Received May 15, 2020, accepted May 26, 2020, date of publication May 29, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998532

Recognition of Audio Depression Based on Convolutional Neural Network and Generative Antagonism Network Model

ZHIYONG WANG, LONGXI CHEN, LIFENG WANG, AND GUANGQIANG DIAO
School of Information Engineering, Shandong Youth University of Political Science, Jinan 250100, China

Corresponding author: Zhiyong Wang (yong@sdyu.edu.cn)

This work was supported in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2016AM30, in part by the Social Science Planning Research Project of Shandong Province under Grant 18CLYJ50, in part by the Shandong Soft Science Research Program under Grant 2018RKB01144, and in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J15LN15.

ABSTRACT This paper proposes an audio depression recognition method based on convolution neural network and generative antagonism network model. First of all, preprocess the data set, remove the long-term mute segments in the data set, and splice the rest into a new audio file. Then, the features of speech signal, such as Mel-scale Frequency Cepstral Coefficients (MFCCs), short-term energy and spectral entropy, are extracted based on audio difference normalization algorithm. The extracted matrix vector feature data, which represents the unique attributes of the subjects' own voice, is the data base for model training. Then, based on the combination of CNN and GAN, DR AudioNet is used to build the model of depression recognition research. With the help of DR AudioNet, the former model is optimized and the recognition classification is completed through the normalization characteristics of the two adjacent segments before and after the current audio segment. The experimental results on AViD-Corpus and DAIC-WOZ datasets show that the proposed method effectively reduces the depression recognition error compared with other existing methods, and the RMSE and MAE values obtained on the two datasets are better than the comparison algorithm by more than 5%.

INDEX TERMS Recognition of audio depression, generative antagonism network, convolutional neural network, Mel-scale frequency cepstral coefficients, entropy feature of spectrogram.

I. INTRODUCTION

With the improvement of people's material life, mental health issues have received widespread attention. Depression is a main category of mood disorders, which the major clinical features are significant and persistent mood depression, loss of interest and pleasure. Patients with mild depression show symptoms of depression, anxiety, loss of interest and low self-assessment; while patients with severe depression will be pessimistic, desperate, hallucinatory delusions, physical decline, and even suicide [1]–[3]. Hopefully, people with depression can be relieved and cured by medication, psychological and physical means.

Benefit from the development of biometric recognition technology, researchers can obtain various information including speaker identity information, age, gender, speech content and emotion by analyzing voice. Clinical observations and studies have found that there is a significant correlation between the audio characteristics and the depression degrees [4], [5]. The language features of depression patients are being slow, monotonous and overcast, which are different from normal population [6], [7]. Currently, the Beck Depression Inventory II (BDI-II) is most widely used self-assessment scale for depressive symptoms and is the tool to assess the degrees of depression [8]. The objective evaluation and rapid identification for identifying the degrees of depression based on computer technology are conducive to diagnosis and treat the early depression patients. Therefore, depression recognition technology based on speech signals

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

has been the focus of scholars due to its advantages of low cost, easy collection and non-contact [9], [10].

Specifically, Automatic Speech Depression Detection (ASDD) attempts to explore the speaker's inner emotions and psychological activities by analyzing their speech signal and changing process based on computer digital signal. The current ASDD methods can be divided into two categories, traditional machine learning methods and deep learning methods. Compared with traditional machine learning methods, deep learning models can extracting high-level semantic features based on the neural network framework, which has brought breakthough progress in recent years. In this paper, we proposed a novel deep learning algorithm which combine convolutional neural network (CNN) and generative antagonism network (GAN) and for ASDD. There are three key problems in the study of speech based depression recognition: (1) how to design experiments to obtain high-quality speech data; (2) how to determine the effective features in many complex speech features; (3) how to build an efficient recognition model. This paper focuses on these three problems, and proposes an audio depression recognition method based on convolution neural network and generative antagonism network model.

II. RELATED WORK

For the audio recognition problem, scholars have proposed many methods, in [11] they constructed a one-dimensional long-short term memory (LSTM) and a two-dimensional LSTM to extract local and global emotion related features in speech, which can improve the accuracy of original model by combining the two features. Chao *et al.* [12] extracted the features of audio and video, and fused them into signs of abnormal behavior, then used long short term memory recurrent neural network (LSTM RNN) to describe dynamic time information. They used multi task learning to improve the accuracy of the results. The audio recognition method based on matrix theory proposed in [13] can be roughly divided into two steps. Firstly, their method extracted features through Mel frequency cepstrum coefficients (MFCC), and then used matrix theory to characterize the energy mode band defined for a specific time frame and frequency, which furture improves the accuracy of the model. Guo *et al.* [14] proposed a joint learning model based on label relaxation low-rank ridge regression (JOLL4R), which commonly learned the conversion of each audio by coupling the regression loss of audio matrix. However, these methods are prone to overfitting in data training and lack generalization in practical applications.

In [15] work, they proposed an audio recognition method based on transfer learning. When the model is applied to new samples, the algorithm can convert the data to the specific class by identifying the category that is statistically close to the associated data item in limited datasets. Wang *et al.* [16] proposed a layered sparse coding framework for language emotion recognition, which can automatically represent audio features. The model proposed in [17] can extract feature vectors for environmental sound classification by utilizing discrete features and support vector machine to obtain the correlation of coefficients through linear relationship with the scattering transform. which utilized discrete features. In [18] work, they proposed an audio recognition method based on line spectrum frequency and extreme machine learning. Their model can reduce the computational amount, which replaced the non-speech part of the input signal by using the statistical characteristics of line spectrum frequency.

The model in [19] applied visual words on the audio fragment spectrum to recognize speech information. Their method can mine the representative features by utilize the visual bag of words method to accelerate the extraction process of robust features of audio, and then quantify it into a set of visual words and image histograms. In [20] work, they proposed a Bayesian audio recognition method based on convolutional non-negative matrix factorization, which imposed certain features on the time-frequency components of the recovered signal and reverberation component through the prior distribution. Hoirin *et al.* [21] proposed a feature contribution network (feature convolutional neural network, FCNN), whose output layer is composed of sigmoid contribution gates and learned the element-level contribution of input features. However, when the samples in data are not clean and exists many noisy samples, these models cannot show better classification performance.

