Northumbria Research Link

Northumbria
University
NEWCASTLE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 4, 2020, accepted May 3, 2020, date of publication May 12, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994124

Attention-Block Deep Learning Based Features Fusion in Wearable Social Sensor for Mental Wellbeing Evaluations

JIKUN JIN[1], BIN GAO[1], (Senior Member, IEEE), SIHAO YANG[1], BINGMEI ZHAO[1,2], LIZHU LUO[2], AND WAI LOK WOO[3]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]MOE Key Laboratory for Neuroinformation, Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8SB, U.K.

Corresponding authors: Bin Gao (bin_gao@uestc.edu.cn) and Lizhu Luo (834191311@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 31800961, in part by Sichuan Science and Technology Program under Grant 2018JY0361, and in part by China Postdoctoral Science Foundation under Grant 2018M633336.

ABSTRACT With the progressive increase of stress, anxiety and depression in working and living environment, mental health assessment becomes an important social interaction research topic. Generally, clinicians evaluate the psychology of participants through an effective psychological evaluation and questionnaires. However, these methods suffer from subjectivity and memory effects. In this paper, a new multi- sensing wearable device has been developed and applied in self-designed psychological tests. Speech under different emotions as well as behavior signals are captured and analyzed. The mental state of the participants is objectively assessed through a group of psychological questionnaires. In particular, we propose an attention-based block deep learning architecture within the device for multi-feature classification and fusion analysis. This enables the deep learning architecture to autonomously train to obtain the optimum fusion weights of different domain features. The proposed attention-based architecture has led to improving performance compared with direct connecting fusion method. Experimental studies have been carried out in order to verify the effectiveness and robustness of the proposed architecture. The obtained results have shown that the wearable multi-sensing devices equipped with the attention-based block deep learning architecture can effectively classify mental state with better performance.

INDEX TERMS Mental health assessment, wearable device, attention-based feature fusion.

I. INTRODUCTION

Mental health evaluation is an important topic for human safety analysis. Wearable device, acquiring data of related social-speech and behavioral activity, provides a new approach to understand mental health better by establishing the interrelationships of Social Signal Processing (SSP) and Physical Mental Health (PMH). Traditional methods had been proposed to measure and evaluate social behavior. However, they are of limited effectiveness for continuous monitoring of mental health. The key point is to use a more comprehensive analysis by combining multi-sensor features available from the wearable device. These features assist to determining the potential relationship between human activities and mental health. Moeslund *et al.* [1] summarized technologies in in automatic visual analysis of human behavior including automatic initialization, tracking, pose estimation, and movement recognition. However, these technologies have many restrictions in daily life and equipments are expensive. Thus, sensor-based social signal processing has become an active research topic [2]–[4] which attracted researchers on the relationship generation between the multi-sensor data and healthcare. In [5], Pentland proposed the wearable intelligent devices which was developed to objectively sense and gain an understanding of human wellbeing. In order to capture social signals with high quality, reliability, and validity, the first priority is to create an appropriate collection environment or experiment. Long-term wellbeing monitoring [5] is able to achieve high accuracy for analyzing long-term daily behaviors for human. Long-term monitoring requires the expenditure of a long duration and this results in significant challenges in recruiting and retaining sufficient number of

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

participants. In addition, there is a need to protect the privacy of the participants [6]. The use of wearable devices in short-term for targeted psychological tests is a possible solution to offer an efficient and low-cost method to analyze social signals for mental wellbeing monitoring.

