Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5565967, 8 pages
https://doi.org/10.1155/2021/5565967

Research Article

A Novel Smart Depression Recognition Method Using Human-Computer Interaction System

Lijun Xu[1], **Jianjun Hou**,[1] **and Jun Gao**[2]

[1]*Institute of Art and Design, Nanjing Institute of Technology, Nanjing, Jiangsu 211167, China*
[2]*Siemens Ltd., China Jiangsu Branch Co., Ltd., Nanjing, Jiangsu 211100, China*

Correspondence should be addressed to Lijun Xu; xulijun@njit.edu.cn

Received 11 April 2021; Revised 8 May 2021; Accepted 17 May 2021; Published 26 May 2021

Academic Editor: Shan Zhong

In recent years, depression not only makes patients suffer from psychological pain such as self-blame but also has a high disability mortality rate. Early detection and diagnosis of depression and timely treatment of patients with different levels can improve the cure rate. Because there are quite a few potential depression patients who are not aware of their illness, some even suspect that they are sick but are unwilling to go to the hospital. In response to this situation, this research designed an intelligent depression recognition human-computer interaction system. The main contributions of this research are (1) the use of an audio depression regression model (DR AudioNet) based on a convolutional neural network (CNN) and a long-short-term memory network (LSTM) to identify the prevalence of depression patients. And it uses a multiscale audio differential normalization (MADN) feature extraction algorithm. The MADN feature describes the characteristics of nonpersonalized speech, and two network models are designed based on the MADN features of two adjacent segments of audio. Comparative experiments show that the method is effective in identifying depression. (2) Based on the research conclusion of the previous step, a human-computer interaction system is designed. After the user inputs his own voice, the final recognition result is output through the recognition of the network model used in this research. Visual operation is more convenient for users and has a practical application value.

1. Introduction

Depression is a hidden mental illness, and it also involves mental health problems. Symptoms may be manifested in emotional and emotional disorders and physical discomfort. Depression not only harms patients [1, 2] but also brings a heavy economic burden to patients' families and society. Depression can lead to increased expenditure on medicines, psychotherapy, rehabilitation, and other aspects, and it can also make patients inefficient or even unable to work. The report in Reference [3] pointed out that in 2002, the United States suffered as much as 44 billion U.S. dollars in economic losses due to depression causing workers to be unable to work normally or reducing work efficiency. Although there is no relevant data in China, the huge economic losses caused by depression can still be seen from the above figures.

Although depression is hugely harmful, it is a disease that can be effectively treated and improved. In clinical practice, medication can promote the recovery of depression patients. In addition, psychotherapy and physical therapy can also achieve better results. However, a considerable part of the patient population has not been diagnosed in time. More than half of the people affected by depression at home and abroad have not received treatment. There are many reasons, such as insufficient precision in quantifying the prevalence of depression [4–6]. In addition, the diagnosis of depression is misdiagnosed. For example, depression and bipolar disorder are two relatively similar mood disorders. The depressive periods of depression and bipolar disorder are often difficult to distinguish [7]. Global Burden of Disease (GBD) conducted multiple epidemiological surveys for middle school students in 2016 [8]; the number of depressive symptoms in middle school students ranges from one-fifth to one-half.

The key to the treatment of depression is the preliminary diagnostic screening. If it is possible to quickly diagnose whether an individual is suffering from depression under rel-

atively safe conditions without too much privacy involved, it will greatly reduce the difficulty of clinical screening for depression and encourage patients to receive treatment as soon as possible. Voice is a noninvasive, clinically accessible information. At present, there have been a large number of studies on speech and depression [9, 10], which provides the possibility to explore speech as a tool for automated diagnosis of clinical depression. According to existing studies, the speech of patients with depression has the following characteristics: slower speaking rate, frequent pauses, long pauses [11, 12], reduced changes in voice characteristics [13], lack of circumflex and frustration, and dull voice [14]. Compared with normal individuals, individuals in the depression group have more pronounced breath sounds [15]. From the perspective of prosody characteristics, the changes in fundamental frequency (F0) of depression patients decrease, such as bandwidth, amplitude, and energy [16], which indicates that the changes in voice frequency of depression patients decrease. The spectral characteristics are also related to the degree of depression of the patient. Studies have found that the degree of change in the sound spectrum energy below 500 Hz and 500-1000 Hz is related to the severity of depression [17]. It can be seen that feature extraction of speech helps to better understand depression.

