ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.024

Acoustic and Linguistic Analyses to Predict Cognitive Decline Progression in Alzheimer's Disease

Chen Qian, Jian Wen and Jiyun Li +

School of Computer Science and Technology, Donghua University Shanghai, China

Abstract. Alzheimer's Disease (AD) is a type of dementia that progressively destroys cognitive functions and, eventually, the ability to carry out daily tasks. Since current treatments aim to delay the disease, there is an urgent need to monitor and predict the disease progression in a straightforward manner. In this paper, we extract acoustic and linguistic features from patients' spontaneous speeches and then analyze them within various models, respectively.

Keywords: Alzheimer's disease, acoustic features, linguistic features, cognitive decline detection

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease with insidious onset and irreversible progression. It causes cognitive decline, which gradually affects patients' memory, language, and the ability to take care of themselves. The earlier the disease is diagnosed, the better the chance of delaying its progression [1]. However, for different patients, the decline rate and time are dissimilar. Current research is mainly concerned with detecting early signs before an individual meets the criteria for Alzheimer's dementia but lacks in the prediction of disease progression. While memory impairment is deemed as the main symptom of AD, language is also considered valuable clinic information [2]. The ubiquity of speech has led to a number of researches about exploiting speech features for the detection of AD. The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge provides a platform for researchers working on the prediction of disease progression based on a speech to build their own models using a new standardized dataset. Here in this paper, we develop approaches to predict cognitive decline progression based on acoustic and linguistic features extracted from the speech data provided by the ADReSS challenge.

2. Related Work

Over the past few years, there has been increasing research on speech technology for dementia detection. Machine learning models like linear discriminant analysis (LDA) and LSTM have been adopted in this area. Weiner et al. [3] used an LDA classifier with a set of acoustic features. However, machine learning cannot always get good performance because of their poor learning abilities. With the rapid development of deep learning, more models like convolutional neural network (CNN), BERT, and Wav2Vec have been used in speech classification [4-5]. But limited to the scale of medical datasets, deep learning models are rarely used for the prediction of disease progression.

In principle, there are mainly two kinds of features in speech recognition: one is acoustic features and the other is linguistic features. COVAREP defines some acoustic features, which can be extracted via an acoustic analysis framework software, namely openSMILE [6]. Other defined acoustic features, such as the number of pauses, pause proportion, phonation time, phonation-to-time ratio, speech rate, articulation rate, and noise-to-harmonic ratio, are also useful. Currently, VGGish becomes a widely accepted choice to extract these features. Linguistic features from spontaneous speech have been proved more applicable in cognitive decline detection, due to the built-in information can be straightforwardly analysed. At present, speech

+ Corresponding author. Tel.: +021-67792293;
E-mail address: jyli@dhu.edu.cn

transcripts are almost mandatory in ASR, such as Google cloud-based speech recognizer. It is worth noting that extracting linguistic features directly from audio is rarely performed.

3. Data and Features

3.1. Dataset

With the aim of predicting cognitive decline over a two-year period, the ADReSS dataset collects a range of speech recordings from a longitudinal cohort study, and the audios therein refer to Alzheimer's patients performing a category (semantic) fluency task at their baseline visit [7]. The overall length of recordings exceeds 2 minutes, with a maximum of 3 minutes. There is a total of 105 audio recordings split into a training set and a test set (70% and 30%, respectively). The former contains 15 audio recordings for the decline group and 58 ones for the no-decline group.

Apparently, the dataset is so small and imbalanced that further processing must be implemented. Firstly, we use an audio analysis tool called librosa to split audio recordings into 6-second segments (including 3-second overlap). As a result, we now have 713 recording segments for the decline group and 2106 for the no-decline group. Then, we balance the number of recordings across the two groups by discarding specific segments only containing the interviewer's speech. Eventually, the training set is made up of 713 and 704 recording segments for the decline group and no-decline group, respectively.

3.2. Acoustic Features

We use the openSMILE v2.3 toolkit to extract acoustic feature sets, such as IS12, ComParE16, and eGeMAPS. The theoretical significance, as well as the practical usefulness of the feature sets, has been discussed by existing works. IS12 contains 5757 features per 6000ms frame, while ComParE16 includes 6373. As an augmented version of the GeMAPS set, eGeMAPS contains 88 features. In brief, these feature sets contain F0 semitone, loudness, and other most common statistical functionals, which enable patient identification.

