Exploring MMSE Score Prediction Model Based on Spontaneous Speech

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

Jieyuan Zheng
School of Computer Science and Technology
Donghua University
Shanghai, China
zjy123h@163.com

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

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

***Abstract*—The Mini Mental State Examination, referred to as MMSE, is a screening tool for cognitive dysfunction in the elderly, and it is also one of the most influential screening tools for cognitive impairment. It is usually managed by a well-trained doctor, but this is time-consuming and expensive. An effective method is to detect whether cognitive function has declined through the conversation between them. From the perspective of acoustics and linguistics, using 108 subjects provided by the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) 2020 Challenge, using speech to predict the MMSE score, the acoustic Root Mean Squared Error (RMSE) is 5.49. The RMSE in linguistics is 4.51. Integrating the acoustic model and the linguistic model, and assigning different weight ratios to their final predicted scores, the RMSE is 4.18.**

***Index Terms*—Alzheimer's disease, acoustic features, linguistic features, MMSE**

I. Introduction

Alzheimer's Disease (AD), also known as Alzheimer's, is a neurodegenerative disease. According to epidemiological studies, the incidence of AD increases with age, about 5% of people over 65 years old, and up to 20% of people over 85 years old. According to statistics from Western countries [1], it is estimated that between 2000 and 2050, the population over 65 will triple, which will undoubtedly greatly increase the burden on families and the country. Because it is an irreversible disease, drug treatment may temporarily change the symptoms of the disease, but it cannot reverse its progress. For these reasons, there is an increasing need for this additional, non-invasive detection tool to enable preliminary identification of AD at an early stage.

At present, the more popular detection methods are to use Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), but this is undoubtedly more expensive. In the process of cognitive decline, the appearance of language barriers [2] is an important sign, which includes naming [2], difficulty in finding words, repetition, and improper use of pronouns [3]. This makes it possible to use speech to evaluate the AD process.

Cognitive assessments are often used for clinical validation, such as the Mini-Mental Status Examination (MMSE) [4]. Although simple to administer MMSE, it is burdensome for subjects and may also be influenced by various demographic factors [5]. Preliminary evidence [6] shows that automated methods can predict MMSE scores from open communication. Based on this, the main contributions of this paper are as follows. First, uses opensmile-3.0 to extract acoustic features, including ComParE16, emobase, eGeMAPS and Is09-13, and put the extracted features into the acoustic model. Second, there are two types of linguistic features. The first is to use BERT [7] to extract sentence vectors; the second is to use n-grams to vectorize text, combine psycholinguistic features, and put them into machine learning models. Third, comprehensively consider the acoustic model and the linguistic model, and fuse the models at the decision-making level.

II. Related work

In recent years, people have paid more and more attention to speech and language disorders in AD. However, most of the work is focused on dementia classification tasks [8, 9], rather than more detailed prediction of MMSE scores [10]. Aparna Balagopalan [11] demonstrated the use of domain knowledge-based methods to extract linguistic features from text and acoustic features from corresponding audio files, and combine two regression models, namely linear model and ridge regression model. The RMSE obtained is 4.56. Morteza Rohanian [12] and others used the LSTM with gating multi-modal fusion model, combined with multi-modal features, and the RMSE obtained was 4.54. Utkarsh Sarawgi [13] used transfer learning and ensemble models and got an RMSE of 4.60.

Although the RMSE of these papers is lower than the baseline, there is also a problem, that is, the impact of acoustic and linguistic models on the final results is not fully considered. Especially after the prediction results of the two models are obtained, the respective influences on the final results are comprehensively considered, and different weights are given respectively when the decision-making layer is fused.

III. Dataset and Features

A. Overview of the Dataset

The data set of this paper comes from the ADReSS Challenge [6], the subjects provided are theft pictures, which are provided by the Boston Diagnostic Aphasia Exam [14, 15].

DOI reference number: 10.18293/SEKE2022-062

During the recording process, the subject asked to describe the content in the picture, there is no time limit (the interviewer may stimulate the subject to add details). The provided .cha file is a manual transcription of the audio, using the CHAT encoding system [16], which contains non-verbal clues such as adding false starts, pauses, discourse markers for word repetition, and incomplete sentences. For the ADReSS Challenge, the original speech is also divided into standardized segments with a maximum length of ten seconds.

