Depressive Tendency Recognition by Fusing Speech and Text Features: A Comparative Analysis

Yimin He, Xiaoyong Lu*, Jingyi Yuan, Tao Pan, and Yafan Wang
Northwest Normal University Lanzhou, Gansu, China
E-mail: luxy@nwnu.edu.cn

***Abstract*— Depression will be accompanied by long-term depression, loss of interest, excessive guilt, and other states, seriously affecting people's physical and mental health, causing certain harm to the individual family and society. Depressed people speak slowly, single tone, and the pause time is longer. At the same time, the expression of emotion is often accompanied by many negative words. Therefore, the combination of speech and text information using Gated Recurrent Neural Network for depression prediction, this fusion method from the signal layer to the language layer to analyze the data. It is more comprehensive than only use a single speech feature or text feature model. At the same time, the comparative analysis of different speech types, three kinds of emotional stimulus corpus, and gender was carried out. The multimodal system is more effective than a single modality, and speech type, emotional stimulus, and gender have certain effects on depressive recognition.**

Keywords—Depression recognition, Deep learning, GRU, Multimodal

I. INTRODUCTION

Depression is one of the most common mental disorders. The incidence rate, the duration, and the danger of the disease are the biggest challenges in the diagnosis of mental illness [1]. Therefore, mental health has been paid more and more attention by people from all walks of life. However, the diagnosis of depression is mainly based on doctors' judgment, which may bring a series of subjective biases to the diagnosis process. The degree of cooperation of patients and the professional level and experience of doctors will have a certain impact on the accuracy of the diagnosis of depression [2]. In recent years, with the increase of work pressure, more and more people will lack interest in daily activities, lose funny, lose weight or gain significantly, lose sleep or sleep too much, lack energy, feel worthless or excessive guilt, and repeatedly think of death or suicide. With the increase in depression, how to diagnose depression quickly and accurately is a major problem for medical staff. Therefore, early detection, early intervention, and early change are of great significance to improve the well-being of patients with depression and the whole family. A reliable and effective assessment of individuals with depression is essential for their treatment and recovery. The current evaluation methods are limited to subjective evaluation of patients' self-report and clinical interviews [3]. It does not take into account the observable behavioral indicators, which can better inform the occurrence and severity of depression. In recent years, machine learning has been widely used in the medical field because of its powerful data processing and mining ability, and gradually applied to the early prediction, recognition, and auxiliary diagnosis of depression [4].

II. RELATED WORK

With the increase in depression, the burden of doctors' time and energy is becoming heavier. It is almost impossible to identify patients with depression in a large range through complex psychological measurement tools. Therefore, more and more researchers pay attention to the method of automatic detection of mental illness through speech. In an automation system, speech is a very effective evaluation index [5]. Besides acoustic features, text features are also a way of information, and their syntactic and semantic features can be used as effective indicators to detect depression [8]. Zhao Xiaoli [9] analyzed the common features of microblog emotion and behavior of patients with depression, combined with knowledge base and experimental corpus, constructed a depression domain dictionary by using two semantic similarity algorithms and extracted dictionary features, semantic features, and extended features closely related to depression. Shi Zhiwei et al. [10] Constructed a model to identify depression tendency based on microblog text data, and defined a depression index to measure the degree of individual depression tendency in a period of time.

It is not thorough to predict whether the patient is depressed by using speech or text features alone. For example, using a happy tone and an angry tone to say the same sentence may have opposite meanings. Thus, understanding text content or speech information alone is not enough to explain that individuals want to express complete semantics. So increasingly studies have found that text and speech can be fused to predict the degree of depression. And many ways to fuse speech and text., Fraser [11] and others have established a depression detection system using abundant text features and acoustic features. Morales and Levitan [12] fused speech and text in the feature layer, compared the fused and single speech features or text features found that multimodal system led to the best performance. Meanwhile, use the ASR system to transcribe speech automatically. And the text features transcribed by ASR are also effective for the detection of depression. But there is information redundancy. Meanwhile, it also needs to solve the problem of time synchronization between different data sources and the problem of feature incompatibility [13]. If multiple features are fused at the decision level, this fusion method can solve the shortcomings of the feature level fusion, but it can not make full use of the

*Corresponding author.

correlation between multiple modal data. Therefore, this paper chooses to fuse speech features and text features at the model level, which not only deviates from the basic paradigm of feature level and decision level fusion, but also reduces the problem of information source synchronization, and fully utilize the correlation between each modal data.

