, pathological speech and language, automatic analysis of speaker states and traits, machine learning, natural language proceeding (NLP)

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Marko Meža,
University of Ljubljana, Slovenia
Michal Przemyslaw Muszynski,
Université de Genève, Switzerland

****Correspondence:***
Zehra Shah
zshah@ualberta.ca

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *08 January 2021*
Published: *09 February 2021*

Citation:
Shah Z, Sawalha J, Tasnim M, Qi S, Stroulia E and Greiner R (2021) Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech. Front. Comput. Sci. 3:624659. doi: 10.3389/fcomp.2021.624659

1 INTRODUCTION

Alzheimer's Dementia (AD) has recently become one of the leading causes of death in people over 70 years (Alzheimer's Association (2019)). With life expectancy increasing, the prevalence of AD among older adults is also rising. Currently, the number of cases among people over the age of 60 is doubling every 4–5 years, and currently, one in every three individuals over the age of 80 is likely to develop AD (Ritchie and Lovestone (2002)). AD is a progressive neurodegenerative disorder that is characterized by the loss of subcortical neurons and synapses that begins in areas such as the hippocampus and the entorhinal cortex (Braak and Braak (1991); Terry et al., (1991)). Over time, more associative areas begin to show amyloid deposition and neurofibrillary tangles in addition to neuronal and synaptic loss. As it spreads, patients develop additional cognitive and functional deficits in domains such as attention, executive function, memory and language (Nestor et al., (2004)). Current theories maintain that clinical symptoms are preceded by subtle cognitive deficits that worsen over time. Early recognition of these deficits could prove valuable for treating pre-stage AD, allowing for a better quality of life for the patient and their caregivers.

Currently, clinical diagnostic methods for determining who has AD include cognitive assessments (e.g., Mini-Mental State Examination [MMSE]), self-report questionnaires and neuroimaging (e.g., Positron Emission Tomography [PET]) (Weller and Budson (2018)). While these methods have proven useful, they suffer from several shortcomings. Cognitive assessments can be tedious and suffer from low test-retest reliability based on practice effects; self-report questionnaires also lack reliability and validity; and neuroimaging is an expensive, invasive, and time-consuming procedure.

By contrast, speech analysis is a simple, non-invasive and inexpensive approach. There are several reasons why it may be useful for detecting AD. Early identification, especially in the prodromal stages, can significantly reduce the progression of various cognitive deficits (Dubois et al., (2009)). There is evidence that therapeutic interventions are most efficacious before neuronal degeneration occurs in the brain (Nestor et al., (2004)). Thus, an emphasis on early detection is imperative for the prognosis of AD. As such, episodic memory, visuospatial ability, and confusion are some of the first signs of cognitive decline in AD patients (Arnáiz and Almkvist (2003); Jacobs et al., (1995)). These deficits can be observed through verbal communication in a structured task, motivating the recent use of speech data for diagnostic screening of AD in elder patients (Chien et al., (2019)). In our study, we used machine learning (ML) approaches to distinguish between AD and control patients, using acoustic and linguistic features from spontaneous speech produced by a subject describing a picture.

The current literature on detecting AD from spontaneous speech samples can be divided into two main categories. One class of systems analyzes linguistic features (lexicon, syntactic and semantic information), while the other deals with acoustic-dependent features. In the acoustic domain, AD patients exhibit longer and more frequent hesitations, lower speech and articulation rates, and longer pauses compared to control participants in spontaneous speech tasks (Hoffmann et al., (2010); Szatloczki et al., (2015)). Some have attempted to apply ML approaches to learn models that use acoustic features to distinguish AD from control participants. Tóth et al., (2018) learned a model for distinguishing early stage AD patients from control patients using spontaneous speech from a recall task. Their classification model found significant differences in speech tempo, articulation rate, silent pause, and length of utterance. Mirzaei et al., (2017) tried to improve on previous models by examining temporal features (jitter, shimmer, harmonics-to-noize ratio, Mel frequency cepstral coefficients [MFCCs]).