This research is based on speech data to recognize depression. In the early stage, the speech signal characteristics of the 2014AVEC dataset were extracted, and algorithmic modeling was used to identify the degree of depression of the subject. The model with relatively good experimental effect is selected to realize the human-computer interaction system in the later depression recognition. Finally, the experimental subject of research and application of speech-based depression recognition is completed. The work of this research includes the following:

(1) The MADN method is used for speech feature extraction. Since the extracted conventional speech features include the sample's own personalized speaking features, this personalized speaking feature will affect the training of the depression recognition model, resulting in poor generalization of the trained model. Therefore, this study selects the MADN method for feature extraction, and the features extracted based on this method have better adaptability

(2) The purpose of traditional depression recognition is to identify whether you have depression. The purpose of this study is not only to identify whether you have depression but also to determine the degree of depression. Only by identifying the specific degree of depression can the most appropriate treatment plan be given. The depression recognition model used in this study is DR AudioNet

(3) Design a human-computer interaction system where users can input audio data online to identify depression. The system can help users understand their own depression status and can also assist doctors in initial examinations

2. Speech-Based Depression Recognition

In order to identify depression more conveniently and quickly, this paper designs a set of human-computer interaction system. First, the system will collect voice data. Secondly, the sample set is preprocessed and feature extracted at the front end. Again, machine learning algorithms [18, 19] and deep learning algorithms [20, 21] are used to model the samples in the backend. Finally, the trained model is used to obtain the recognition result of the test sample, thereby obtaining the final depression recognition result. The speech-based depression recognition process is shown in Figure 1.

The detailed process of identifying depression is as follows:

(1) Perform preprocessing and feature extraction on the original data. The extracted features include Mel frequency cepstrum coefficients (MFCCs), zero-crossing rate, energy, and other speech features. These feature data will be used for model training

(2) The front end is used for feature data collection (back-end auxiliary data processing). The data processing models mainly include deep learning algorithms such as CNN and deep convolutional neural network (DCNN), and you can also choose a suitable machine learning algorithm

(3) The results identified by the back-end can be output to the visual display page according to the interaction requirements and displayed to doctors and patients. Doctors and patients have a preliminary understanding of the condition based on the output of the system

Although the research on the recognition of depression based on speech signals has achieved good research results at this stage, it is inevitable that there are still challenges in this research field.

Mainly reflected in the following aspects:

(1) Effectiveness of voice feature selection: the classification of depression and nondepression based on speech signals is based on the premise that the speech features with significant differences between the two can be extracted. Different people have different characteristics in terms of tone, loudness, and energy of voice signals. The key point is to be able to find the characteristics that distinguish depressed individuals from nondepressed individuals in order to achieve better recognition results. However, relevant research results show that this effective voice feature has not been found yet. The most effective speech features on different speech datasets may be different

(2) The accuracy of the results of depression recognition: due to different data characteristics, experimental algorithms, and hardware equipment conditions, recognition results with different accuracy will naturally be obtained. To improve the accuracy of the recognition results of depression by continuously improving