III. Collection of Depression Corpus

A. Subjects

This study recruited 54 university students with depression tendency, including 27 males and 27 females. And 54 students in health including 27 males and 27 females, participated in the experiment as the control group. All subjects have to go through the Putonghua test to ensure that their Putonghua standard is clear. Before the experiment, the subjects were tested with BDI-II and PHQ-9 scales [14] to exclude those who did not meet the requirements. Before the formal experiment, all subjects signed the informed consent and voluntarily participated in the experimental study.

TableI. Basic information of the subjects

Group	Health		Depression	
gender	Female	male	Female	male
number	27	27	27	27
age	19.78 ± 1.485		19.52 ± 1.682	
	19.80 ± 1.669	19.77 ± 1.305	19.57 ± 1.716	19.47 ± 1.676
BDI-II	4.20 ± 3.210	4.67 ± 4.444	16.73 ± 3.107	17.87 ± 4.343
language	Chinese			

The details of the subjects are shown in Table I. The independent sample t test showed that there was no significant difference in age between depression group and healthy group ($t = 0.920$, $P = 0.359 > 0.05$), there was no significant difference in age between healthy women and depression women ($t = 0.534$, $P = 0.595 > 0.05$), and there was no significant difference in age between healthy men and depression men ($t = 0.774$, $P = 0.442 > 0.05$).

B. Task Design

The content of the depression corpus includes picture description and question-answering. The pictures in the picture description corpus are modified from the Chinese facial expression picture system [15]. The picture described is three female images, as shown in Fig.1. The subjects described it according to three pictures on the computer screen. For example, what is the female expression you see in this picture? What is the reason behind this expression? Or what do you think of when you see the female expression in the picture? The question-answering corpus is used to answer questions according to the questions displayed on the computer screen. There are three kinds of questions: neutral questions, positive questions, and negative questions [16]. For example, how much sleep have you had recently? This is a neutral issue. Are you optimistic recently? This is a positive question. Do you often feel disappointed recently? This is a negative problem. The interview consists of 30 questions.

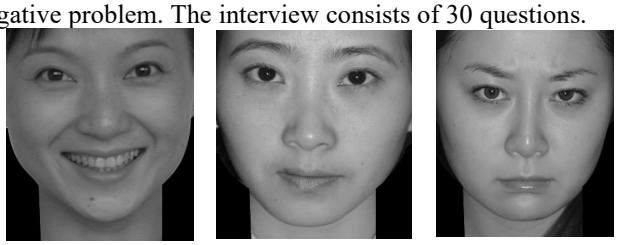


Fig.1. Picture description task

C. Data Collection

There are about 9 minutes for each subject to complete the two recording tasks of picture description and question-answering in a quiet recording room, and the total length of interviews is 972 minutes. Professional recording equipment Roland R-26 with a sampling rate of 44.1KHz, 16-bit quantization, and a two-channel format as a .wav file. is applied in this experiment. And the data are collected and converted into a single-channel .wav format.

IV. Experimental Methods

A. Experiment Design

Firstly, the corpus is preprocessed, and then the data set is divided into training, verification, and test data sets according to the ratio of 8:1:1. Finally, through model training, validation, and testing, the data are classified into high depression tendency and health. The detailed process is shown in Fig. 3.

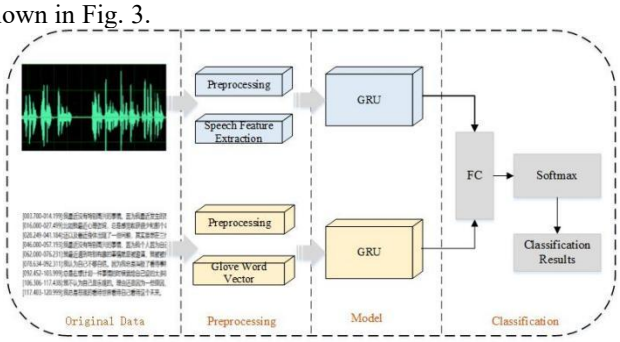


Fig.2. Overall block diagram of the experiment

B. Experiment Setup

For the experimental equipment conditions and the network structure, the relevant parameters are shown in TableII.

TableII. Experimental environment and related configuration

Hardware / Software		Model / Parameter	
Operating system	CentOS Linux	Word vector dimension	300
Word vector Training tool	Gensim	Learning rate	0.001
Word segmentation tool	Jieba	Batch size	128
Programming tools	Python	Speech encoder step size	750
Deep learning framework	Tensorflow	Text encoder step size	128

C. Preprocessing

In this experiment, the speech signal was preprocessed to cut out the coughing and indistinct sound. MFCC and prosody features were extracted from speech signals by the openSMILE2.3.0 toolkit [17]. The MFCC feature set contains a total of 39 features, which includes 12 MFCC parameters (1-12) from the 26 Mel frequency bands and log-energy parameters, 13 delta, and 13 acceleration coefficients. The frame size is set to 25 ms at a rate of 10 ms with the Hamming function. According to the length of each wave file, the sequential step of the MFCC features is varied. The prosody features are composed of 35 features, which include the F0 frequency, the voicing probability, and the loudness contours.