In addition, we extract sound spectrograms from the audios. A spectrogram is defined as a time-varying spectral representation, which provides a visual overview of a signal's loudness, or amplitude, as it varies over time at different frequencies. A Mel spectrogram is another spectrogram where the frequencies are converted to the Mel scale. It is generated to classify different signals and audio events. An MFCC, as the abbreviation of Mel-Frequency Cepstral Coefficients, uses a matrix to illustrate an accurate representation of the shape of the vocal tract. In this paper, we convert recording segments to spectral images so that CNN can be further applied to predict the decline progression.

3.3. Linguistic Features

Previous studies have pointed out that extracting linguistic features via pre-trained models is usually unsatisfactory, especially on classification tasks. In this paper, we put forward two methods to achieve the extraction. On the one hand, we use a device to transcribe the recordings automatically. Then, the resulting transcripts are converted into word embedding and sentence embedding by a pre-trained BERT model for text classification. On the other hand, we skip the transcription. Instead, the raw recordings are directly input to a Wav2Vec 2.0 model. The outcoming linguistic features are embedded and saved in a CSV file.

4. Acoustic-based Experiment

4.1. Proposed Approach

As aforementioned in Section 3.2, we proposed two types of acoustic features extracted by openSMILE and spectrogram, respectively. As Figure 1 shows, the former is classified by three machine learning models that can deal well with a small dataset: support vector machine (SVM), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost). Notably, we do not reduce feature dimensions, because our previous work has proven it cannot improve models' performance.

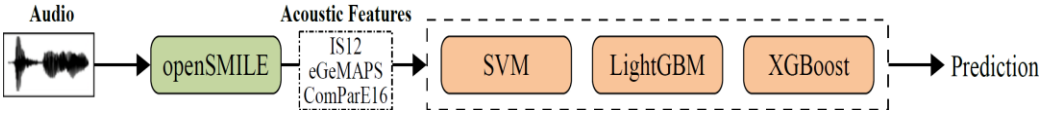

Fig. 1: Extracting acoustic features by SVM, LightGBM, and XGBoost.

We hereby propose another approach to extract acoustic features. Firstly, we generate Mel and MFCC spectrums from each 6-second speech segment, respectively. Then, we build a CNN model that takes spectrums as inputs. It is worth noting that attention is introduced into the CNN architecture in order to enhance its performance. Technically we adopt three attention methods referring to squeeze-and-excitation block, efficient channel attention, and convolutional block attention module [8]. The revamped CNN model is illustrated in Figure 2.

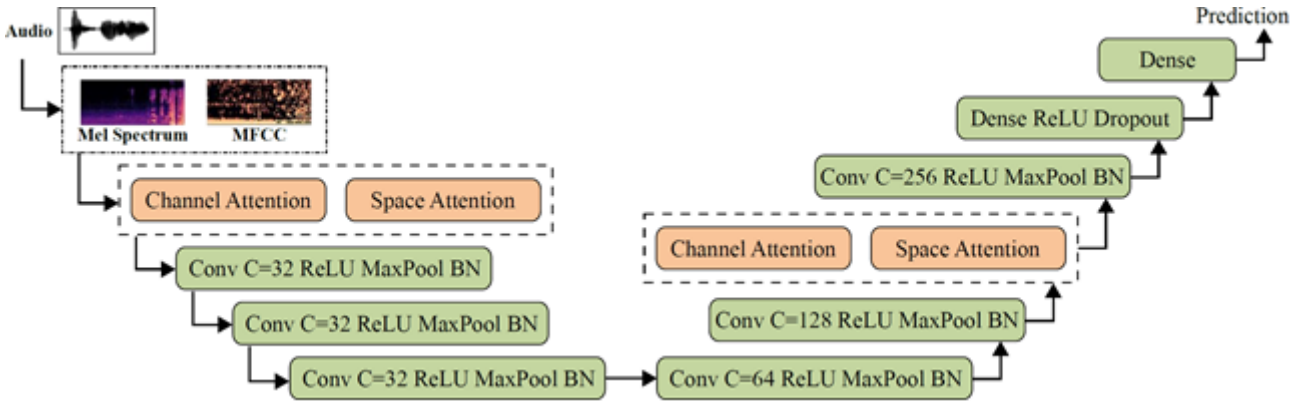

Fig. 2: Extracting acoustic features by CNN with an attention mechanism.