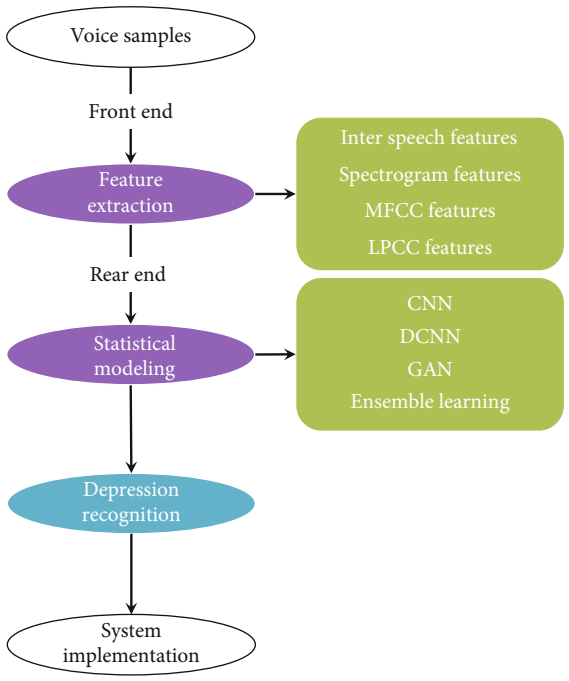

Figure 1: Depression recognition process.

the experimental conditions in terms of software and hardware is a problem to be solved in this research

(3) The authority of the depression speech dataset: at the present research stage, there is no unified standard dataset as the speech source material for depression recognition. Some of the datasets used for the research are the voices of the subjects collected by the experimental researchers themselves, and some use the international open source voice datasets. The voice file collection process is also affected by the field environment, the denoising effect of the equipment, and the immediate state of the subjects. For this reason, seeking a unified standard for the integration of speech data related to depression is one of the challenges facing this field

3. Voice-Based Depression Recognition Method Used

3.1. Depression Recognition Model Used. The network model used in this study integrates CNN and LSTM. The type of data processed is voice data, and the output is whether it is a depression patient. The deep learning model used is DR AudioNet, and its network structure is shown in Figure 2.

As can be seen from Figure 1, the DR AudioNet network mainly consists of 2 convolutional layers, 1 pooling layer, 2 LSTM layers, and 1 fully connected layer. The input of the network is voice feature data. The construction process of speech feature data is as follows: extract MFCCs, zero-crossing rate, energy, and other features for each frame of speech in each speech, and select 60 consecutive frames of speech characteristics in each speech to construct a two-dimensional matrix. The X axis of the matrix is time, and the Y axis is frequency information. The role of the convolutional layer is to extract the semantic information of Gao. The role of the pooling layer is to reduce the dimension of features. The two-dimensional matrix is converted into one-dimensional data after passing through the convolutional layer and the pooling layer. The LSTM layer is used to extract long-term dependency information. The fully connected layer is used to encode changes in speech on the X axis and give prediction results.

In order to make full use of speech information to improve the recognition rate, this paper uses an improved network model based on the DR AudioNet network. The improved network mainly contains 3 models, namely, M1, M2, and M3. These 3 models are all DR AudioNet networks. The execution process of the improved network is to first extract the feature V1 of the current speech segment. Feature V1 mainly includes MFCCs, short-term energy, short-term zero-crossing rate, and formant frequency. Input feature V1 into the M1 model. Considering that the personalized interference information mixed in each sample will affect the recognition effect of the model, the feature V2 of the previous segment of the current speech segment is used to train M2. Feature V2 is obtained using the MADN method. In order to further reduce the interference of the personalized information carried by the sample, the feature V3 of the last segment of the current speech segment is used to train M3. Feature V3 is obtained using the MADN method. The structure of the improved network is shown in Figure 3.

3.2. Multiple Feature Extraction

3.2.1. Mel Frequency Cepstrum Coefficients (MFCCs). MFCCs are one of the most widely used and basic voice features. Equation (1) gives the conversion relationship between ordinary frequencies and MFCCs.

$$f_{\text{mel}} = 2595 \log\left(1 + \frac{f_{\text{Hz}}}{700}\right), \tag{1}$$

where f_{mel} represents Mel frequency scale and f_{Hz} stands for normal frequency. Use P filters to calculate MFCCs. The center frequencies of these filters can be evenly arranged according to the Mel frequency, and the frequencies of the two bottom points of each filter are the center frequencies of the two adjacent filters. The output is $L(p), p = 1, 2, \cdots, P$ after filtering. Let $l(p)$ be the lower limit frequency, $c(p)$ be the center frequency, and $h(p)$ be the upper limit frequency of the p-th triangle filter. Then, the relationship between them is as follows:

$$c(p) = h(p-1) = l(p+1). \tag{2}$$

The output of the filter is transformed as follows to obtain MFCCs.