V. Experimental Model

Because the consideration of single-mode is too single and one-sided, there are some shortcomings. Therefore, a model combining speech and text is adopted. In this experiment, multiple modes, such as MFCC, prosody, and transcripts, are considered, which include sequential speech information, statistical speech information, and text information respectively.

For the speech signal, once the MFCC feature is extracted from the speech signal, a subset of the sequence feature is input to the GRU, which leads to the formation of the time sequence pattern of the internal hidden state h_t to the model. This internal hidden state will be updated every time data X_t is input and the hidden state of the previous step h_{t-1} is as follows:

$$h_t = f_\theta\left(h_{t-1}, A_t\right) \tag{1}$$

Where f_θ is the RNN function with the weight parameter θ, h_t represents the hidden state of the t-th time step, and X_t represents the t-th MFCC feature in $X = \{x_1 : t_\alpha\}$. After the speech signal X is encoded by the RNN, the last hidden state h_{t_α} of the RNN is considered as a representative vector containing all sequential speech data. Then, this vector is connected with another prosody feature vector P to generate a more informative vector representation of the signal, $E = concat\left\{h_{t_\alpha}, P\right\}$. Using the openSMILE2.3.0 toolkit, $X_t \in R^{39}$ and $P \in R^{35}$ extract MFCC and prosody features. The last hidden state of the speech RNN is concatenated with the prosody feature to form the final vector representation E, and then the vector is passed through the fully connected neural network layer to form the speech coding vector A.

The corresponding transcription text is obtained through speech recognition. In order to use the text information, Jieba [19] is used to segment the text to remove stop words and low-frequency words. Tokenized words are expressed, and the words are converted into corresponding 300-dimensional word vectors through the Glove [20] model, which contains additional contextual meaning between words. The marked sequence vector is fed to the text recursive encoder, and the text RNN uses equation (1) to encode the word sequence of the transcript. In this case, X_t is the t-th token from the text input. The final hidden state of the text RNN also forms a text encoding vector T through another neural network layer. Finally, the softmax function is applied to the cascade of vector A and vector T to predict depression/health classification. The predicted probability distribution of the target class is as follows:

$$\begin{aligned} A = g_\theta(e), T = g'_\theta(h_{last}) \\ \hat{y}_i = soft\max\left(concat(A,T)^T M + b\right) \\ L = -\log\prod_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \end{aligned} \tag{2}$$

In equation (2), g_θ and g'_θ are feedforward neural networks with weighted parameters θ, A and T are the final encoding vectors from speech RNN and text RNN, respectively. $M \in R^{d\times c}$ and bias b is the learned model parameter. C is the total number of categories, and N is the total number of samples used in training.

VI. Experimental Results and Analysis

This experiment aims at two speech types: picture descriptions and question-answering corpus. Distinguish between males and females. Three kinds of emotional stimulation: positive, neutral, and negative.

A. The Overall Experimental Results

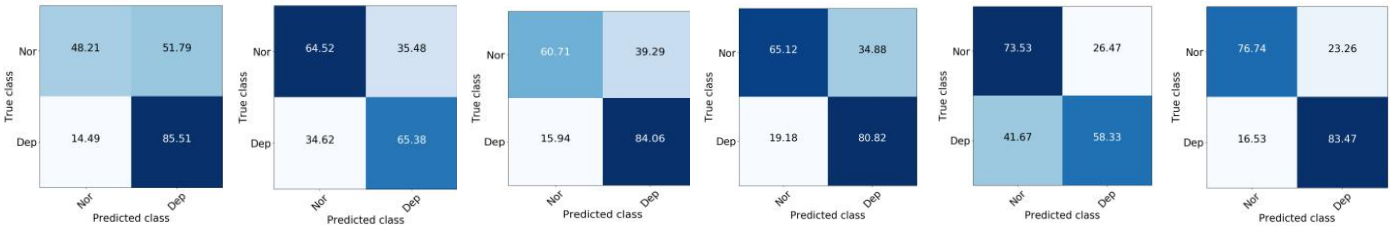
Experiments are carried out on two speech patterns of the corpus, and the specific experimental results are shown in Table III :