Conversational transcripts contain rich information about the speaker, such as the wealth of their vocabulary, the complexity of their syntactic structures, and the information and meanings they communicate. Previous research has shown that language changes in patients who suffer from AD (Wankerl et al., (2017); Kempler (1995))–e.g., these patients often have difficulty naming objects within specific categories, replacing forgotten words with pronouns and repeating certain words or phrases (Kirshner (2012); Adlam et al., (2006); Nicholas et al., (1985)). This has motivated numerous research projects on conversation samples in AD and control patients. Fraser et al., (2016) examined picture description transcripts from demented vs. control individuals. Subsequently, they also analyzed acoustic features in addition to natural language, and achieved an accuracy of 81%. They found that semantic information was one of the best features (syntactic fluency, MFCCs and phonation rate) for separating AD from control participants.

Our paper is motivated by the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, hosted by the INTERSPEECH 2020 conference (Luz et al., (2020)). The data set provided in this challenge is a carefully curated subset of the larger DementiaBank corpus (Becker et al., (1994)). Among the various challenge submissions, the top-performing models analyzed both linguistic and acoustic features, and many of these top submissions used deep learning methods (including some pre-trained models) to generate their results. For example, Koo et al., (2020) used an ensemble approach with bi-modal convolutional recurrent neural networks (cRNN), applied to a variety of feature sets from pre-trained acoustic and linguistic algorithms in addition to some hand-crafted features. They achieved an accuracy of 81.25% on their classifier evaluation and an RMSE score of 3.75. Another study by Balagopalan et al., (2020) achieved an accuracy of 83.33% and an RMSE of 4.56 by adding a binary classification layer to a pre-trained language algorithm developed by Google: Bidirectional Encoder Representations from Transformers–BERT. The Sarawgi et al., (2020) submission applied RNNs and multi-layered perceptrons (MLP) to various types of acoustic and linguistic features in an ensemble approach. They also used transfer learning from the classification models to the MMSE scores by modifying the last layer structure, achieving an RMSE of 4.6 and an accuracy of 83.33%. Lastly, Searle et al., (2020) used linguistic features only, with pre-trained Transformer based models, and achieved their best performance using features computed from the full transcripts (including both participant and interviewer speech). They obtained a classification accuracy of 81% and an RMSE of 4.58. The commonality among these top submissions was the use of deep-learning methods, along with pre-trained acoustic and/or language models.

Our study hopes to improve further by applying simple, computationally inexpensive ML techniques to natural language and acoustic information. In particular, we train models that use both acoustic and language features to distinguish AD from healthy age-matched elders and predict their MMSE scores. Our system feeds the acoustic features into one pipeline, and the linguistic ones in another. Each pipeline preprocesses the features, then uses internal cross-validation to tune the hyperparameters and select the relevant subset of features. We use ensemble methods to combine the various learned models, to produce models that can 1) label a speech sample as either AD or non-AD, and 2) predict the associated MMSE scores.

2 METHOD

For this study, we were given a training set of 54 AD patients and an age- and gender-matched set of 54 healthy controls (this is a subset of the larger DementiaBank data set; see Becker et al., (1994)). This subset of DementiaBank contained spontaneous speech samples of participants asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam (Goodglass et al., (2001)). For each participant, we obtained 1) the original recorded speech sample, 2) the normalized speech segments extracted from the full audio sample after voice activity detection, audio normalization and noise removal, as well as 3) the speech transcript files annotated using CHAT (Codes for Human Analysis of Transcripts) transcription format (MacWhinney (2017)). Additionally, some descriptive features were given about these individuals, including age, gender, binary class label (AD/non-AD; the target for the classification task), and their MMSE score (which we try to predict in the regression task). The MMSE has a maximum score of 30, and lower MMSE scores are generally associated with progressively more severe dementia. The challenge organizers withheld a test set containing data from 24 AD and 24 control participants for final evaluation. For further details of this data set, we refer the reader to Luz et al., (2020).