$$C_n = \sum_{p=1}^{P} \log L(p) \cos\left((p-0.5)\frac{h\pi}{P}\right) \quad h = 1, 2, \cdots, H, \tag{3}$$

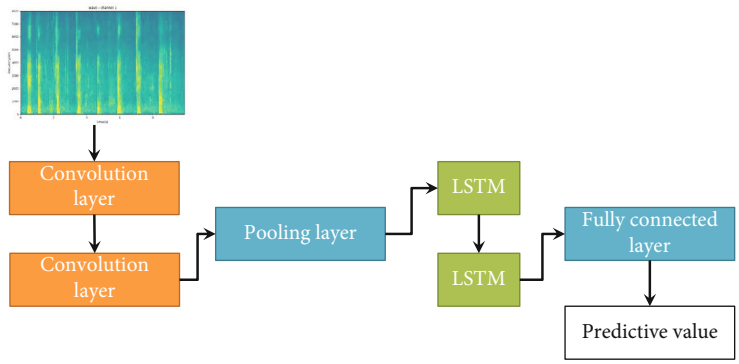

Figure 2: DR AudioNet network structure.

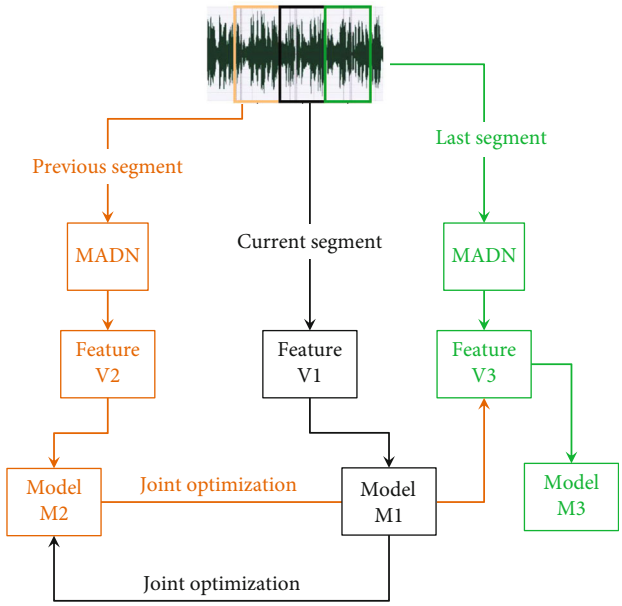

Figure 3: Structure diagram of the network used in this research.

where H represents the number of MFCCs, $H \in \{12, 13, 14, 15, 16\}$, $20 \le P \le 40$.

3.2.2. Resonant Peak, Energy, and Zero Crossing Rate. Energy usually refers to the average energy of a frame of speech signal, which represents the changing trend of the signal. It is possible to analyze whether the voice signal contains sound through energy analysis. By analyzing the voice characteristics of depression patients, it can be observed that the voice characteristics of depression patients mostly include intermittent, unclear speech, and long pauses. Therefore, energy characteristics can be used to identify depression. The so-called zero crossing rate refers to the number of times the signal value passes through the zero value per second. The zero-crossing rate feature has been widely used in the fields of speech recognition and music information retrieval and is a key feature for classification of percussive sounds. The combination of this feature and energy feature is one of the most commonly used feature combinations in speech recognition.