TableIII. The overall experimental results of two speech types

Model	Accuracy		Precision		Recall		F1-score	
Speech types	picture descriptions	question-answering	picture descriptions	question-answering	picture descriptions	question-answering	picture descriptions	question-answering
Speech	0.669	0.730	0.623	0.699	0.855	0.808	0.721	0.750
Text	0.650	0.659	0.648	0.688	0.654	0.583	0.651	0.631
Speech+Text	0.724	0.801	0.681	0.782	0.841	0.835	0.752	0.808

According to the experimental results of two speech types corpus in Table III. The accuracy of speech + text reached 0.724 and0.801, which was significantly higher than the accuracy of speech and text.Meanwhile, the precision and F1 scores are significantly higher than speech and text. However, the recall of the speech+text model is slightly lower than that

of the speech model, because the recognition result of the speech model for the depression category is higher than the recognition result of the speech+text model for the depression category.



Speech Text Speech+Text

Fig.3. Confusion matrix of picture description

Speech Text Speech+Text

Fig.4. Confusion matrix of question-answering

It can be seen from the confusion matrix in Figure 3 that the speech model has the best recognition effect on depression, reaching 85.51%. The recognition result of the text model for the health category reaches 64.52%. However, the speech + text model has the best overall recognition effect on depression and health categories.

As can be seen from the confusion matrix in Figure 4, the recognition results of the speech + text model for the depression category are 83.47%, and the recognition results for the health category are 76.74%. It is higher than the recognition results of speech and text.

Through the analysis of the overall experimental results of picture description and question-answering corpus, it can be found that the recognition result of question-answering corpus is higher than that of picture description corpus, which indicates that speech style also has a certain influence on the recognition result. According to the three models of speech, text, and speech + text, we can see that the recognition effect of combining speech and text is higher than that of a single speech or text.

B. Results of different gender experiments

The depression corpus used in this experiment distinguishes the gender of the subjects and balances the ratio of male to female. At the same time, some studies have shown that gender also has a certain impact on the identification of depression. Therefore, we experimented according to the gender types, and the results are as follow

TableIV.The gender types experimental results of picture description

Model	Accuracy		Precision		Recall		F1-score	
Gender	female	male	female	male	female	male	female	male
Speech	0.674	0.621	0.634	0.510	0.814	0.632	0.713	0.564
Text	0.664	0.619	0.645	0.613	0.727	0.647	0.684	0.630
Speech+Text	0.746	0.702	0.699	0.660	0.862	0.833	0.772	0.736

By comparing the experimental results described in the pictures in Table IV, it can be found that in terms of accuracy, precision, recall, and F1-score, the recognition results of females are significantly higher than those of males and slightly higher than the overall experimental results.

TableV.The gender types experimental results of question-answering

Model	Accuracy		Precision		Recall		F1-score	
Gender	female	male	female	male	female	male	female	male
Speech	0.780	0.699	0.854	0.657	0.675	0.833	0.754	0.735
Text	0.685	0.621	0.722	0.650	0.600	0.524	0.655	0.580
Speech+Text	0.818	0.786	0.792	0.760	0.864	0.835	0.826	0.796

By comparing the results of the question-answering groups in Table V, it can be found that the results of females recognition are significantly higher than that of males in accuracy, precision, recall, and F1 score, and slightly higher than that of the overall experiment.

Through the experiment of gender discrimination on picture description and question-answering corpus, it is found that the recognition result of females is higher than that of males. The reason may be that females are more inclined to show their inner feelings, while males are more inclined to hide their inner feelings. It also fully confirmed the influence of gender on the recognition of depression.

C. Results of different emotional stimulation experiments

The depression corpus used in this experiment distinguishes positive, neutral, and negative emotions, so the experiment is carried out on the corpus of different emotional stimuli, and the experimental results are as follows:

TableIV.Results of three emotional stimulation experiment

Model	Accuracy			Precision			Recall			F1-score		
Emotional Stimulation	positive	neutral	negative	positive	neutral	negative	positive	neutral	negative	positive	neutral	negative
Speech	0.664	0.704	0.725	0.657	0.694	0.702	0.686	0.729	0.783	0.671	0.711	0.740
Text	0.630	0.661	0.687	0.624	0.650	0.672	0.657	0.700	0.731	0.640	0.674	0.780
Speech+Text	0.692	0.744	0.793	0.675	0.732	0.770	0.743	0.771	0.836	0.783	0.751	0.802

By comparing the experimental results of three emotional stimuli in Table VI, it is found that the recognition results of negative emotional stimuli are the highest in terms of accuracy, precision, recall, and F1 score, followed by the results of neutral emotional stimuli, while the recognition results of positive emotional stimuli are the worst. This is because the depressive people's response to the positive stimulus is weakened and the response to the negative stimulus is enhanced. When people with depression observe facial expression pictures and answer questions, they pay more attention to negative expression pictures. It is also found that the recognition result of the speech + text model is higher than that of the speech model and text model.