In both acoustic-based approaches, the final classification is made through a majority vote of all 6-second segments.

4.2. Experimental Setup

After exploring different speech-cutting methods, each long audio is split into 6-second segments with a 3-second overlap. In the former experiment, the SVM uses linear kernel and auto gamma, while the boosting type of LightGBM is set to 'GBDT' with 50 leaves. The max depth of XGBoost is 3, and the number of leaves is 50, too. In light of the small size of the training dataset, we choose leave-one-out cross-validation in the end.

In the latter experiment, all convolution kernels are 3x3, the moving stride is one, and the padding is two. In different layers, there are varying quantities of kernels. The CNN model has five convolution blocks, each of which contains a convolution layer, a ReLU function, a maximum pooling layer, and a batch normalization layer. The five convolution layers are defined with 32, 32, 32, 64, and 128 output channels, respectively. After several explorations, we embed the attention modules behind the first and the last convolution block, respectively. Dropout is a common stochastic regularization technique to improve generalization. We implement the spatial dropout for all five CNN layers with a dropout probability of 0.5. In the training process, the learning rate is set to 0.001, and the batch size is set to 128. A five-fold cross-validation method is used in this approach.

4.3. Experimental Result and Analysis

The experimental result of our proposed model using conventional acoustic features is shown in Table 1. From the result, we can see that SVM with IS12 achieves the best mean-F1 63%, which is an improved 2% compared to the baseline system. For the same machine learning model, IS12's performance is much better than other feature sets. It means IS12 gets better classification information, which helps to train the models well. Apparently, XGBoost and LightGBM do not live up to our expectations, which suggests that GBDT-based models are not appropriate for this classification task.

The experimental result of the CNN model using audio spectrums is shown in Table 2, which indicates that embedding the attention mechanism into CNN architecture can greatly improve the model's performance. When CBAM is embedded, the model achieves the best mean-f1 of 63%. However, we realize that 1D-CNN with MFCC can reach the same performance as the baseline, even though the attentions are not added to the basic model. It is caused by MFCC dislodging some irrelevant factors from the Mel spectrum.

Table 1: Comparing results of machine learning models

Models	**Features**	**Mean-F1**
SVM	**IS12**/eGeMAPS/ComParE16	**0.63**
XGBoost	IS12/eGeMAPS/ComParE16	0.56
LightGBM	IS12/eGeMAPS/ComParE16	0.41
LDA(baseline)	eGeMAPS	0.61

Table 2: Comparing results of CNN

Models	**Features**	**Mean-F1**
2D-CNN with SENet	Mel Spec	0.48
2D-CNN with ECA	Mel Spec	0.51
2D-CNN with CBAM	**Mel Spec**	**0.63**
LDA(baseline)	eGeMAPS	0.61

In conclusion, the results of both experiments point out that the acoustic features make the classifiers more prone to overfitting due to the speech segments remaining overlapped for 3 seconds.

5. Linguistic-based Experiment

5.1. Proposed Approach

Our approach is to model patients' speech as a sequence to predict whether they will have cognitive decline or not. Since pre-trained transformer-based models have potent abilities to extract semantic information from transcripts, we choose to use BERT and Wav2Vec 2.0 as the linguistic feature extractors. BERT has 768 embedding dimensions, whereas Wav2Vec2.0's is decided by the audio length. As shown in Figure 3, three machine learning models are classifiers.

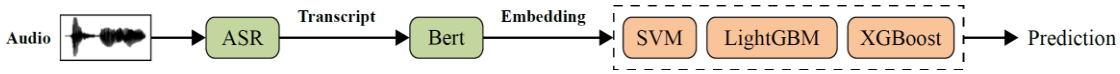

Fig. 3: Extracting linguistic features by SVM, LightGBM, and XGBoost.

Inspired by the acoustic spectrums, we consider using 1D-CNN and 2D-CNN to deal with BERT embedding and Wav2Vec 2.0 embedding, respectively. The output embedding of Wav2Vec 2.0 is located at the character level, while BERT's is at the sub-word level. It may result in a different performance of classification. Figure 4 illustrates the same task using pre-trained Wav2Vec 2.0 and a 2D-CNN model. It is worth noting that audio can be directly processed by Wav2Vec 2.0, so we do not need to use ASR to transcribe in Figure 4.