3.3. MADN Method. In the actual scene, each person has a different tone color and voice size, some people have a high pitch, and some people have a low voice. The existence of this voice feature reduces the accuracy of depression recognition, and the personalized speech feature will weaken the generalization ability of the depression recognition model. This is because many features such as MFCCs extracted from each frame of audio include not only depression-related features but also static characteristics of voice personality. In response to this, this article uses the MADN algorithm to identify and remove personalized features of depression. The calculation process of the above MADN algorithm is as follows:

(1) Read the original audio

(2) Preprocess all audio

(3) Extract features such as MFCCs, denoted by $V(k, s)$, s is the number of frames of speech, and each frame contains k elements

(4) $D(k, s)$ is obtained by differential calculation through the feature $V(k, s)$ of two adjacent audio segments, and the differential processing is used to eliminate the personalized information in the voice. $D(k, s)$ represents the time sequence change of audio, and the distribution of its characteristic value is relatively stable. The calculation is described as follows: S is the total number of frames of speech

$$D(k, s) = V(k, s+1) - V(k, s), \quad s = 1, 2, \cdots, S-1 \tag{4}$$

(5) Normalize different scales for different features

$$F(k, s) = \frac{D(k, s) - D_{\min}(k, W_k : S_k)}{D_{\max}(k, W_k : S_k) - D_{\min}(k, W_k : S_k)}, \quad k = 1, 2, \cdots, K \tag{5}$$

The values of S_k and W_k are different scales and sliding

windows, and the calculation formula is

$$W_k = \begin{cases} \max(0, k-5) & k = 1, 2, \cdots, 12, \\ \max(0, k-10) & k = 13, 14, \\ \max(0, k-15) & k = 15, 16, 17, \end{cases} \quad (6)$$

$$S_k = \begin{cases} \min(60, k+5) & k = 1, 2, \cdots, 12, \\ \min(60, k+10) & k = 13, 14, \\ \min(60, k+15) & k = 15, 16, 17 \end{cases} \quad (7)$$

(6) Output normalized features $F(k, s)$ of various scales

4. Simulation Experiment and Analysis

4.1. Experiment-Related Settings. In order to analyze the superiority of the network used in identifying depression, the evaluation indicators used are Mean Absolute Error (MAE) [22] and Root Mean Square Error (RMSE) [22]. The experiment mainly conducts research from two aspects; one is the influence of different feature extraction methods on the results. The second is to use different regression algorithms. The feature extraction methods used include separate MFCCs and multifeature combination. Comparison algorithms are logistic regression (LR) [23], CNN [24], and DCNN [25]. The parameter settings of the network used in this research are shown in Table 1.

4.2. Experimental Results and Analysis

4.2.1. Performance Comparison of 3 Models in the Network Used. In order to verify whether the three models in the used network have the expected optimization function, this study conducted experiments on the 2014AVEC test set. Figure 4 gives the recognition results of the three models on the voice data in the network. In Figure 4, from M1 to M3, the values of RMSE and MAE are gradually decreasing, which shows that after the optimization of the previous model, the performance of the latter model has indeed been improved. The reason for the gradual improvement of the recognition performance of the three models is that on the basis of the DR AudioNet network; the upper-level model is optimized using the features of adjacent voices, thereby improving the recognition effect.

4.2.2. Comparison of the Recognition Effect of Different Models on Depression. After comparing the performance of the three models in the improved network, we will further explore the effects of other features and models in the recognition of depression. Through comparative experiments, to demonstrate the effectiveness of the feature V1 and DR AudioNet network used in this article, Table 2 shows the depression recognition effect of different features and different models.

From the experimental results of the LR, CNN, and DCNN models, it can be seen that although the V1 feature is a combination of multiple features, the experimental results on the V1 feature are not better than the MFCCs alone. On DCNN, the recognition effect of the two features is not much different. On the AudioNet model, the results of the V1 feature are significantly better than the MFCCs alone. Furthermore, the network used in this paper recognizes the three characteristic data of V1, V2, and V3; the recognition result is obviously improved; and the performance advantage is obvious. Through comparative experiments, it can be concluded that the characteristics of the network used in this article V1 and the DR AudioNet network can effectively predict the degree of depression. When extracting features, different scales are used to normalize features, which effectively merges different features and retains audio depression features. At the same time, V2 and V3 are used to jointly optimize DR AudioNet, which effectively integrates the nonpersonalized depression characteristics of the MADN feature to the speaker.