The application of machine learning and deep learning algorithms in wearable devices is crucial. In most wearable devices, they extract 6-axis behavior data in the classification of complex movements such as gestures or dances [7], [8]. In addition, by fusing with speech and behavioral features, it is possible to design wearable devices with machine learning algorithms for monitoring mental health wellbeing. Efficient speech segmentation and classification methods help to analyze social audio. Audio features mainly include Mel-frequency cepstral coefficients (MFCCs) and spectral features. Log-mel spectrograms are used as audio features, which can be processed by using image classification and segmentation model [9], [10]. Speech classification methods can be divided into supervised and unsupervised models. Unsupervised models include Hidden Markov Models (HMMs)[11], Gaussian Mixture Models (GMMs)[12], and Nonnegative Matrix Factorization (NMF) [13], [14] which have advantages of fast computations and do not require human annotation for the data. In recent years, deep learning model significantly improves the classification performance despite the long-duration training process. For instance, convolution neural networks (CNNs) can extract high level speech features and achieve high classification accuracy by using spectrogram [15], [16]. Another network with high performance accuracy in audio classification is the Long Short-Term Memory (LSTMs) which is a variant of recurrent neural network with good results in analyzing time series signals. Chernykh *et al.* [17] achieved emotion classification by using LSTM, and Han *et.al* [18] built a LSTM network through the DenseNet structure to further improve the accuracy. Deep learning model often requires large datasets while the annotation is a complicated task [19]. Transfer learning [20], [21] enables the deep model to perform better in a small datasets. The model can first learn abundant information in a large public dataset and then fine-tuning in the small target dataset. Transfer learning achieves remarkable results in natural language processing [22] and image classification [23].

For multi-sensor wearable devices, speech pattern is one of the most effective cues for analyzing mental health. This is usually accomplished by speech segmentation from the wearable users. However, single speech segmentation has severe limitations. It does not comprehensively consider the relationship between speech signals under various emotions nor can it relates to behavioral data such as natural limb movements under stress. Thus, multi-sensor data is considereds as a way forward in assisting speech segmentation to further enhance the classification accuracy. Appropriate feature fusion method or model can effectively fuse different categories of features and learns the intrinsic association of different features. Chen *et al.* [24] constructed a deep feature fusion model for CTR(Click-Through Rate) prediction whereby they fused image features with one hot features and obtained good performance. Yu *et al.* [25] proposed a model to fuse deep learning and traditional image features which yielded better results than single CNNs. Janani and Ramanan [26] presented a feature fusion framework to connect traditional Bag-of-Features and CNN features in the object classification task. Feature fusion method achieves good performance in processing speech data. Hasan *et al.* [27] proposed an audio-visual feature fusion via deep neural networks and implemented speech recognition with low error rate. In addition, the audio-visual feature fusion was used to recognize lip language [28]. Xu *et al.* [29] constructed the deep model which fused MFCCs and spectrograms, and resulted in high score in the DCASE-2017 audio scene classification challenge. Therefore, the effective feature fusion method can help to utilize features to improve the classification performance of the classifier.

In this paper, we propose an effective features fusion method that fuses multiple sensor features of the wearable device for mental health evaluation. The contributions can be summarized as follows:

(i) Designing wearable devices with multiple sensors and developing an efficient collection process of the voice and behavioral data for wearer. In addition, we design an objective psychological test for depression/anxiety and recruit participants among the university students. The collected data generates the dataset for training and testing the proposed wearable device.

(ii) Proposing attention-based features fusion block to fuse behavior features and speech features under various emotions. It improves the performance compared with direct connection fusion method. Based on the block, we construct a mental state classifier.

(iii) Presenting and analyzing classification results for depression/anxiety level of participants and exploring the relationship between multi-sensor data and mental health.

The paper is organized as follows, the framework of wearable device, classification and model fusion are presented in Section II. Results and analysis are shown in Section III. Section IV is the conclusion of the paper.

II. IPROPOSED SYSTEM DESCRIPTION

A. DESCRIPTION WEARABLE SOCIAL SENSING PLATFORM AND ANALYSIS FRAMEWORK

The block diagram of wearable social-sensing and data analysis is presented in Figure 1. It indicates the various signals collected by the wearable device and describes how the feature fusion model can be used in the system. The proposed system is illustrated in four parts: (i) audio signal processing, (ii) activity signal processing, (iii) feature fusion system. (iv) prediction and analysis for social sensing results. The wearable device collects audio and activity data. The audio data consists of 5 speech fragments of different emotions for every participant. The system analyzes 5 speech features