In [22] work, they proposed a human language emotion recognition framework based on time information and deep learning, their algorithm can assess emotional statements in time series information by combining long and short-term memory networks. Bavu *et al.* [23] proposed an audio recognition method based on a learnable second-order filter and a deep separable one-dimensional convolutional residual network, which can learn representation of audio information by learning the time correlation between the sample level and the frame level. In [24] work, they proposed an audio recognition method (Multi-Stream HMM, MSHMM) based on concealed Markov model. Their model can extract the corresponding anti noise audio features through deep noise reduction automatic encoder. In addition, Hossein *et al.* [25] proposed a single-stage hidden Markov audio recognition method based on smooth start, which adopted a complete Biphone to achieve flat-start context-sensitive model. However, these methods cannot apply the memory information in the previous training to assist the current tasks. Although previous researches have achieved certain results, there are still many areas for improvement. Studies have shown that depression patients have significant differences in emotions from normal people, like depression, sadness, anxiety and worry. To alleviate the problem of less training data and make full use of emotional features, we employ multi-scale audio differential normalization features based on contextual emotional information and propose a novel convolutional neural networks and generative confrontation networks framework. The comparison experiment results with current methods show that our proposed audio depression recognition algorithm has better performance in the diagnosis of depression, overall the main contributions of our model are as follows:

(1) A novel audio depression regression prediction network based on CNN and GAN is used to construct a more efficient classification and recognition model, experimental results have shown that our model achieve the best BDI-II value in training data.

(2) For the shortcoming of only using single MFCCs as the input of the network, which cannot make full use of audio features. We adopt new forms of features based on spectrogram entropy, which is furtherly to measure the change of speech signals. Our model first select the corresponding speech frame through the sliding window and normalize it, then extract the entropy features of MFCCs and short-term energy spectrogram to obtain local non-personalized depression features.

III. RECOGNITION OF AUDIO DEPRESSION BASED ON DEEP LEARNING

A. THE FRAMEWORK OF OUR MRTHOD

In our work, we first perform the front-end preprocessing and feature extraction on the voice sample files, and send the features into the Depression AudioNet framework for training, and classify and identify depression in the final fully connected layer. The specific flow of our proposed depression recognition algorithm based on CNN and GAN is shown in Figure 1.

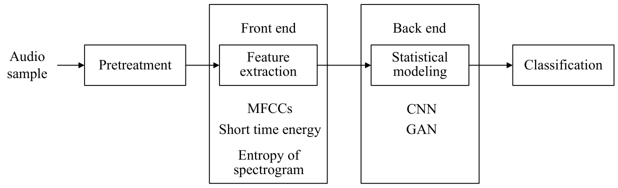

FIGURE 1. The framework of our method.

The specific processes of depression recognition are:

1)Pre-processing the original voice sample file to remove long-term silent fragments in the data set, and splice the rest parts into a new audio file.

2)There are three types of features in our model, Mel frequency cepstrum coefficients of speech signal based on audio difference normalization algorithm, short-term energy and entropy features of the spectrogram. The extracted feature data of the unique attributes of the patient's own speech is the data basis for model training

3)The back end adopt the combination of CNN and GAN to construct the depression recognition research framework, the applied methods is the expansion of the existing research work.

4)Based on the experiment results of the combination of different features and classification models, we select the most discriminative features and classifiers to construct the Depression AudioNet recognition framework.

B. DATA PREPROCESSING

Performing pre-processing operations on audio samples to obtain representative features. Cause the subjects have no audio information when listening to the question, the non-voice segment part should be cut out. Firstly, for each audio file, the long silence section need to be removed and the rest is spliced into a new audio file. Then, each valid audio file is divided into the same length and no overlapping audio segment, which is composed of 60 frames. Hamming window is selected for the audio frame and each frame contains 1024 data points. The overlapping part of the previous frame and the next frame is the frame length, the audio sampling rate is 44100Hz, so the time covered by an audio clip is:

$$[(60 + 1) \times 1024/2]/44100 = 0.708s \tag{1}$$

C. AUDIO FEATURE EXTRACTION

1) MFCCs

MFCCs are the most commonly used features in speech signal processing, which have the advantages of being consistent with human hearing and low dimensions. MFCCs effectively combine the human ear's auditory perception characteristics with the voice signal generation mechanism. The following formula explains how to convert the common frequency domain scale of audio to MFCCs frequency scale:

$$f_{mel} = 2595 \log(1 + \frac{f_{H_z}}{700}) \tag{2}$$

In the formula: f_{mel} represents Mel frequency scale, f_{H_z} represents common frequency. Generally, the calculation for MFCCs uses a set of filters whose center frequencies are arranged at even intervals according to the Mel frequency scale, and the frequencies of the two bottom points of each filter triangle are equal to the center frequencies of two adjacent filters. Suppose the number of filters is M, the output after filtering is $X(m)$, $m = 1, 2, \cdots, M$; and set $l(m)$, $c(m)$, $h(m)$ be the lower limit frequency, center frequency and upper limit frequency of m triangular filters. Then the lower limit, center and upper limit frequency of the adjacent triangle filter have the following relationship:

$$c(m) = h(m-1) = l(m+1) \tag{3}$$

Performing logarithmic operation on the output of the filter bank d, and then inversing discrete cosine transform to obtain MFCCs:

$$C_n = \sum_{m=1}^{M} \log X(m) \cos((m-0.5)\frac{n\pi}{M}) \quad n=1, 2, \cdots, L \tag{4}$$

In the formula, L is the coefficient of MFCCs. The numbers of filter M are between 20-40, in our paper, we set $M = 32$。

2) ENTROPY CHARACTERISTICS OF SPECTROGRAM

We attempt to construct a new feature form the characteristic that depression patients speak more "monotonously" than common people. It is assumed that the monotonous sounds are more concentrated in frequency distribution, and