We considered a set of possible base learners, each over a subset of the features–the (1), (2), and (3) mentioned above. We used internal 5-fold cross validation to identify which of these base learners was best. Due to the size of our data, we chose to use a 5-fold CV procedure. 10-fold CV or Leave-one-out CV procedure would result in small partitions, leading to possible overfitting (low bias, higher variance). To ensure consistent and reliable comparison between our models, we defined and used a common set of folds that were balanced in terms of class labels (or MMSE scores) as well as gender. For each model, we evaluated performance metrics (average accuracy for classification, and average RMSE for regression) based on these test folds, as well as on the final hold-out test set.

2.1 Language and Fluency Features

The organizers provided transcripts that were annotated using the CHAT coding system (MacWhinney (2017)). First we extracted only the participant's speech from these transcripts (removing the interviewer's content). Then, using the CLAN (Computerized Language Analysis) program for processing transcripts in the CHAT format, we computed the following set of global syntactic and semantic features for each transcript: type-token ratio (TTR)–the number of unique words divided by total number of words; mean length of utterance (MLU), where an utterance is a speech fragment beginning and ending with a clear pause; number of verbs per utterance; percentage of occurrence of various parts of speech (nouns, verbs, conjunctions, etc.); number of retracings (self-corrections or changes); and number of repetitions. We also computed a number of fluency features, including percent of broken words, part-word and whole-word repetitions, sound prolongations, abandoned word choices, word and phrase repetitions, filled pauses, and non-filled pauses. In total, we computed 62 such informative summary features for each transcript.

2.2 N-Gram Features

We processed the raw (unannotated) transcripts to compute bag-of-words and bigram features. First, we standardized the transcripts by converting them into a list of word tokens. Next, we used the WordNet lemmatizer (Miller (1998)) to find and replace each word with the corresponding lemma; for example, words like "stands", "standing" and "stood" were all replaced by the common root word "stand". Finally, we removed stopwords from each transcript, where stopwords are highly common (and presumably uninformative) words that may add noise to the data (such as "I", "am", "was", etc.), using a predefined stopwords list from the Python natural language toolkit (NLTK) package.

Next, we used the standardized transcripts to compute bag-of-words vectors (using words seen in the training set only)–that is, a vector of 514 integers for each transcript, where the *k*th value is the number of times the *k*th word occurred–and normalized these vectors with the Term Frequency-Inverse Document Frequency (TF-IDF) function, which is a normalization procedure that reflects how important a word is to a document in a corpus–effectively penalizing words that occur frequently in most of the documents in the corpus. For example, in our case the word "boy" might occur frequently in all transcripts, so it may not be very informative. Finally, we also computed bigram vectors in a manner similar to bag-of-words–where each bigram is a *pair* of words that appeared adjacent to one another. We found a set of 2,810 bigrams.

2.3 Acoustic Features

Using the speaker timing information provided in the transcripts, we extracted the participants' utterances (removing the interviewer's voice) from the audio recordings, for a total of 1,501 participant utterances from the training set, and 592 from the test set. We then normalized the audio volume across all speech segments. We computed four different sets of features from each audio segment using OpenSMILE v2.1 (Eyben et al., (2010)). Note that our overall learner will consider various base-learners, each running on one of these feature sets.

(FeatureSet#1) The **AVEC 2013** (Valstar et al., (2013)) feature set includes 2,268 acoustic features including 76 low level descriptor (LLD) features and their statistical, regression and local minima/maxima related functionals. The LLD features include energy, spectral and voicing related features; delta coefficients of the energy/spectral features, delta coefficients of the voicing related LLDs and voiced/unvoiced duration based features.

(FeatureSet#2) The **ComParE 2013** (Schuller et al., (2013)) feature set includes energy, spectral, MFCC, and voicing related features, logarithmic harmonic-to-noize ratio (HNR), voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, leading to a total of 6,373 features.