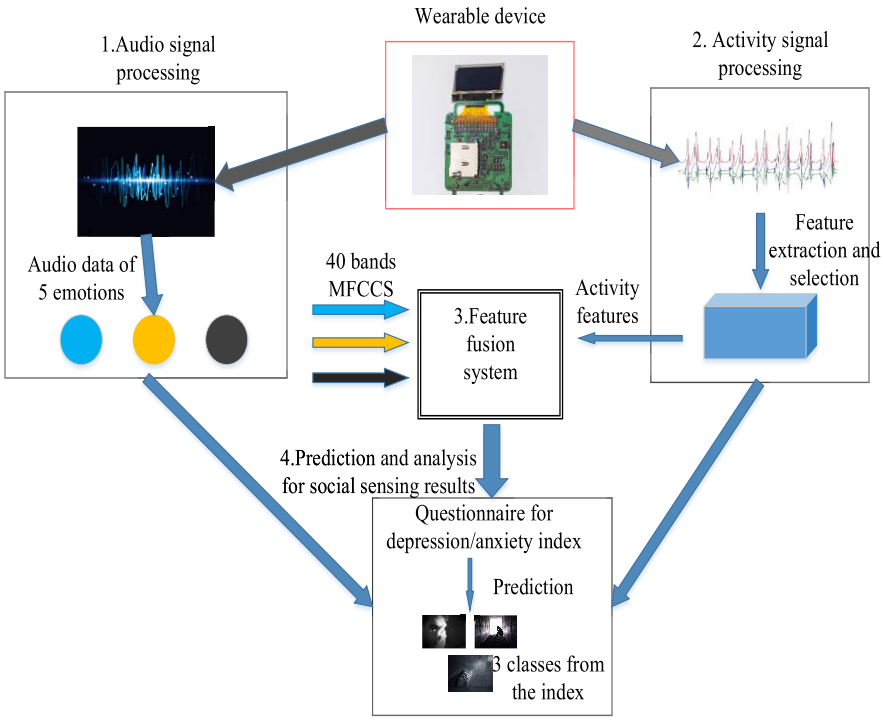


FIGURE 1. Block diagram of the data analysis system.

from different emotions as well as activity features and find their relationship with social sensing results. Finally, it makes a fusion on these features and predicts the level of depression. In addition, the wearable device collects data from participants, performs feature extractions and stores data. The training and prediction process of machine learning model runs on the local server.

The proposed wearable device and its relevant hardware platform is shown in Figure 2. The microprocessor of the wearable device is an ARM-Cortex4 microcontroller with DSP function for audio feature calculation and the model is STM32F405. Besides, the sensors system consists of 6D acceleration and angular sensor (MPU6050), temperature and humidity sensor (SI7021) to collect multi-modal data from the environment, physiological signals and behavioral activity. The sampling frequencies of MPU6050 and SI7021 are 100hz and 0.1Hz, respectively.

FIGURE 2. Hardware platform of the self-designed wearable device.

The audio collecting system contains MEMS microphones as well as audio code unit (WM8978), the audio signals collected and amplified by an inter-integrated-circuit (I2C) bus with 8 KHz sampling frequency. The display module is an OLED screen. In order to record large amounts of data, the wearable device contains power management unit with a 2200mAh lithium battery and a micro SD card.

B. DESCRIPTION EXPERIMENT PROTOCOL AND SOCIAL DATA ACQUISITION

The dataset is collected from an autobiographical memory test which involves the participation of 60 students (30 males and 30 females; age range $= 18 - 26$) at the University of Electronic Science and Technology. All students signed informed consent before the experiment and we have signed a confidentiality agreement with the participants on their speech content. Prior to the experiment, the level of depressive symptoms and state as well as trait anxiety of all participants were assessed by using the Beck Depression Inventory (BDI-II) [30], and State-Trait Anxiety Inventory (SAI, TAI) [31]. The scores of the questionnaires were used to calibrate the data. For the autobiographical memory test, the participants were initially asked to think of six specific events for each emotion (happy/angry/sad/fearful/neutral) that had happened to them rather than being told by others. Meanwhile,

the participants can write them down for each event to give a clue for the following recording session. During the experiment, the participants were shown the prompt words for 30s during which they verbalized the events coupled with emotion as specifically as possible.