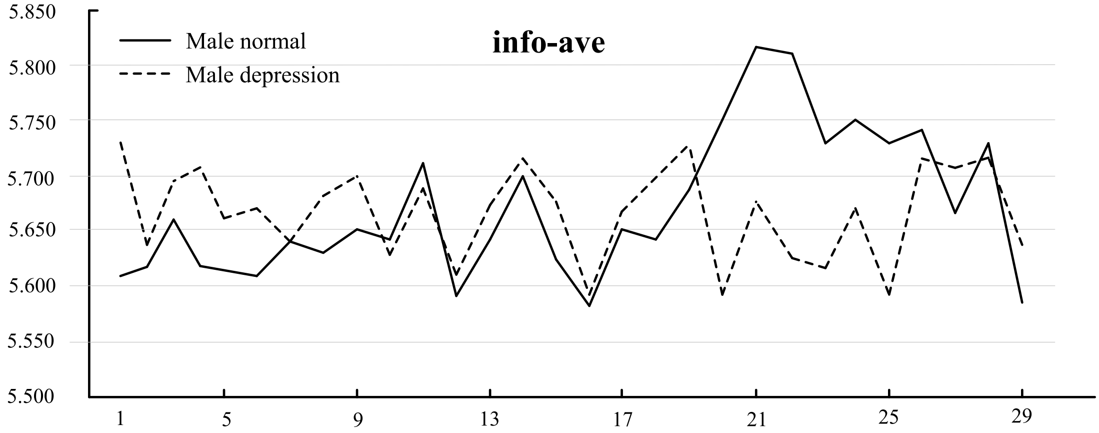

FIGURE 2. **Spectral entropy characteristics (male).**

the frequency components of rich sounds are more diverse. Based on the measurement method of signal harmonics in Omori [26] *et al.* work, we can use the concept of entropy to describe this hypothesis. Entropy is usually used to measure the uncertainty of the signal source, the greater the uncertainty, the greater the entropy, and the corresponding probability distribution will tend to be "scattered". Otherwise, if the distribution tends to be "centralized", the uncertainty is small and the corresponding entropy is also small. In speech signals, people usually apply spectrograms to express the changes of speech in dimensions of time and frequency, and we can directly calculate the entropy of the spectrogram. The hypothesis is: monotonic sound corresponds to a smaller spectrogram entropy value, while rich sound corresponds to a larger spectrogram entropy value. Wth clinical observation, the above hypothesis can be simply expressed as: the spectrogram entropy of common people should be greater than that of depression patients. The calculation formulas of the spectrogram entropy are as follows:

1. Retain the audio segment. The subband entropy spectrum method is used to distinguish the voiced and unvoiced segments, and only the voiced segments are used for subsequent calculations.

2. Calculation of the spectrogram. The Short Time Fourier Transform (STFT) is used to calculate the spectrogram, the formula is as follows:

$$X(t,f) = \int_{-\infty}^{+\infty} w(t-\tau)x(\tau)e^{-j2\pi f\tau}d\tau \tag{5}$$

$x(t)$ is the input voice signal, $w(t)$ is the window function, $X(t,f)$ is the time-frequency domain representation of the input speech signal. In calculation process, the adopted $w(t)$ is a rectangular window with a length of 25ms, time step $d\tau$ is 5ms. Meanwhile, in our work, we noly retain the amplitude information in the time-frequency domain while ignore phase information: $A(t,f) = \|X(t,f)\|$, The normalized $A(t,f)$ is the spectrogram.

3. Calculation of the spectrogram entropy: Converting the obtained spectrogram to an 8-bit grayscale image, and then calculating the grayscale entropy according to the following formula:

$$H_{info} = -\sum\nolimits_{i=0}^{255} p(i)\log_2 p(i) \tag{6}$$

$$H_{renyi} = -\frac{\log_2 \sum_{i=0}^{255} p^n(i)}{n-1} \tag{7}$$

$p(i)$ is the probability density function of the gray value, H_{info} is the information entropy, H_{renyi} is *renyi* entropy. In the calculation, only a specific frequency of the entire spectrogram band is intercepted, and the spectrogram is cut into multiple parts at 50ms intervals to calculate separately. The obtained sequence of time entropy: $H_{info}(t)$ and $H_{renyi}(t)$。

4. Feature construction. Calculating the maximum, minimum, mean, median and variance of $H_{info}(t)$ and $H_{renyi}(t)$ as feature values。

The entropy characteristics of male and female spectrograms are shown in Figure 2 and 3 respectively. From the observations in Figure 2 and 3, it can be found that the feature of the mean value of the spectrogram entropy (info_ave) does not perform well in the male population, only showing differences in the speech segments read by some vocabulary, while women shows difference in almost all the speech segments.

In this paper, the T test is used for the verification of the entropy characteristics of the spectrogram. T-test results show that the median and mean values of entropy have significant differences in some speech segments ($p < 0.05$). Notably, the frequency range of speech segment interception has a greater impact on the number of speech segments with differences between groups (it can be seen in Table 1). In terms of information entropy characteristics, when the input spectrogram contains 0 ~ 2000Hz, there are only a few speech segments are significantly different between normal and depression people. When it becomes 1000 ~ 2000Hz, the paragraphs with significant differences further reduced, and women are even 0. And when choosing 250 ~ 1250Hz, there are relatively many different paragraphs, the paragraphs in men and women are 11 and 12 respectively. The specific frequency range selection experiment results are shown in Table 1.

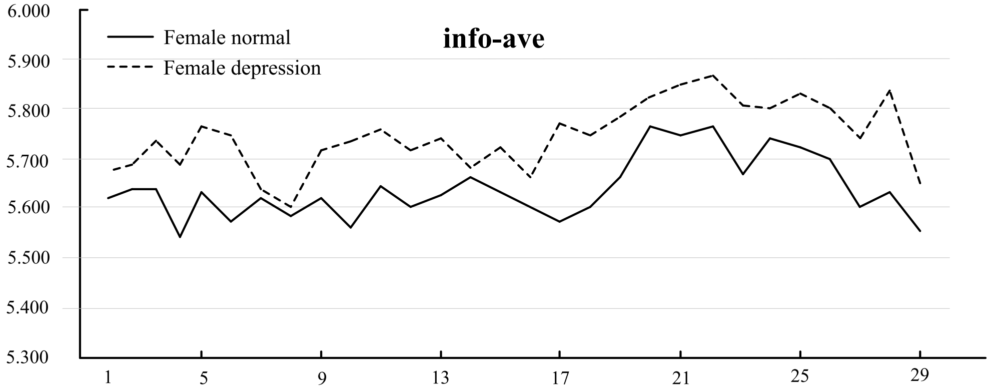

FIGURE 3. Spectral entropy characteristics (female).

TABLE 1. In different frequency ranges, the number of speech segments with H_{info} characteristics that with different groups.