The ADReSS challenge data set is a balanced subset consisting of 156 subjects. Each subject provides a speech. Between AD and non-AD, age and gender are evenly distributed. The following two tables (TABLE I and TABLE II) respectively show the basic situation of the training set and test set.

TABLE I
ADReSS Training Set: Basic characteristics of the patients in each group (M=male and F=female)

	AD			Non-AD		
Age	**M**	**F**	**MMSE**	**M**	**F**	**MMSE**
[50,55)	1	0	30.0	1	0	29.0
[55,60)	5	4	16.3	5	4	29.0
[60,65)	3	6	18.3	3	6	29.3
[65,70)	6	10	16.9	6	10	29.1
[70,75)	6	8	15.8	6	8	29.1
[75,80)	3	2	17.2	3	2	28.8
Total	24	30	17.0	24	30	29.1

TABLE II
Characteristics of the ADReSS test set

	AD			Non-AD		
Age	**M**	**F**	**MMSE**	**M**	**F**	**MMSE**
[50,55)	1	0	23.0	1	0	28.0
[55,60)	2	2	18.7	2	2	28.5
[60,65)	1	3	14.7	1	3	28.7
[65,70)	3	4	23.2	3	4	29.4
[70,75)	3	3	17.3	3	3	28.0
[75,80)	1	1	21.5	1	1	30.0
Total	11	13	19.5	11	13	28.8

B. Acoustic Features

The ComParE16 feature set [17] is extracted using opensmile-3.0. The feature set contains 6373 static features, which are obtained by calculating various functions on LLD (low-level descriptors, LLD). LLD includes logarithmic harmonic noise ratio, voice quality characteristics, F0 Viterbi smoothing, spectral harmonics and psychoacoustic spectral sharpness. This feature set encodes human speech and has been used as an important non-invasive marker for AD detection. We remove the mean and normalize the variance of the obtained feature set. Standard deviation standardization makes the processed data conform to the standard normal distribution, that is, the mean is 0 and the standard deviation is 1. The transformation function is as follows:

$$X^* = \frac{\chi - \mu}{\sigma} \tag{1}$$

Where μ is the mean of all sample data, and σ is the standard deviation of all sample data.

In addition to ComParE16, features of emobase, eGeMAPS, and Is09-13 are also extracted for comparison experiments.

C. Linguistic Features

The ADReSS data set provides a corresponding .cha file for each subject, which contains the conversation between the interviewer and the subject (beginning with *INV and *PAR, respectively). First, extract the subject's speech fragments according to certain rules, and then use TF-IDF (term frequency—inverse document frequency) for the text. The calculation formula is as follows:

$$TFIDF = TF \times \frac{1}{DF} \tag{2}$$

Where TF is the term frequency in the text, and DF is the number of documents containing the current term. In addition to considering the frequency of a vocabulary in the text, it also pays attention to the number of all texts that contain this vocabulary. This can reduce the impact of high-frequency meaningless vocabulary and dig out more meaningful features.

For psycholinguistic features, four classic psycholinguistic attributes (age of acquisition, concreteness, familiarity, and imageability) and emotion scores are considered. They are obtained from the Medical Research Council (MRC) psychology database and Natural Language Toolkit (NLTK) respectively.

Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT), this model mainly uses the Encoder structure of Transformer, but the model structure is deeper than Transformer. The Transformer Encoder contains 6 Encoder blocks, the BERT-base model contains 12 Encoder blocks, and the BERT-large model contains 24 Encoder blocks.

Through the pre-training model BERT, the text of each subject is converted into a 768-dimensional sentence vector, and the obtained sentence vector is normalized.

IV. Experiment

A. Acoustic Model

Multilayer Perceptron (MLP) is a neural network with a forward structure that maps a set of input vectors to a set of output vectors. MLP can be regarded as a directed graph, composed of multiple node layers, and each layer is fully connected to the next layer. Except for the input node, each node is a neuron with a nonlinear activation function. MLP is the promotion of traditional perceptrons, which overcomes the weakness that traditional perceptrons cannot recognize linearly inseparable data.

For the acoustic model, as shown in Fig. 1, this article uses opensmile-3.0 to extract ComParE16, emobase, eGeMAPS and is09-13 features, and then put them into MLP, which contains five fully connected layers, and uses L1 regularization in the layer. Add the Dropout layer to the second layer and the penultimate layer, randomly remove some neurons in the network, thereby reducing the dependence on the weight of w, so as to reduce the effect of fitting. The activation function

sigmoid is added to the fully connected layer, and the ReLu activation function (max_value = 30) is added to the output layer to predict the MMSE score.