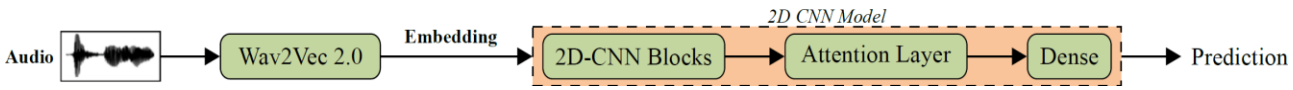

Fig. 4: Extracting linguistic features using Wav2Vec 2.0.

5.2. Experimental Setup

In the former experiment, the hyper-parameters are the same as the ones defined in the acoustic-based experiment. In the latter experiment, we use *Hugging Face* to implement the Wav2Vec 2.0 base model (Wav2Vec2-base-960h). This model is pre-trained and fine-tuned on 960 hours of LibriSpeech on 16kHz sampled speech audio. We also use Hugging Face to develop the BERT base model (BERT-base-uncased).

5.3. Experimental Result and Analysis

As shown in Table 3, SVM with BERT embedding leads to the best performance of 67% as the baseline result. The result confirms that BERT embedding produces a positive impact. However, 1D-CNN with BERT embedding cannot achieve the same performance. It is possible that embedding is produced by a DNN whilst the 1D-CNN is also a deep network, which potentially makes the model overfit. We figure out that 2D-CNN with Wav2Vec 2.0 embedding also performs terribly, suggesting that character-level embedding fails to capture a valuable representation of speech's semantics.

Table 3: Comparing results of models with linguistic features

Models	**Features**	**Mean-F1**
SVM	**BERT embedding**	**0.67**
2D-CNN with CBAM	Wav2Vec2.0 embedding	0.51
1D-CNN	BERT embedding	0.61
SVM (baseline)	**EVAL+ FREQ**	**0.67**

6. Discussion and Conclusion

In this paper, we present our work in detecting the cognitive decline progression in two manners. One is an acoustic-based approach, which is realized by two models; while another is a linguistic-based approach, which is implemented by three models. In general, the experimental results show that the latter (a mean-F1 of 67% which remains equal to the baseline) performs better than the former (a mean-F1 of 63% compared to the baseline of 61%).

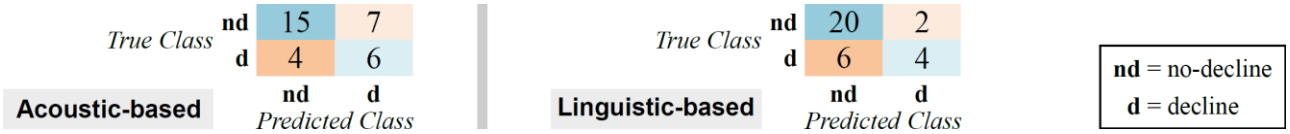

Fig. 5: Analysis of the experimental results.

We analyse the underlying rationales and depict our conclusion in Figure 5. As a matter of fact, the acoustic-based approach not only leads to stable prediction results but also makes more correct predictions in the declining class. Contrariwise, the linguistic-based approach excels at prediction in the no-decline class. As a natural consequence, we believe that making use of both features in conjunction can result in a much better model. In the future, we will research feature fusion and model fusion. As for the acoustic-based approach, we will consider more usage of pre-trained models to extract more helpful acoustic features. As for the linguistic-based approach, we will concentrate on how to transcribe audio recordings more correctly.

7. Acknowledgements

This work has been supported by the National Key R&D Program of China under Grant 2019YFE0190500, the Fundamental Research Funds for the Central Universities of Ministry of Education of China (Grant No.2232021D-22), and the Initial Research Funds for Young Teachers of Donghua University.

8. References

[1] J. Rasmussen and H. Langerman, "Alzheimer's disease - why we need early diagnosis," *Degenerative neurological and neuromuscular disease*, vol. 9, p. 123, 2019.

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

[3] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational German." in *Interspeech*, 2016, pp. 1938–1942.

[4] 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.

[5] 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, pp. 12 449–12 460, 2020.

[6] 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.

[7] 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.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.