Gender	Frequency Range（Hz）				
	0-2000	0-750	50-750	250-1250	1000-2000
Male	6	6	11	11	4
Female	7	1	1	12	0

From Table 1, it can be seen that the number of segments where the two types of people differ on different frequency bands is inconsistent, and the frequency range where males and females differ significantly is also different. In this paper, we suspect that there are two cases for this phenomenon: (1) The information of human speech is mostly concentrated in the range of 1 ∼ 3 times the fundamental frequency, and some noise components dominate outside this frequency range, so effective information cannot be extracted by entropy in the full frequency band. (2) According to the principle of STFT, in time-frequency analysis, when the time resolution and frequency resolution are fixed, the range with the highest accuracy will be concentrated in a specific frequency band. The selected framing method and window function may make the best resolution range is concentrated around 50 ∼ 1000Hz. The above assumption need to be verificated by experimental results, and we attempt to improve the discriminating ability of this feature in subsequent studies.

3) Short term energy. Short-term energy refers to the average energy of a frame of speech signal, reflecting the amplitude change of the speech signal. Short-term energy can be used to distinguish between voiced and silent, depression patients have symptoms of slurring and pauses, and these symptoms will become more prominent with the degree of depression increases. Therefore, short-term energy characteristics can be furtherly used to analyze speech pause information.

D. THE NORMALIZATION ALGORITHM FOR MULTISCALE AUDIO

When doctor gets more audio information, the more accurate of the diagnosis result for depression patients. However, the audio data of the depression database is limited. Since the number of audio segments is inversely proportional to the audio duration of a single sample, therefore with the increase of a single sample duration will result in a reduction in the total number of sample segments. The increase in the data dimension of a single sample will also greatly increase the complexity of the calculation, affecting the calculation speed and recognition accuracy of the model. This is also an urgent problem to be solved in the current research on audio-based depression recognition. This is an urgent problem to be solved in research on audio-based depression recognition. In life, different speakers have different volume and timbre characteristics, and some people are born with a higher voice while some people are born with a deep voice. The personalized speech characteristics of the speaker will negatively affect the accuracy of depression recognition model. The MFCCs, short-term energy and spectrogram entropy features extracted for each frame of audio contain a large number of depression-related features, and are also mixed with the speaker's personality speaking characteristics, which is caused by its static properties. The speaker's personalized speech characteristics will weaken the generalization ability of the depression recognition model, therefore, a multi-scale audio normalization algorithm is used to obtain local non-personalized depression features [29].

Features based on audio differences reflect audio change information during the same speaker's speaking process, and are not easily affected by personalized speaking characteristics. Since the data levels of various features are different, the features are applied to be normalized with different scales. In order to obtain the local change information of the speaker's audio, the corresponding speech frame is selected according to the sliding window for normalization, instead of comparing with a whole section of audio. The selected speech frame according to the sliding window can enhance the dynamics of local audio changes [30], which more effectively reflects the non-personalized audio features. The steps of multiscale audio difference normalization algorithm framework are as follows:

1 Input the original audio file.

2 Reading and preprocessing all audio files.

3 Extracting MFCCs, Short-term energy, using $V(n,f)$ to represent zero crossing rate and formant frequency characteristics, f is the number of voice frames, each frame contains n elements.

4 Obtaining $D(n,f)$ through the difference calculation of the audio features $V(n,f)$ of two adjacent frameso $D(n,f)$ represents the temporal change of audio, which weakens the personalized information of the speaker's speech, and the distribution of feature values is relatively stable under the same degree of depression. The calculation method is as follows:

$$D(n,f)=V(N,f+1)-V(n,f) \quad f=1,2,\ldots,F-1 \tag{8}$$

5 Normalizing different features at different scales:

$$F(n,f)=\frac{D(n,f)-D_{\min}(n,f_n:F_n)}{D_{\max}(n,f_n:F_n)-D_{\min}(n,f_n:F_n)} \quad n=1,2,\ldots,N \tag{9}$$

In the formula: The values of F_n and f_n represent different scales and sliding windows, the formula is as follows:

$$f_n=\left\{\begin{array}{l} max(0,n-5)n=1,2,\ldots,12 \\ max(0,n-10)n=13,14 \\ max(0,n-15)n=15,16,17 \end{array}\right\} \tag{10}$$

$$F_n=\left\{\begin{array}{l} min(60,n+5)n=1,2,\ldots,12 \\ min(60,n+10)n=13,14 \\ min(60,n+15)n=15,16,17 \end{array}\right\} \tag{11}$$

6 Output: $F(n,f)$ is the normalized feature at different scales.

E. AUDIO DEPRESSION REGRESSION PREDICTION NETWORK

Recently, deep learning technology has developed rapidly and achieved good results in the field of speech signal processing. It can learn to generate advanced semantic information and enrich hand-designed features. Huang *et al.* [27] proposed a binary classification network structure for identifying depression in the 2016 AVEC competition, which is mainly composed of CNN and LSTM. The input of this model is audio information, and the output is whether the corresponding individual is depressed audio. It is audio information, and the output is whether the corresponding individual is depressed audio or other. Our work make two aspects of optimization and improvement bases on this network model: (1) In order to further extract patient audio features, adopting the complementary combination of MFCCs, short-term energy and spectrogram entropy features as the model input; (2) The model based on the classification of depression is improved to the model of regression prediction of depression. Since patients with various degrees of depression require different treatments, it is necessary to predict the BDI-II value of depression patients. Our proposed deep learning model, namely the network structure of audio Depression Regression AudioNet (DR AudioNet) is shown in Figure 4.

In our work, we firstly combine the CNN and GAN for audio depression regression prediction. CNN method has better classification performance for large-scale image processing, and the GAN method can generate samples close to real sample distribution to expand the train data set. Our proposed deep learning model can improve the classification ability of neural network, and better identify depression patients.