(FeatureSet#3) Our third feature set consists of the following three feature sets. The **emo_large** (Eyben et al., (2010)) feature set consists of cepstral, spectral, energy and voicing related features, their first and second order delta coefficients as LLDs; and their 39 statistical functionals. The functionals are computed over 20 ms frames in spoken utterances. This produced 6,552 acoustic features across the utterances. The **Jitter-shimmer**

feature set is a subset of INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al., (2010)) feature set, consisting of three pitch related LLDs and their delta coefficients. We also computed 19 statistical functionals of the LLDs on the voiced sections of the utterances, resulting in 114 features. Finally, we extracted seven speech and articulation rate features by automatically detecting syllable nuclei (De Jong and Wempe (2009)), and used a script from the software program Praat to detect peaks in intensities (dB) followed by sharp dips. We also calculated other features, such as words per minute, number of syllables, phonation time, articulation rate, speech duration and number of pauses for each speech sample (Chakraborty et al., (2020)).

(FeatureSet#4) We computed the **MFCC 1–16** features and their delta coefficients from 26 Mel-bands, which uses the fast Fourier transform (FFT) power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Inclusion of statistical functionals resulted in 592 features. This feature set is a subset of AVEC 2013 feature set (Valstar et al., (2013)).

We also added age and gender of the participants to each set of features.

2.4 Language-Based Models

Given our two sets of linguistic features above (**Sections 2.1 and 2.2**), we explored various dimension reduction techniques and base learning algorithms to find the best performing pipeline. The dimension reduction techniques include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and univariate feature selection using ANOVA F-values. The base learning algorithms explored for the classification task are logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). For the regression task, the regression versions of the same algorithms are trained (except logistic regression is replaced by linear regression). Internal 5-fold cross-validation was used to tune the hyperparameters for each model based on accuracy. The hyperparameters explored were:

Dimension reduction: For classification models, dimension reduction with PCA using {10, 20, 30, 50} components, and LSA using {100, 200, 500} components; for regression models, dimension reduction with PCA using {20, 30, 50} components, and LSA using {200, 500, 800} components.

Models: SVM (regularization parameter C: {0.1, 1, 10, 100, 1,000}, kernel: {linear, RBF, polynomial}); LR (regularization parameter C: 20 values spaced evenly on a log scale in the range $[10^{-4}, 10^{4}]$, loss function: {L1, L2}); RF (number of trees: {100, 300, 500, 700}, maximum features at each split: {5, 15, 25, 35, 45, 55}, minimum samples at leaf node: {1, 2, 3, 4}); and XGB (maximum depth: {5, 6, 7, 8}, learning rate: {0.02, 0.05, 0.07, 0.1}, number of trees: {50, 100, 200, 500, 1,000}). The same hyperparameters were explored for the regression models as well (with the exception of replacing LR with linear regression).

Our internal cross-validation found the best-performing *language-based classification* model, which consisted of the following steps:

Step1: 5-component PCA transformation of the *dense* language and fluency features described in **Section 2.1** (after standardizing using z-scores);
Step2: 50-component LSA transformation of the *sparse* unigram and bigram features described in **Section 2.2** (after standardizing using TF-IDF transform); and
Step3: L1-regularized logistic regression.

The best language-based regression model involved the following:

Step1: 30-component PCA transformation of the *dense* language and fluency features described in **Section 2.1** (after standardizing using z-scores);
Step2: 100-component LSA transformation of the *sparse* unigram and bigram features described in **Section 2.2** (after standardizing using TF-IDF transform); and
Step3: Random Forest Regressor, using 100 trees, minimum of four instances at each leaf node, and 25 features considered for each split.