4.2.3. Comparison of Noise Immunity. Noise immunity is a key indicator to measure the performance of an algorithm, because the actual data will inevitably be mixed with some noise data. These noisy data are bound to have a certain impact on the recognition results. An algorithm with good performance should have good noise immunity. In order to explore whether the algorithm in this paper is susceptible to noise, here, we add different amounts of Gaussian white noise to the dataset. During the experiment, this study added Gaussian white noise with a mean value of 0 and a variance of 0.01, 0.02, 0.03, 0.04, and 0.05 to the original dataset. Each model uses V1 features, and the evaluation indicators still use MAE and RMSE. The recognition results of each model on the noisy dataset are shown in Table 3.

The data in Table 3 shows that as the noise variance increases, the recognition performance of all models shows a downward trend. This is consistent with the theory. Among them, the performance decline trend of LR, CNN, and AudioNet is obvious. DCNN is not significantly affected by noise, and its noise immunity is the best overall. When the variance is less than 0.03, the network used is not sensitive to noise and has strong robustness. However, as the variance gradually increases, the performance of the network used begins to decline, and the downward trend is obvious, resulting in a lower performance than the DCNN model. In summary, the network used is suitable for processing micronoise data. Although the network used has certain antinoise performance, its antinoise performance needs to be further improved.

5. Design of Human-Computer Interaction System for Intelligent Recognition of Depression

In order to assist doctors in the preliminary diagnosis of patients with suspected depression, based on the above research results, this paper designs a human-computer interaction system for intelligent recognition of depression. The overall design requirements of the system are shown in Figure 5.

Table 1: Parameter settings.

Parameter	Value
Feature size	$17 * 60$
Batch size	32
The number of convolution kernels in the convolution layer	64
Convolution kernel size	$3 * 1$
The number of cells in the LSTM layer	128
Number of nodes in the first fully connected layer	128
The last fully connected layer	1

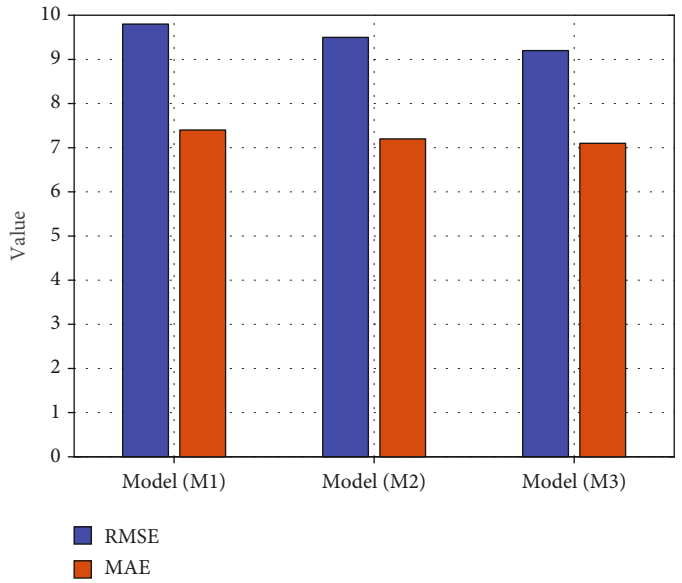

Figure 4: The recognition results of the three models in the network.

The depression recognition system designed in this paper is mainly divided into three modules: user voice recording module, depression recognition module, and About Us. The functional module is shown in Figure 6. Different modules corresponding to different functional requirements are as follows:

(1) User voice recording module: the main function of this module is to realize the function of recording and storing the user's own voice files. The corresponding functional operations can be realized by clicking different button controls. This module is mainly to realize the collection of user voice signal materials for subsequent depression recognition

(2) Depression recognition module: after the user selects the voice file to be recognized, the selected voice file is recognized for depression based on the trained algorithm model and the result is returned to the interface. The module also realizes the function of playing the selected voice file, so that the user can clearly know the sound content of the selected voice file

Table 2: Comparison of experimental results.