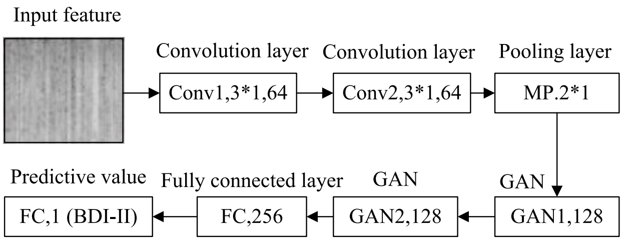

FIGURE 4. **The framework of audio depression regression prediction network model.**

The core formula of the theoretical basis of the GAN algorithm to achieve sample generation and identification of the entire process is:

$$\min_G \max_D V(D,G)=E_{X\sim P_{data(x)}}[\log D(x)] +E_{z\sim P_{z(z)}}[\log(1-D(G(z)))] \tag{12}$$

The formula shows that the realization process of the establishment and implementation of the GAN algorithm is the process of game and confrontation between generator G and discriminator D. Theoretically, when the probability that the generator generates samples is equal to the probability of real sample data, the training process reaches the ideal state. That is $p\left(G\left(x\right)\right)=P_{data}\left(x\right)$.

$$E_{x\sim P_{data}(x)}\left[\log D\left(x\right)\right] \tag{13}$$

To optimize the modeling results, the discriminator D can accurately predict the probability distribution of real sample data. Meanwhile, the variable x represents the real sample data in the experiment, $P_{data}\left(x\right)$ represents the probability distribution corresponding to the real sample data

$$E_{z\sim P_{z(z)}}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \tag{14}$$

According to the relevant theoretical knowledge of the logarithmic function, when the independent variable is less than 1, making the mean value $D\left(G\left(z\right)\right)\approx 0$ to maximize equation (14). By maximizing $V\ (D, G)$, the the discriminant model confuses the difference between the generated sample and the real sample. The variable z represents the input random noise from the generative network G, then $G\left(z\right)$ is the sample data generated by the built network. The entire network achieve its functional purpose through this extremely small confrontation.

The operation of convolution is the process of weighted summation of range variables, the expression is shown in equation (15). In the deep learning field research process, x in equation (15) is input, ω is convolution kernel, s is the output feature map.□

$$s\left(t\right)=\left(x*\omega\right)\left(t\right) \tag{15}$$

In practice operation, the case that the independent variables in Equation (15) are continuous values may not be satisfied. The input of the convolutional neural network is usually multi-dimensional feature data, then the corresponding convolution kernel is also the same dimensions. The expression of convolution in machine learning algorithms is

$$s\left[i,j\right]=\left(I*K\right)\left[i,j\right]=\sum_{m}\sum_{n}I\left[i+m,j+n\right]K\left[m,n\right] \quad (16)$$

The combination of GAN and CNN methods has made some changes for GAN network in the stability perspective. The specific network structure changes are as follows:

1) In the discriminant model, the mobile window is used to replace the space pooling to reduce dimensions. Using corresponding function to correct the deviation caused by reducing the dimension without using pooling directly, the discriminant network in this method is a convolutional neural network without a pooling layer. Generative model is a deconvolution process.

2) Connecting the input and output of the generative model and discriminant model with a convolutional layer to make the network a fully convolutional network

3) Except the input of the discriminant model and the output of the generative model, batch normalization is used in remaining parts of the network construction to stabilize the process of balanced learning.

4) The generative model uses ReLU as the activation function, and the last output layer uses tanh as the activation function.

5) Discriminant model uses LeakyReLU as activation function.

In the early stage, the GAN algorithm was used to train the model to generate sample data that was similar to the real sample for later classification tasks, the formula is:

$$\begin{aligned} &P\left(H\left(x\right)\neq f\left(x\right)\right) \\ &\quad=\sum_{k=0}^{\lfloor T/2\rfloor}\binom{T}{K}\left(1-\varepsilon\right)^{k}\varepsilon^{T-k}\leq \exp\left(-\frac{1}{2}T\left(1-2\varepsilon\right)^{2}\right) \end{aligned} \quad (17)$$

The mathematical relationship of the above formula shows that the larger the number of individual classifiers that represent each other in the algorithm, the error rate of the entire result will show an exponential decline, which greatly improves the final correct rate of the classification result. Where $H\left(x\right)\neq f\left(x\right)$ represents the case where the experimental result is not equal to the real result, and the final rate is the error rate with above situation.

IV. RECOGNITION AND CLASSIFICATION OF AUDIO DEPRESSION BASED ON DR AUDIONET NETWORK

In deep convolutional neural networks, the shape of the input image and the convolution kernel are often square, while the data dimension of the speech signal is one-dimensional. Therefore, deep learning model cannot be directly used as image processing methods. To solve this problem, MFCCs, short-term energy and spectrogram entropy features are extracted for each frame of speech in the audio segment in the experiment, and then form a two-dimensional matrix based in 60-bit speech features of each segment. In the two-dimensional matrix, the horizontal axis represents time and the vertical axis represents frequency information. The same spectrum model can represent completely different audio in different frequency intervals, and the square convolution kernel and pooling operation used by CNN for image processing will cause confusion between different audio and weaken the recognition ability [28], [29]. In our work, we attempts to use a one-dimensional convolution instead of a square filter on the entire frequency axis to solve this problem. The convolutional layer can effectively capture rich high-order semantic information. The pooling layer is used to reduce the dimension of the feature map and introduce invariance to small changes in relative position, which effectively improve classification accuracy and reduce operation complexity. Through convolution and pooling operations, two-dimensional input features will become one-dimensional deep features. Then, these features are imported into the GAN layer to extract long-term dependency information. Finally, the end of the network architecture are two fully connected layers, which are used to encode long-term changes in audio on the timeline and predict depression scores.

The Depression AudioNet network only uses the features of the current audio segment. To make full use of the non-personalized depression features of the two audio segments based multi-scale audio difference normalization algorithom, we adopt several commonly used network model fusion Methods and characteristics. Traditional neural network fusion is usually to linearly weight the prediction values of several networks or use a random gradient to perform weighted fusion. The number of samples involved in training has not increased, but the type of features has been increased, which is a parallel fusion method. [30].

The non-personalized depression features based on the multi-scale audio difference normalization algorithm are obtained on the two adjacent audio segments input features in the Depression AudioNet network, they have the same size, BDI-II value, and the temporal Relevance. The proposed novel network model architecture is shown in Figure 5. It can be seen that the latter model is trained on the basis of the former model, which the parameters of adjacent model are shared. Firstly, the model use the Depression AudioNet network to extract MFCCs, short-term energy, and spectrogram entropy features (called feature V1) from the data set for training. The Depression AudioNet network can separately predict the depression score, but it is not well-processed on the speaker 's personalized information. Therefore, on the basis of this model, the multi-scale audio difference normalization algorithm feature (called feature V2) of the previous segment is used to train model two, the features related to depression are learned and reduced the interference of the speaker's personalized voice. Finally, on the basis of model two, algorithm can extract the multi-scale audio difference

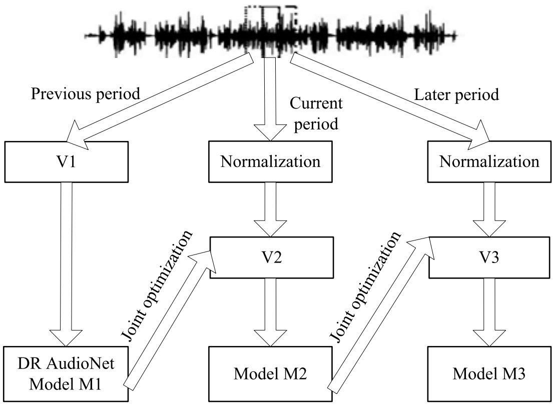

FIGURE 5. The overall architecture diagram of Identification and classification network model.