2.5 Acoustic Models

All acoustic features were real values and were therefore standardized using z-scores. We used PCA to reduce the dimensionality of the features sets. For FeatureSet#1 and FeatureSet#2, we used PCA, and kept the minimum number of features capable of retaining 95% of the variance. In case of FeatureSet#3 and FeatureSet#4, the number of principals were determined through internal 5-fold cross-validation. Therefore, the dimension of FeatureSet#1 is reduced from 2,268 to 700, FeatureSet#2 from 6,373 to 1,100, FeatureSet#3 from 6,552 to 1,000 and FeatureSet#4 from 592 to 50. Next, we selected the best 50 principal components from FeatureSet#1, and the best 70 from FeatureSet#3 applying univariate feature selection method based on ANOVA F-value between the label and each feature. For FeatureSet#2, we calculated feature importance weights using a decision-tree regression model, and selected only the features with importance weight higher than the mean.

After this pre-processing stage, our system fed these audio features to various machine-learning algorithms, that each identify patterns of features that can distinguish dementia patients from healthy controls (the classification task), and can compute a subject's MMSE score (the regression task). We explored several learning algorithms, including Adaboost, XGB, RF, gradient boosting (GBT), decision trees (DT), hidden Markov model (HMM) and neural network (NN). Internal 5-fold cross-validation was performed to tune the hyperparameters of the classifiers and regressors. The predictions were made in two steps. In the first step, the classifiers and regressors were trained and tested with acoustic features, age and gender to predict whether the speech segment was uttered by a health control or an AD patient and to predict that subject's MMSE score. Next, weighted majority vote classification was performed to assign each subject a label of health control or AD, based on the majority labels of the segment level classification. The predicted MMSE scores on all the segments of one subject were averaged to calculate the final MMSE score of that subject. The best performing classifiers on acoustic data are the following:

(1) Neural network with one hidden layer, trained on FeatureSet#1.

TABLE 1 | Results of our best performing classification models distinguishing AD from non-AD subjects. The "Baseline (Acoustic)" model is described in Luz et al. (2020). The right-most column shows accuracy on the held-out test set of 48 subjects (24 AD and 24 non-AD). The rest of the table lists model performance using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Classifiers	Class	Precision	Recall	F1 score	Accuracy	Accuracy (hold-out set)
	AD	1.0	0.60	0.75		
Logistic regression (NLP)	HC	0.71	1.00	0.83	80% ± 0.00%	85%
	OVR	0.86	0.80	0.79		
	AD	0.68	0.84	0.75		
SVM (NLP)	HC	0.79	0.60	0.68	72% ± 1.85%	73%
	OVR	0.73	0.72	0.72		
	AD	0.74	0.96	0.83		
Majority vote (NLP + Acoustic)	HC	0.94	0.66	0.78	81% ± 1.17%	83%
	OVR	0.84	0.81	0.81		
	AD	0.71	0.78	0.74		
Majority vote (Acoustic)	HC	0.76	0.68	0.72	73% ± 1.36%	65%
	OVR	0.73	0.73	0.73		
	AD	0.57	0.52	0.54		
Baseline (Acoustic)	HC	0.56	0.61	0.58	57%	63%
	OVR	0.57	0.57	0.56		

AD, Alzheimer's dementia; HC, Healthy control; OVR, Overall rating.

(2) AdaBoost Classifier with 50 estimator and logistic regression as base estimator, trained on FeatureSet#4.
(3) Adaboost with 100 estimators and DT as the base estimator trained on FeatureSet#3.

The three regressors with the lowest RMSE were:

(1) Gradient boosting regressor, trained on FeatureSet#4.
(2) Decision tree with number of leaves 20, trained on FeatureSet#2.
(3) Adaboost regressor trained on FeatureSet#3 with 100 estimators.

2.6 Ensemble Methods

After obtaining our best-performing acoustic and language-based models, we computed a weighted majority-vote ensemble meta-algorithm for classification. We chose the three best-performing acoustic models along with the best-performing language model, and computed a final prediction by taking a linear weighted combination of the individual model predictions. The weights assigned to each model were proportional to that model's mean cross-validation accuracy, such that the best performing model is given the highest weight in the final prediction. For regression, we also computed an unweighted averaging of our best language and acoustic model predictions for MMSE scores.