During data collection, the wearable device is worn at the preferred wrist to collect the voice as well as the behavioral data of the participants. The behavior in the experiment is not a specific movement, it behaves as a hand swing of the participant during the experiment and this movement may be unconscious. In addition, in order to prevent the participants from being disturbed, each subject was tested alone in a quiet room. To prevent potential bias, all experimental procedures were guided by computer programs. The collection of the wearable device is synchronized to the clock of the computer, which allows us to effectively timestamp the collection data. After the experiment, we extracted the speech and behavioral data of the wearable device. In this case, we collected 30 pieces of speech and behavioral data for each subject (six speaking fragments for each emotion). The dataset of the paper is composed of speech and behavioral data of the 60 participants and the ground truth is the scores of the questionnaires. The specification of the dataset is shown on Table 1. Temperature and humidity sensors are embedded in wearable devices, whereas they are not used because the experimental environment for collecting data is fixed and the time is short. Thus, the environmental sensing data changes little.

TABLE 1. Description of dataset.

Data type	Sampling rate	Data fragment length	Data size
Speech signal data	8000Hz	30s	60×5×6
6-axis behaviour data	100Hz	30s	60×5×6

There are two main limitations with the experiment. Each emotion is generated by recalling a specific event instead of generation in the natural state. Besides, all participants are university students with age range from 18 to 26 and our experiments do not cover wider age groups. Thus, at the initial evaluation, our experiments only focus on these age group. Therefore, the impact should be drawn that our system can only test on data of persons on these age groups.

C. DESCRIPTION AUDIO FEATURES AND BEHAVIOR FEATURES

In the experiment, the speech data is collected by microphones of the wearable device, it is grouped into data segments of 30s with 8kHz sampling frequency. The behavioral data is obtained through 6-axis sensor in which it consists of three axes acceleration data and three axes angular velocity data. This 6-axis data is used to extract behavioral features and they are calculated from the sliding window. The time interval between the sliding windows is 3 seconds. In addition, the features are divided into time domain features and frequency domain features as listed in Table 1.

The input audio features of the network are Mel Frequency Cepstral Coefficients (MFCCs) [32] which mimic the human auditory system. Firstly, the audio signal is divided into several frames with 512-points and take the Short-Time Fourier Transform (STFT) of each frame. It then maps the power of the spectrum onto the Mel scale and take the discrete cosine transform of the Mel log power. The MFCCs are the amplitudes of the resulting spectrum. Feature extraction operations are conducted by using Librosa [33] which is an open-source library for audio analysis. For speech classification tasks, the raw speech is used as input. However, for the training tasks, the dimensions of the raw speech signal are huge (30s speech segment has 240,000 data points), which cannot be directly used as the input of the network because of the excessive calculation. Furthermore, it is more complicated to learn effective speech features in the network from the raw speech signal since it requires large amount of training data. MFCC is th e commonly used effective speech features and it can be used as input for deep learning model. the features are extracted by framing the raw signal, which reduce the dimension of the input signal.

D. PROPOSED FRAMEWORK OF FEATURE FUSION SYSTEM AND CLASSIFICATION METHOD

1) LSTM BASED NETWORK AND FINE-TUNING METHOD

The basic network of the proposed framework is the LSTM (Long Short-Term Memory). The effect of LSTM on time series learning is profound. A significant attribute of LSTM is the ability to map from the entire history of inputs to each output [34]. Besides, LSTM solves the vanishing gradient and context access problems commonly plague the RNN [35], [36]. The basic unit of the LSTM architecture consists of a memory block with different types of memory cells and three adaptive multiplications named input gate, forget gate as well as output gate. LSTM contains information outside the normal flow of the RNN in a gated cell, which helps to avoid the vanishing gradient problem. The training loss of LSTM can be back-propagated through time and layers.

The audio data can be divided into time series segments and each segment has 470-time steps. The proposed model makes use of a multi-layered LSTM structure to extract high level emotional features on time steps.