TABLE 2. AViD-Corpus data set distribution.

Dateset	Non-depressive individuals	Depressed individuals
Training sets	28	22
Test sets	26	24

normalization algorithm features (called feature V3) of the subsequent segment to train model three. By combining the advantages of two models, the algorithm can extract more discriminative depression feature information.

V. EXPERIMENTAL RESULTS AND ANALYSIS

To verify the effectiveness of the proposed audio depression recognition method based on convolutional neural network and generative confrontation network model, we conduct several comparison methods, including FCNN algorithm proposed in [21], the TMSN algorithm proposed in [23], and the MSHMM algorithm proposed in [24]. The hardware environment configuration used in the experiment is: 3.5GHz Intel Xeon E5-2697 v2 CPU, 32GB memory and 1T hard disk.

A. DATASETS

The data set used in the experiment is the AViD-Corpus and the DAIC-WOZ.

AViD-Corpus data set voice material collection is done in a relatively quiet environment through a laptop with a built-in sound card to complete the voice collection task. Each participant will be recorded one to four times, and the interval between each recording is about two weeks. The voice content of the data set is embodied by giving the subjects different specific tasks to stimulate the emotional state of the subjects under different language environment conditions to mobilize the individual emotions of the subjects. Subjects with depression tended to express different emotional states from non-depressed individuals through the source voice files.

The distribution of depressed and non-depressed individuals in the AViD-Corpus data set is shown in Table 2 below.

The DAIC-WOZ data set is mainly for audio, video and feedback data collected from the diagnosis of anxiety, depression and other psychological diseases. This data set was cited in the depression recognition task module in the 2016 and 2017 AVEC competitions. The virtual interviewer Ellie and the subject's situational dialogue are used to realize the interview to stimulate the subject's corresponding emotional response and record the linguistic and non-verbal emotional characteristics in real time.

The distribution of depression and non-depression individuals in the DAIC-WOZ data set is shown in Table 3.

TABLE 3. DAIC-WOZ sample distribution.

Dataset	Non-depressive individuals	Depressed individuals
Training sets	77	30
Test sets	23	12

The experiment is based on a larger data set, and Part of the data set is used in the competition This article and the competition used part of the voice material in this data set. A total of 189 samples in dataset and the official standard is divided into 107 training samples, 35 verification samples, and 47 test samples. There is part of no publicly labeled samples indicating whether the individual is depressed in the test set, in order to ensure the accuracy and validity of the experimental results, we only uses 142 samples in the training set and the verification set. The details are shown in Table 3. The label division in the experiment is based on the value of the patient's own health survey score evaluation. The data set gives the specific score of the PHQ8_Score questionnaire survey result, and also gives the PHQ8_Binary binary tag which 0 is a non-depressed individual, and 1 is a depressed individual.

B. EVALUATING INDICATOR

To evaluate the effectiveness of the proposed method on the recognition of depression, we take Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as evaluation indicators. The smaller of values of MAE and RMSE, the more accurate of proposed algorithm predicts depression. MAE can well reflect the error between the predicted value and the true value. Assuming that there are N samples, the true label value of each sample is $y_i(i = 1, 2, \ldots, N)$, and the predicted value is $\hat{y}_i(i = 1, 2, \ldots, N)$, then the average absolute error represents the average value of the absolute error between all predicted and true values. The calculation formula is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{18}$$

RMSE represents the root mean square of the error between all predicted values and the true value, and is used to measure the deviation between the predicted value and the true value.

The calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{19}$$

C. MODEL TRAINING

The feature size of the input data of the DR AudioNet network is set 17 × 60, the batchsize is 32, both convolutional layers have 64 convolution kernels, the size of the convolution kernel is 3 × 1, and the number of cells in the GAN layer is set to 128 The number of nodes in the first fully connected layer is also 128, and only one node in the last fully connected layer to outputs the prediction score. The number of filters M is between 20-40, and we set $M = 40$ according to [8]. L is the MFCCs coefficient, which the value is usually 12-16. In order to obtain the best recognition performance, the value of L is experimentally discussed in this paper. The experimental results are shown in Table 4. From Table 4, it can be seen that the optimal recognition performance is obtained when the value of L is 13.

TABLE 4. Comparison results of the three models on the AViD-Corpus and DAIC-WOZ test sets for depression identify.

Datasets	L	Models	RMSE	MAE
AViD-Corpus	12	M1	9.13	7.86
	13		9.02	7.78
	14		9.04	7.82
	15		9.11	7.86
	16		9.21	9.94
DAIC-WOZ	12	M1	10.35	8.06
	13		10.21	7.96
	14		10.54	8.21
	15		10.83	8.43
	16		11.04	8.67

After determining the value of L in the MFCCs coefficient, the overall performance of the DR AudioNet network model is evaluated on the test sets of the AViD-Corpus data set and the DAIC-WOZ data set, respectively.

The results are shown in Table 5, it can be seen that the model M2 has better performance than the model M1. In the model M2, the multi-scale audio difference normalization algorithm is used to process the previous audio of the current audio segment in the model M1 to obtain the feature V2, which expresses the depressive change characteristics of the previous audio. Model M2 uses feature V2 to fine-tune model V1. The two sets of RMSE and MAE in model M3 were 8.32, 7.14, and 8.56, 7.32, which were higher than M2. In model M3, feature V3 is selected to jointly optimize model M2. The experimental results prove that the DR AudioNet network in this paper can indeed improve the classification performance.

TABLE 5. Comparison results of the three models on the AViD-Corpus and DAIC-WOZ test sets for depression identify.