During the training process, we set epsilon to 1e-07, learning-rate to 0.01, batch_size to 16, epochs to 2000, and loss and metrics to use Mean Square Error (MSE).

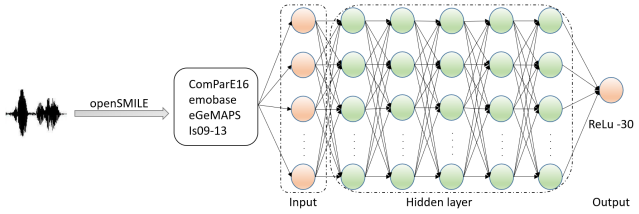

Fig. 1. Our proposed for acoustic model.

B. Linguistic Model

For the linguistic model, as shown in Fig. 2, there are two types. The first is to use the pre-trained BERT model to extract sentence vectors and standardize the obtained features. Taking into account the large difference in MMSE scores, the corresponding score of each subject is divided by 30, and then standardized, combined with the machine learning model (due to the low dimensionality of the feature, machine learning is used), better results can be obtained on ridge regression. During the training process using the ridge regression model, we set alphas=numpy.linspace(1,0.05), store_cv_values=True. The second is to use lexical features to combine emotional factors. First, use TfidVectorizer to extract syntactic features, and then obtain emotional scores from NLTK (Natural Language Toolkit). The obtained features are selected using random forest regression algorithm. The processing of the MMSE score is the same as above, and finally combined with the SVR model to get a better result. In the process of using SVR training, we set the kernel to poly, c to 100, gamma to scale, degree to 3, epsilon to 0.01, and coef0 to 1.

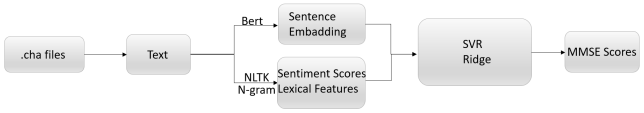

Fig. 2. Our proposed for linguistic model.

C. Fusion of Acoustic and Linguistic Model

Considering the acoustic model and the linguistic model, the results obtained by the acoustic model and the linguistic model are combined. Considering that the effect of the linguistic model is better than that of the acoustics, the weight of the acoustic model is appropriately reduced. In the experiment, the weights were evenly distributed first, and the results obtained by the linguistic model and the acoustic model were multiplied by a weight of 0.5, and then according to the weight of 0.1, the weight of the results obtained by the linguistic model was increased, and the weight of the results obtained by the acoustic model was decreased.After many experiments, it is found that multiplying the result obtained by the acoustic model by a weight of 0.3 and the result obtained by the linguistic model by 0.7, the optimal result of the combination of the two models can be obtained.

D. Experimental Environment

The experiment proceeded from three different perspectives, first using MLP to process the acoustic features, then using machine learning to process the linguistic features, and finally integrating the results obtained from the acoustic and linguistic models, and comparing and analyzing with the baseline.

The environment used in this paper is python3.6, the deep learning framework is tensorflow-based keras framework, the machine learning library is scikit-learn, and the operating system used is windows 10. The experiment set up a five-fold crossover, and used the trained model to predict the test set.

V. Results and Analysis

In order to effectively evaluate the features extracted in this article and the effectiveness of the models adopted, RMSE is proposed as an evaluation index.

Root Mean Squard Error (RMSE) is the square root of the ratio of the square of the deviation between the predicted value and the true value to the number m of the test set. It is used to measure the deviation between the predicted value and the true value. The smaller the value, the better the prediction effect of the model. The specific formula is as follows:

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2} \tag{3}$$

Among them, y_i represents the predicted value, and $\hat{y_i}$ represents the actual value.

In the baseline experiment, it has been proved that the acoustic and linguistic features in spontaneous speech have a certain correlation with the detection of cognitive impairment, and the relevant results have been provided. For the test set, the RMSE of acoustic features is 6.14, and the RMSE of linguistics is 5.21.