3 RESULTS

3.1 Classification

Table 1 presents the results for the classification task. The model that obtained the highest average cross-validation accuracy (81% ± 1.17%) is a weighted-majority-vote ensemble of the best language-based model and three of the best acoustic-based models. The second highest accuracy (80% ± 0.00%) was obtained by the language-based logistic regression. However, a McNemar test reveals that these two models do not exhibit a statistically significant difference in performance (McNemar test statistic = 4.0, $p > 0.05$). This is also evident by the performance of these two models on the final held-out set, where the language-based logistic regression gives the highest accuracy (85%) and the weighted-majority-vote ensemble gives a slightly lower accuracy (83%). Using McNemar's test to compare these two models on the held-out test set, we obtain a test statistic of 3.0, with $p > 0.05$, indicating that the performance difference between these models is not statistically significant.

Note that our ensemble model, which uses only acoustic features, performs significantly better than the "baseline model" (provided by the organizers), which also uses acoustic features only.

3.2 MMSE Prediction

Table 2 shows the RMSE of various regression models; columns 2 and 3 show the average RMSE and R^2 scores over the five cross-validation folds, and columns 4 and 5, on the hold-out test set (provided by the organizers of the challenge). These results show that the language-based model obtains the best RMSE of 6.43 on the cross-validation set and 5.62 on the hold-out set. The combined language-acoustic model did not perform as well as the standalone language-based model, with an average RMSE of 6.83 on the cross-validation set and 6.12 on the hold-out set.

Further, the Wilcoxon test between the RMSEs of the two best models (best acoustic + best language-based combination vs. best stand-alone language-based), returns a test statistic of 66.0 with $p < 0.05$ on the hold-out set, and a test statistic of 1,375.0 with $p < 0.05$ on the cross-validation set. This means we cannot reject the claim that these two models are significantly different in performance.

We also report the coefficient of determination (R^2) for all our models: the best R^2 was 0.17 on the validation folds and 0.14 on the held-out test set. These low numbers are expected, given the relatively small size of this INTERSPEECH challenge data set and

TABLE 2 | Results of our best performing regression models predicting a subject's MMSE score (ranging from 0 to 30, with lower values indicating more severe dementia). The 'Baseline (Acoustic)' model is described in Luz et al. (2020). As in **Table 1**, the columns on the right show RMSE and R^2 on the held-out test set of 48 subjects (24 AD and 24 non-AD). The middle columns list RMSE and R^2 using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Regressors	RMSE	R^2	RMSE (hold-out set)	R^2
Random forest (NLP)	6.43 ± 0.18	0.17	5.62	0.14
Gradient boosting (acoustic)	6.89 ± 0.17	0.06	6.67	−0.21
Random forest (NLP) + gradient boosting (acoustic)	6.66 ± 0.18	0.13	6.01	0.02
Majority vote (all models)	6.85 ± 0.16	0.10	6.12	−0.02
Baseline (acoustic)	7.30	–	6.14	–

the complexity of the condition. Interpreting this statistic in an absolute sense is problematic, especially as we did not find any other study using the same data set that reported this metric. We note that models based on language features achieved the best R^2 values, which further supports our claim that language features are very important for this task.

4 DISCUSSION

We investigated a variety of ML models, using language and/or acoustic features, to identify models that performed well at using speech information to distinguish AD from healthy subjects, and to estimate the severity of AD. Our results, of over 85% accuracy for classification and approximately 5.6 RMSE for regression, demonstrate the promise of using ML for detecting cognitive decline from speech. In our investigation, we explored multiple different combinations of features and ML algorithms; in the future, it would be interesting to delve deeper into the behavior of our best models, to determine the contribution of individual (or groups of) features to the model's ability to distinguish AD patients from healthy controls. Further, although we have currently used the full set of standard stopwords for removing noise in our language models, it may be worthwhile to see whether using a reduced set of stopwords (for example, preserving pronouns) might be more advantageous.