Datasets	Models	RMSE	MAE
AViD-Corpus	M1	9.02	7.78
	M2	8.64	7.45
	M3	8.32	7.14
DAIC-WOZ	M1	10.21	7.96
	M2	9.46	7.53
	M3	8.56	7.32

Figures 6 and 7 are the loss function change curves of the three models on the AViD-Corpus and DAIC-WOZ data sets, respectively.

From Figure 6 and 7, the loss function of model M3 converges faster on the two data sets than other models. The experimental results of these three models show that the

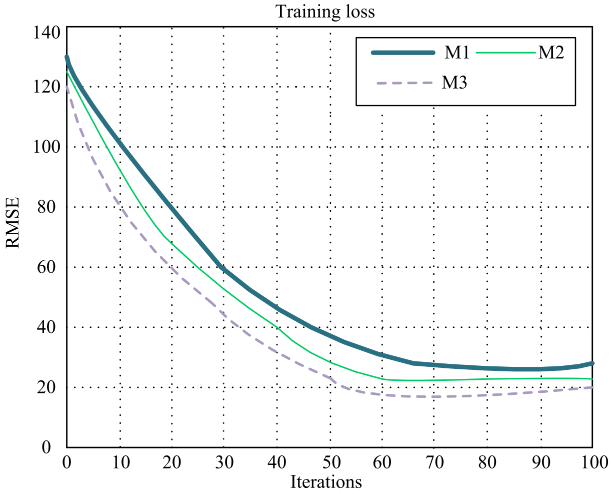

FIGURE 6. Curve of loss function trained by three models on AViD-Corpus dataset.

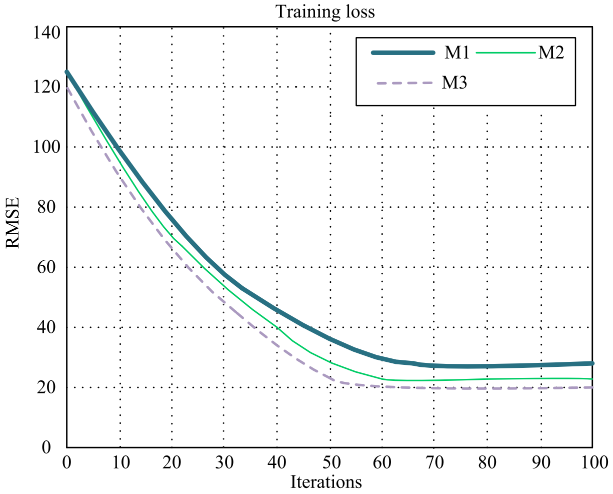

FIGURE 7. The loss function variation curve of three models trained on DAIC-WOZ Dataset.

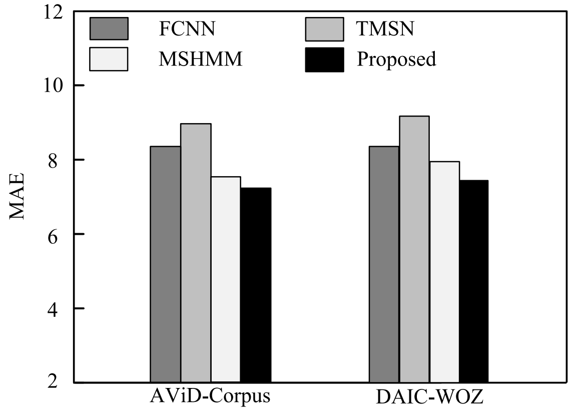

FIGURE 8. Comparison of MAE results of different models on AViD-Corpus and DAIC-WOZ datasets.

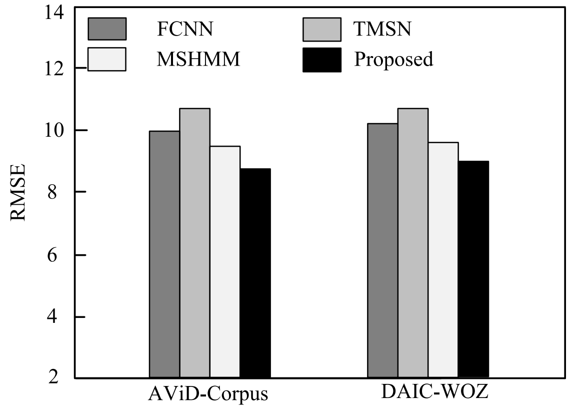

FIGURE 9. Comparison of RMSE results of different models on AViD-Corpus and DAIC-WOZ datasets.

proposed method uses the multi-scale audio features of two adjacent audio segments can jointly optimize the network model, which furtherly reduce the audio depression recognition error, and effectively integrates the audio features to the speaker's non-personalized. The depression features are more conducive to regression prediction of depression recognition models.

D. COMPARISON WITH CURRENT ADVANCED ALGORITHMS

To verify the superiority of the proposed depression regression prediction model based on convolutional neural networks and generative adversarial networks, we conduct experiments on the FCNN algorithm proposed in [21], the TMSN algorithm proposed in [23], and the MSHMM algorithm proposed in [24]. The experimental results are shown in Figures 8 and 9.

From Figure 8 and 9, compared with several existing methods, the proposed method effectively reduces the depression recognition error. The mean values of RMSE and MAE on the two data sets of AViD-Corpus and DAIC-WOZ are (8.36, 7.15) and (8.58, 7.35) which are high more than 5% with the comparison algorithms. Existing methods have poor learning ability when the samples are unbalanced, which leads to high MAE and MSE values of the model. The proposed method uses the multi-scale audio differential normalization algorithm to select the corresponding speech frame according to the sliding window to extract the MFCCs, short-term energy and spectrogram entropy features of the training data set to normalize at different scales. Based on the representative features, the model can further improve the classification accuracy.

VI. CONCLUSION

This paper proposed a novel deep learning model to obtain non-personalized depression features (differential normalization features) of two adjacent sections of local audio, and extract speech signals such as Mel frequency cepstrum coefficients MFCCs, short-term energy and spectrogram entropy features. The features based on audio timing changes reflect the speaker's audio change information, reduce the speaker's personalized speech characteristics, and show a strong correlation with the BDI-II value. With the combination of CNN and GAN, DR AudioNet is used to build a depression recognition model. Furtherly, the algorithm uses DR AudioNet to optimize the previous model through the normalized features of the two adjacent segments of the current audio segment. During the optimization process, our model can extract the most discriminative features and optimize itself, which improve the classification accuracy and BDI-II values.