Model	Feature	MAE	RMSE
LR	MFCCs	8.112	10.314
	V1	8.240	10.453
CNN	MFCCs	7.984	10.168
	V1	8.067	10.282
DCNN	MFCCs	7.922	9.978
	V1	7.913	9.966
AudioNet	MFCCs	7.785	9.801
	V1	7.504	9.717
Network used	V1+V2+V3	7.230	9.215

(3) About our module: this part is mainly for product introduction and function description of the speech-based depression recognition application system. It is the software manual to help users understand the main functions and practical uses of the system

Table 3: Recognition results of noisy datasets by models.

Noise\model	Index	LR	CNN	DCNN	AudioNet	Network used
Mean 0, variance 0.01	MAE	8.382	8.764	8.226	8.251	7.976
	RMSE	11.103	11.180	10.375	10.469	9.893
Mean 0, variance 0.02	MAE	8.921	9.082	8.886	9.335	8.007
	RMSE	11.872	12.127	10.932	10.903	10.432
Mean 0, variance 0.03	MAE	9.644	9.828	9.273	10.272	8.865
	RMSE	12.824	12.794	11.412	11.583	11.224
Mean 0, variance 0.04	MAE	10.720	10.532	10.023	11.091	9.942
	RMSE	13.983	13.627	12.349	12.622	12.180
Mean 0, variance 0.05	MAE	11.365	11.284	10.892	12.022	10.957
	RMSE	14.921	14.532	13.200	14.177	13.284

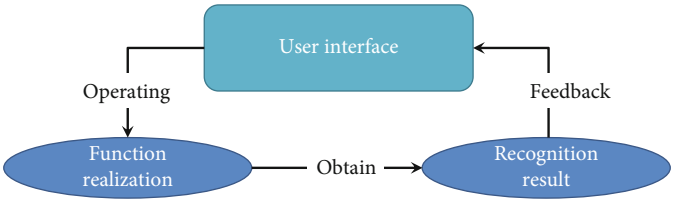

Figure 5: Analysis of system design requirements.

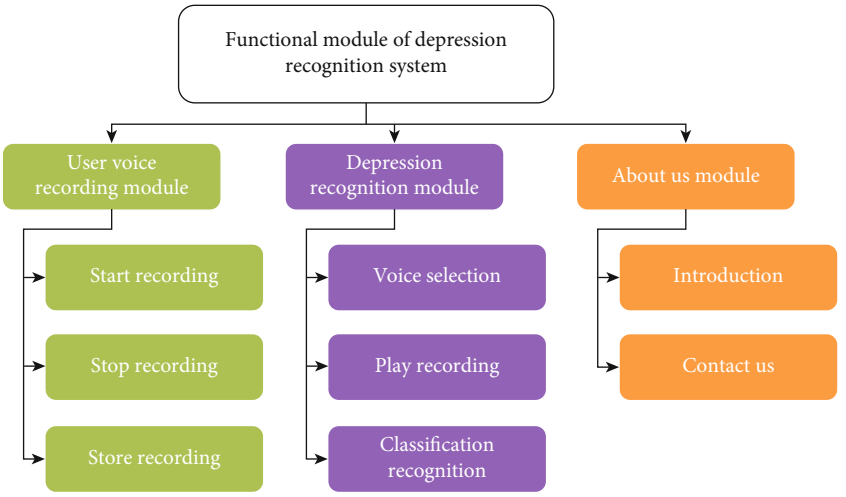

Figure 6: System function module.