Our current best-performing models outperform recent results reported in the literature and provide evidence that, for discriminating between subjects with AD vs. healthy controls, features based on language (semantics, fluency and n-grams) are very useful. Compared to other top ranked results, our methods do not involve complex, computationally expensive algorithms. Instead, we used an ensemble approach with simple models to produce competitive results. Furthermore, a weighted majority vote of acoustic and language based models demonstrates competitive performance, implying that a combination of acoustic and language features also holds potential. Finally, comparing only acoustic models, we find that accuracy improves significantly compared to the baseline model (Luz et al., (2020)) for both the classification and regression tasks.

Our competitive performance, obtained using simple feature engineering along with classical machine learning algorithms, indicates that putting together an efficient machine learning pipeline from basic building blocks can achieve nearly state-of-the-art results for the learning tasks explored in this study. This result suggests that, for detecting AD from speech, it may be useful to explore traditional feature engineering and machine learning tools, especially in a limited data setting, as this will additionally provide for better interpretability and reproducibility compared to more complex deep learning based methods.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the DementiaBank corpus of the TalkBank repository [https://dementia.talkbank.org/].

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TalkBank Code of Ethics Carnegie Mellon University [https://talkbank.org/share/ethics.html]. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of this study. ZS prepared the validation sets, and developed and tested the language based models. JS, MT, and SQ developed and tested the acoustic models. ZS, JS, MT, and SQ wrote sections of the manuscript. All authors contributed to manuscript editing and revision, and approved the submitted version.

FUNDING

RG and ZS gratefully acknowledge the funding from NSERC (Canada's Natural Science and Engineering Research Council) and Amii (Alberta Machine Intelligence Institute); and JS gratefully acknowledges the funding from IBM and MITACS. MT and ES gratefully acknowledge support from the AGE-WELL NCE; and MT gratefully acknowledges support from the Alberta Graduate Excellence Scholarship (AGES).

ACKNOWLEDGMENTS

The authors thank the ADReSS challenge organizing committee, including FH, SL, BM, and DF, for their efforts and support.

REFERENCES

Adlam, A.-L. R., Bozeat, S., Arnold, R., Watson, P., and Hodges, J. R. (2006). Semantic knowledge in mild cognitive impairment and mild alzheimer's disease. *Cortex* 42, 675–684. doi:10.1016/s0010-9452(08)70404-0

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's Demen.* 15, 321–387. doi:10.1016/j.jalz.2019.01.010

Arnáiz, E., and Almkvist, O. (2003). Neuropsychological features of mild cognitive impairment and preclinical alzheimer's disease. *Acta Neurol. Scand.* 107, 34–41. doi:10.1034/j.1600-0404.107.s179.7.x

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection.Preprint repository name [Preprint]. Available at: arXiv:2008.01551 (Accessed July 26, 2020).

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi:10.1001/archneur.1994.00540180063015

Braak, H., and Braak, E. (1991). Neuropathological stageing of alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi:10.1007/BF00308809

Chakraborty, R., Pandharipande, M., Bhat, C., and Kopparapu, S. K. (2020). Identification of dementia using audio biomarkers. Preprint repository name [Preprint]. Available at: arXiv:2002.12788 (Accessed February 27, 2020).

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., and Fu, L.-C. (2019). An automatic assessment system for alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep.* 9, 1–10. doi:10.1038/s41598-019-56020-x

De Jong, N. H., and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* 41, 385–390. doi:10.3758/BRM.41.2.385

Dubois, B., Picard, G., and Sarazin, M. (2009). Early detection of alzheimer's disease: new diagnostic criteria. *Dialog. Clin. Neurosci.* 11, 135–139. doi:10.31887/DCNS.2009.11.2/bdubois

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on multimedia, Firenze, Italy, October 25–29, 2010 (New York, NY: AMC), 1459–1462.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *J. Alzheim. Dis.* 49, 407–422. doi:10.3233/JAD-150520

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston diagnostic Aphasia examination*. 3rd Edn. Philadelphia, PA: Lippincott Williams and Wilkins.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in alzheimer's disease. *Int. J. Speech Lang. Pathol.* 12, 29–34. doi:10.3109/17549500903137256

Jacobs, D. M., Sano, M., Dooneief, G., Marder, K., Bell, K. L., and Stern, Y. (1995). Neuropsychological detection and characterization of preclinical alzheimer's disease. *Neurology* 45, 957–962. doi:10.1212/wnl.45.5.957

Kempler, D. (1995). Language changes in dementia of the alzheimer type. *Demen. Commun.* 7, 98–114.