This experiment sets up five-fold cross-validation. TABLE III shows the results of acoustic features on the training set and test set. It can be seen that using ComParE16+MLP has the best effect on the test set. The RMSE is 5.49, which is 10% lower than the acoustic baseline. TABLE IV shows the results of linguistic features on the training set and test set. It can be seen that using Lexical+sentiment+SVR performs best on the test set, with an RMSE of 4.51, which is 13% lower than the linguistic baseline. TABLE V shows that the results of acoustic model and linguistic model are assigned weights of 0.3 and 0.7 respectively. It can be seen that the combination of acoustic feature ComParE16 and linguistic feature Lexical+sentiment results in the best result, and the RMSE is 4.18. It is 31.9% lower than the acoustic baseline and 19.8% lower than the linguistic baseline.

TABLE III
THE RESULTS OF ACOUSTIC MODEL ON THE TRAINING SET AND TEST SET

Features	Model	RMSE on train set	RMSE on test set
baseline	-	7.28	6.14
ComParE16	MLP	5.46	**5.49**
emobase	MLP	**4.73**	5.82
eGeMAPS	MLP	5.06	5.96
Is09-13	MLP	5.08	6.28

TABLE IV
THE RESULTS OF LINGUISTIC MODEL ON THE TRAINING SET AND TEST SET

Features	Model	RMSE on train set	RMSE on test set
baseline	-	4.38	5.21
Bert embedding	ridge	4.86	5.37
Lexical+sentiment	SVR	**4.01**	**4.51**

TABLE V
THE RESULTS OF FUSION MODEL ON THE TEST SET

Features	Model	RMSE on test set
ComParE16+ Lexical+sentiment	**MLP+SVR**	**4.18**
ComParE16+Bert embedding	MLP+ridge	4.91
eGeMAPS+ Lexical+sentiment	MLP+SVR	4.20
eGeMAPS +Bert embedding	MLP+ridge	4.93
emobase+ Lexical+sentiment	MLP+SVR	4.40
emobase+Bert embedding	MLP+ridge	4.87
Is09-13+ Lexical+sentiment	MLP+SVR	4.32
Is09-13+Bert embedding	MLP+ridge	4.93

VI. CONCLUSIONS

For the use of speech to predict MMSE scores, there are relatively few research papers in this area. The paper starts from acoustics and linguistics, combined with MLP and machine learning models, and finds that linguistics can provide more information such as pauses, word repetitions, incomplete sentences and emotions, which provide us with strong evidence for predicting MMSE scores.

In the follow-up work, on the one hand, we can also start with linguistics to discover more meaningful features. On the other hand, for acoustic features, we can extract spectrograms such as Spectrongram (Spec), Melspectrongram (Melspec), Mel-Frequency Cepstral Coefficients (MFCC), etc., and combine convolutional neural networks (CNN) to learn two-dimensional features. For speech that cannot be transcribed, the pre-training model wav2vec2.0 can also be used to encode the speech information and modify the downstream output terminal to obtain the expected result.

REFERENCES

[1] OMS, "Es mental health action plan 2013 - 2020," 2013.

[2] J. Reilly, J. Troche, and M. Grossman, *Language Processing in Dementia*. The Handbook of Alzheimer's Disease and Other Dementias, 2011.

[3] D. N. Ripich and B. Y. Terrell, "Patterns of discourse cohesion and coherence in alzheimer's disease." *J Speech Hear Disord*, vol. 53, no. 1, pp. 8–15, 1988.

[4] J. R. Cockrell and M. F. Folstein, "Mini-mental state examination (mmse)." *australian journal of physiotherapy*, vol. 51, no. 3, pp. 689–92, 2005.

[5] R. N. Jones and J. J. Gallo, "Education and sex differences in the mini-mental state examination," *Journals of Gerontology*, no. 6, p. 6.

[6] 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.

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

[8] K. C. Fraser, J. A. Meltzer, F. Rudzicz, and P. Garrard, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimerš Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[9] A. K?Nig, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, and P. H. a. Robert, "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, p. 112–124, 2015.

[10] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias," in *Slpat: Workshop on Speech Language Processing for Assistive Technologies*, 2015.

[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," 2020.

[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," 2021.

[13] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity," 2020.

[14] ASHAWeb, "Boston diagnostic aphasia examination-third edition (bdae-3)," *Asha*, 2000.

[15] F. Boersma and J. A. Eefsting, "The natural history of alzheimer's disease," *Journal of the American Geriatrics Society*, vol. 44, no. 6, pp. 734–734, 1996.

[16] J. W. Oller and B. Macwhinney, "The childes project: Tools for analyzing talk, 3rd edition, vol 1, transcription format and programs," *Modern Language Journal*, vol. 86, no. 2, pp. 289–290, 2002.

[17] 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.