Audio depression recognition has broad application prospects in the field of mental health. Although the proposed audio depression recognition method using convolutional neural network and generative adversarial network model has obtained good recognition results, while the choice of speech segments has a great influence on the final recognition results. Therefore, in future work, we will explore text processing in natural language processing. Firstly, analyzing the text information of individuals answering questions, and then fuse the obtained speech features and text features to construct more robust and discriminative model.

REFERENCES

[1] V. Vitriol, A. Cancino, C. Serrano, S. Ballesteros, and S. Potthoff, "Remission in depression and associated factors at different assessment times in primary care in Chile," *Clin. Pract. Epidemiol. Mental Health*, vol. 14, no. 1, pp. 78–88, Mar. 2018.

[2] C. Snell-Rood, R. Merkel, and N. Schoenberg, "Negotiating the interpretation of depression shared among kin," *Med. Anthropol.*, vol. 37, no. 7, pp. 538–552, Oct. 2018.

[3] J. Fjermestad-Noll, E. Ronningstam, B. Bach, B. Rosenbaum, and E. Simonsen, "Characterological depression in patients with narcissistic personality disorder," *Nordic J. Psychiatry*, vol. 73, no. 8, pp. 539–545, Nov. 2019.

[4] S. K. Rajput, O. Matoba, Y. Awatsuji, "Holographic multi-parameter imaging of dynamic phenomena with visual and audio features," *Opt. Lett.*, vol. 44, no. 4, pp. 995–998, 2019.

[5] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[6] L. Polyanskaya, A. G. Samuel, and M. Ordin, "Regularity in speech rhythm as a social coalition signal," *Ann. New York Acad. Sci.*, vol. 1453, no. 1, pp. 153–165, Oct. 2019.

[7] D. Habasinska, E. Skrodzka, and E. Bogusz-Witczak, "Development of the polish speech test signal and its comparison with the international speech test signal," *Arch. Acoust., J. Polish Acad. Sci.*, vol. 43, no. 2, pp. 253–262, 2018.

[8] S. Yao, Y. Zhao, A. Zhang, S. Hu, H. Shao, C. Zhang, L. Su, and T. Abdelzaher, "Deep learning for the Internet of Things," *Computer*, vol. 51, no. 5, pp. 32–41, May 2018.

[9] R. Tang, H. Liu, J. Wei, and W. Tang, "Supervised learning with convolutional neural networks for hyperspectral visualization," *Remote Sens. Lett.*, vol. 11, no. 4, pp. 363–372, Apr. 2020.

[10] T.-F. Zhang, P. Tilke, E. Dupont, L.-C. Zhu, L. Liang, and W. Bailey, "Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks," *Petroleum Sci.*, vol. 16, no. 3, pp. 541–549, Jun. 2019.

[11] X. Mao, J. Zhao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. control*, vol. 47, pp. 312–323, Jan. 2019.

[12] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 271–280.

[13] J. Monge-Álvarez, C. Hoyos-Barceló, K. Dahal, and P. Casaseca-de-la-Higuera, "Audio-cough event detection based on moment theory," *Appl. Acoust.*, vol. 135, pp. 124–135, Jun. 2018.

[14] K. Wu, D. Zhang, G. Lu, and Z. Guo, "Joint learning for voice based disease detection," *Pattern Recognit.*, vol. 87, pp. 130–139, Mar. 2019, doi: 10.1016/j.patcog.2018.09.013.

[15] S. Ntalampiras and I. Potamitis, "Transfer learning for improved audio-based human activity recognition," *Biosensors*, vol. 8, no. 3, p. 60, Jun. 2018, doi: 10.3390/bios8030060.

[16] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Commun.*, vol. 99, pp. 80–89, May 2018.

[17] S. Souli and Z. Lachiri, "Audio sounds classification using scattering features and support vectors machines for medical surveillance," *Appl. Acoust.*, vol. 130, pp. 270–282, Jan. 2018.

[18] H. Mukherjee, S. M. Obaidullah, K. C. Santosh, S. Phadikar, and K. Roy, "Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal," *Int. J. Speech Technol.*, vol. 21, no. 4, pp. 753–760, Dec. 2018.

[19] E. Spyrou, R. Nikopoulou, I. Vernikos, and P. Mylonas, "Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms," *Technologies*, vol. 7, no. 1, p. 20, Feb. 2019, doi: 10.3390/technologies7010020.

[20] F. J. Ibarrola, L. E. Di Persia, and R. D. Spies, "A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation," *Signal Process., Off. Publication Eur. Assoc. Signal Process.*, vol. 151, pp. 89–98, Oct. 2018.

[21] Y. Kim, M. Kim, J. Goo, and H. Kim, "Learning self-informed feature contribution for deep learning-based acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2204–2214, Nov. 2018.

[22] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *Proc. Interspeech*, Sep. 2018, vol. 9, no. 2, pp. 937–940.

[23] E. Bavu, A. Ramamonjy, H. Pujol, and A. Garcia, "TimeScaleNet: A multiresolution approach for raw audio recognition using learnable biquadratic IIR filters and residual networks of depthwise-separable one-dimensional atrous convolutions," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 220–235, May 2019, doi: 10.1109/JSTSP.2019.2908696.

[24] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Int. J. Speech Technol.*, vol. 42, no. 4, pp. 722–737, Jun. 2015.

[25] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-start single-stage discriminatively trained HMM-based models for ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 1949–1961, Nov. 2018.

[26] K. Omori, H. Kojima, R. Kakani, D. H. Slavit, and S. M. Blaugrund, "Acoustic characteristics of rough voice: Subharmonics," *J. Voice*, vol. 11, no. 1, pp. 40–47, Mar. 1997.

[27] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[28] P. Somervuo, "Time–frequency warping of spectrograms applied to bird sound analyses," *Bioacoustics*, vol. 28, no. 3, pp. 257–268, May 2019.

[29] A. A. Nagra, F. Han, Q. H. Ling, M. Abubaker, F. Ahmad, S. Mehta, and A. T. Apasiba, "Hybrid self-inertia weight adaptive particle swarm optimisation with local search using C4.5 decision tree classifier for feature selection problems," *Connection Sci.*, vol. 32, no. 1, pp. 16–36, Jan. 2020.

[30] M. Hrabina and M. Sigmund, "Audio event database collected for gunshot detection in open nature (GUDEON)," *J. Audio Eng. Soc.*, vol. 67, nos. 1–2, pp. 54–59, Jan. 2019.