Kirshner, H. S. (2012). Primary progressive aphasia and alzheimer's disease: brief history, recent evidence. *Curr. Neurol. Neurosci. Rep.* 12, 709–714. doi:10.1007/s11910-012-0307-2

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. Preprint repository name [Preprint]. Available at: arXiv:2009.04070 (Accessed September 09, 2020).

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the adress challenge. Preprint repository name [Preprint]. Available at: arXiv:2004.06833 (Accessed April 14, 2020).

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format. Available at: http//childes.psy.cmu.edu/manuals/CHAT.pdf (Accessed April 2014).

Miller, G. A. (1998). *WordNet: an electronic lexical database*. Cambridge, MA: MIT press, 449.

Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo Senge Muvingi, C., Cristancho-Lacroix, V., et al. (2017). "Automatic speech analysis for early Alzheimer's disease diagnosis," in JETSAN 2017: 6e Journées d'Etudes sur la Télésanté, Bourges, France, May–June 31–01, 2017 (Bourges, France: JETSAN), 114–116.

Nestor, P. J., Scheltens, P., and Hodges, J. R. (2004). Advances in the early detection of alzheimer's disease. *Nat. Med.* 10, S34–S41. doi:10.1038/nrn1433

Nicholas, M., Obler, L. K., Albert, M. L., and Helm-Estabrooks, N. (1985). Empty speech in alzheimer's disease and fluent aphasia. *J. Speech Lang. Hear. Res.* 28, 405–410. doi:10.1044/jshr.2803.405

Ritchie, K., and Lovestone, S. (2002). The dementias. *Lancet* 360, 1759–1766. doi:10.1016/S0140-6736(02)11667-9

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. (2020). Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. Preprint repository name [Preprint]. Available at: arXiv:2009.00700 (Accessed August 30, 2020).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010). "The interspeech 2010 paralinguistic challenge," in Eleventh annual Conference of the international speech communication association, Makuhari, Chiba, Septmber 26–30, 2010 (Makuhari, Japan: ISCA), 3137.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in Proceedings INTERSPEECH 2013, 14th annual conference of the international speech communication association, Lyon, France, August 25-29, 2013 (France, EU: International Speech Communication Association (ISCA)), 3500.

Searle, T., Ibrahim, Z., and Dobson, R. (2020). Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. Preprint repository name [Preprint]. Available at: arXiv:2006.07358 (Accessed June 12, 2020).

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Front. Aging Neurosci.* 7, 195. doi:10.3389/fnagi.2015.00195

Terry, R. D., Masliah, E., Salmon, D. P., Butters, N., DeTeresa, R., Hill, R., et al. (1991). Physical basis of cognitive alterations in alzheimer's disease: synapse loss is the major correlate of cognitive impairment. *Ann. Neurol.* 30, 572–580. doi:10.1002/ana.410300410

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi:10.2174/1567205014666171121114930

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, Barcelona, Spain, October, 2013 (New York, NY: ACM), 3–10.

Wankerl, S., Nöth, E., and Evert, S. (2017). "An n-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in Interspeech 2017, Stockholm, Sweden, August 20–24, 2017 (Stockholm, Sweden: ISCA), 3162–3166.

Weller, J., and Budson, A. (2018). Current understanding of alzheimer's disease diagnosis and treatment. *F1000Res.* 7, F1000. doi:10.12688/f1000research.14506.1

Conflict of Interest: The authors declare that this study received funding from the IBM Center for Advanced Studies. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.