VII. CONCLUSIONS

In this paper, the features of speech and text are encoded by a double recursive coding model, and finally fused in the full connection layer. It uses both text data and speech signals, and the data show that the multimodal model can better predict depression tendency. It is also found that the recognition effect of question-answering corpus is better than that of picture description. The recognition result of women is higher than that of men. This may be because females are more willing to express their inner feelings than meles. Moreover, the recognition results of the negative emotional corpus are higher than those of neutral and positive emotional corpus, which may be because depressive people respond more strongly to negative stimuli than to positive ones.

ACKNOWLEDGMENT

This work was supported by the research fund from the National Science Foundation of China (NSFC) under grant No. 31660281, and No. 31860285. Additionally, part of this work was performed in the Natural Science Foundation Project of Gansu province (Grant No. 21JR7RA116).The authors also thank Dr. Dong Qiangli from Lanzhou University Second Hospital for his contribution in the data collection process.

REFERENCES

[1] M. Briley, Lépine, "The increasing burden of depression," Neur. Dise. and Text. New Zealand, vol. 7, pp. 3-7, May 2011.

[2] A. B. Zhou, X. Y. Lu, and W. Y. Wu, "A review of studies on the diagnosis of depression using speech," Jour. of Chin. Comp. Syst. Chinese, vol. 11, pp. 2619-2624, November 2017.

[3] P. Bech, "Handbook of Clinical Rating scales and assessment in psychiatry and mental health," Acta. Psyc. Scan. Denmark, Vol. 121, pp. 487-488, July 2010

[4] Q. M. Yuan, X. Wang, J. W. Shuai, H. Lin, and Y. P. Cao, "Research progress of depression based on artificial intelligence technology," Chine. Jour. of Clin.l Psyc. Chinese, vol. 28, pp.82-86, April, 2020.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and TF Quatieri, "A review of depression and suicide risk assessment using speech analysis," Spee. Comm. Netherlands, vol. 71, pp. 10-49, July, 2015.

[6] C. Howes, M. Purver, and R. Mccabe, "Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression," Acl Work. on Comp. Ling. & Clin. USA, pp. 7-16, June, 2014.

[7] X. L. Zhao, "Research on depression recognition based on microblog text and deep learning," Beij. Univ. of Tech. China, 2019.

[8] Z. W. Shi, J. B. Gao, W. W. Hu, and Z. Y. Liu, "Text based recognition model of depressive tendency," Appl. of comp. Syst. China, vol. 26, pp. 155-159, 2017.

[9] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in Alzheimer's disease through analysis of speech and language," Work. on Comp. Ling. & Clin. January, 2016.

[10] M. R. Morales, and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," IEEE, February, 2017.

[11] J. He, C. Q. Zhang, X. Z. Li, and D. H. Zhang, "Survey of Research on Multimodal Fusion Technology for Deep Learning," Comp. Engi. China, vol. 46, pp. 7-17, May, 2020.

[12] K. L. Smarr, & A. L. Keefer, "Measures of depression and depressive symptoms: beck depression inventory-ii (bdi-ii), center for epidemiologic studies depression scale (ces-d), geriatric depression scale (gds), hospital anxiety and depression scale (hads), and patient health questionnai," Arth. Care & Rese., vol. 63, pp. S454-S466. 2011.

[13] Y. X. Huang, & Y. J. Luo, "A revision of Chinese face expression image system," Chin. Jour. of ment. Heal., vol. 25(001), pp. 40-46, April, 2011.

[14] Y. Zkanca, C. Demiroglu, A. Besirli, & S. Celik, "Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features". Interspeech. pp. 2-6, September, 2018.

[15] F. Eyben, M. Wllmer, & B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor,: Acm Inte. Conf. on Mult., ACM. 2010.

[16] F. G. Shi, "Implementation of Chinese text corpus preprocessing module based on Jieba Chinese word segmentation," Comp. Scie. and Tech., vol. 16, pp. 254-257+263, May, 2020.

[17] J. Pennington, R. Socher, & C. Manning, "Glove: Global Vectors for Word Representation," Conf. on Empi. Meth. in Natu. Lang. Proc., January, 2014.