6. Conclusion

This paper designs an intelligent depression recognition system. First, an improved network is used to recognize depression in voice data. The network uses the MADN feature extraction method to reduce the impact of the personalized voice characteristics carried by the sample on the sample. And use the adjacent voice features of the current voice segment to optimize the upper-level model to improve the recognition ability and generalization of the entire network. Experiments have verified that the network used can achieve a better recognition effect, and it is feasible and effective. Secondly, based on this basic research, this paper designs a depression recognition human-computer interaction system. After inputting the user's voice data, the result of depression recognition is output. The visual interface improves the patient's satisfaction with the visit and can effectively assist the doctor in the diagnosis. In the near future, we will continue to study the role of other modal data in the recognition of depression to further improve the accuracy of identifying depression. And we will also consider to use multiview learning, transfer learning, and the other advanced machine learning technologies to further develop the performance of the proposed intelligent recognition of depression system.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Jiangsu Province University Philosophy and Social Science Research 2019 Major Project, "Human-computer interaction design research based on artificial intelligence technology "(Project No. 2019 SJZDA118); the Higher Education Research Project of Nanjing Institute of Engineering in 2020, "Research on Cultivating Path of Artificial Intelligence Design Applied Talents" (Project No. 2020YB17); Youth Fund for Humanities and Social Science Research Project of the Ministry of Education (Project No. 20YJC760030); and the National Key R&D Program of China (Grant No. 2017YFB0202303).

References

[1] J.-P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatric Disease and Treatment*, vol. 7, Supplement 1, pp. 3–7, 2011.

[2] V. Manicavasagar, *A review of depression diagnosis and management*, vol. 34, no. 1, 2012In Psych: The Bulletin of the Australian Psychological Society Ltd, 2012.

[3] J. Olesen, A. Gustavsson, M. Svensson et al., "The economic cost of brain disorders in Europe," *European Journal of Neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[4] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, "Cost of lost productive work time among US workers with depression," *Journal of the American Medical Association*, vol. 289, no. 23, pp. 3135–3144, 2003.

[5] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[6] I. Schumann, A. Schneider, C. Kantert, B. Lowe, and K. Linde, "Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in pri-mary care: a systematic review of qualitative studies," *Family Practice*, vol. 29, no. 3, pp. 255–263, 2012.

[7] T. Inoue, Y. Inagaki, T. Kimura, and O. Shirakawa, "Prevalence and predictors of bipolar disorders in patients with a major depressive episode: the Japanese epidemiological trial with latest measure of bipolar disorder (JET-LMBP)," *Journal of Affective Disorders*, vol. 174, pp. 535–541, 2015.

[8] T. Vos, C. Allen, M. Arora et al., "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015," *Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016.

[9] A. Nilsonne, "Acoustic analysis of speech variables during depression and after improvement," *Acta Psychiatrica Scandinavica*, vol. 76, no. 3, pp. 235–245, 1987.

[10] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, 1993.

[11] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[12] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[13] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and Cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[14] D. J. France, R. G. Shiavi, S. Silverman, M. Wilkes, and M. Silverman, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[15] S. Scherer, G. Stratou, G. Lucas et al., "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[16] S. Kuny and H. H. Stassen, "Speaking behavior and voice sound characteristics in depressive patients during recovery," *Journal of psychiatric research*, vol. 27, no. 3, pp. 289–307, 1993.

[17] F. Tolkmitt, H. Helfrich, R. Standke, and K. R. Scherer, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," *Journal of Communication Disorders*, vol. 15, no. 3, pp. 209–222, 1982.

[18] H. Jiang, B. Hu, Z. Liu et al., "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, pp. 39–46, 2017.

[19] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: from classification to treatment outcome prediction," *CNS neuroscience & therapeutics*, vol. 24, no. 11, pp. 1037–1052, 2018.

[20] J. Lorenzo-Trueba, G. Eje Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[21] J. F. Zhao, X. Mao, and L. J. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[22] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE transactions on autonomous mental development*, vol. 10, no. 3, pp. 668–680, 2018.

[23] W. Pan, J. Wang, T. Liu et al., "Depression recognition based on speech analysis," *Kexue Tongbao/Chinese Science Bulletin*, vol. 63, no. 20, pp. 2081–2092, 2018.

[24] Z. Wang, L. Chen, L. Wang, and G. Diao, "Recognition of audio depression based on convolutional neural network and generative antagonism network model," *IEEE Access*, vol. 8, pp. 101181–101191, 2020.

[25] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.