The network with 3 LSTM blocks is used to process audio features. However, the size of collected data is small and hence this presents difficulties in sufficiently extracting the bottom layer features. Extracting rich features and generalizing bottom layer features are vital to learning more efficient high-level abstract features and improving the network performance as well as robustness. Figure 3 shows the audio classification model and fine-tuning method. The EMO-DB audio dataset [37] is chosen as the source dataset

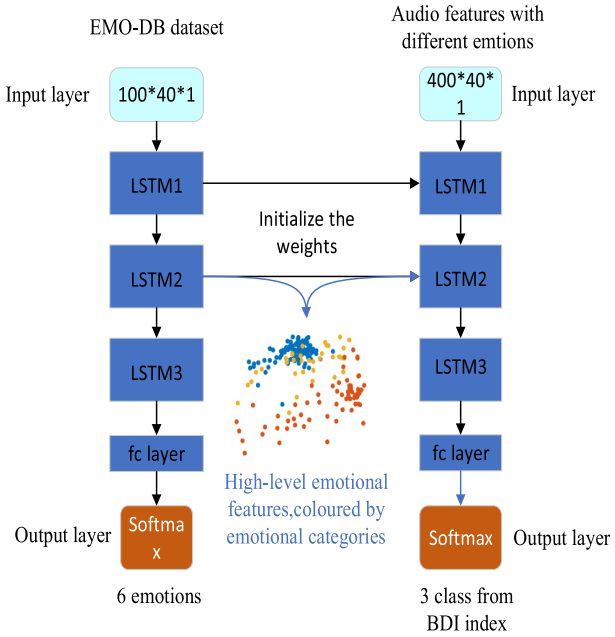


FIGURE 3. Classification model for audio features and fine-tuning method.

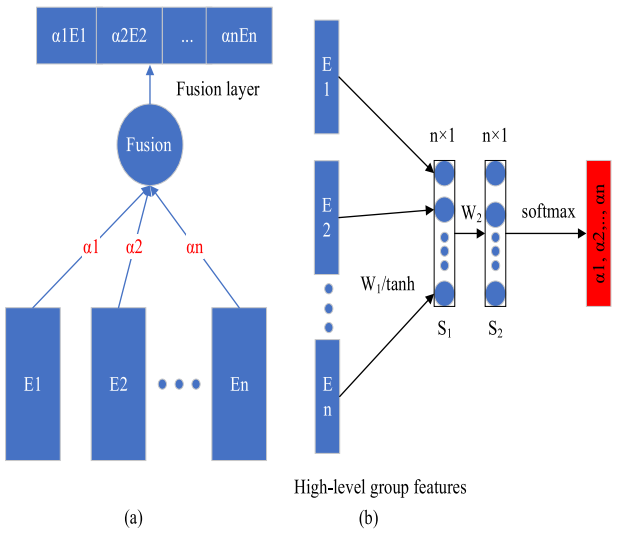


FIGURE 4. Working process of the attention block.

as it consists of audio segments with 6 labeled emotions and our experiment audio data also contains different emotions. Thus, the source task refers to emotions classification and training this task helps the network to learn more of the basic emotional low-level features. In implementation, the weights of the first two LSTM layers are initialized by using source task training weights while other further layers need to be retrained.

For the main task, the input of the network is the audio features of single emotion, which is the 40-length MFCCs with 470-time steps. In the first step, the model is trained on audio features of each emotion separately to analyze and compare the classification results of the speech under each emotion.

2) PROPOSED ATTENTION-BASED FEATURE FUSION

Commonly used feature fusion methods include weighting method as well as direct connecting method. For weighting method, finding the weight value for different features is the crucial part. However, determining the optimal weights combination is a difficult task. In order to effectively integrate the features of different emotions, we designed an attention block to produce better combination weights and this block can make the model focus on relevant emotions.

Attention mechanism was used in the transformer model [38] and word encoder model [39] where this mechanism enables the model to exercise attention to the more related word vectors in the translation task, while reducing the attention to unrelated word vectors. Thus, the attention mechanism can be used in feature fusion tasks as it enables the model to focus on important features. For our task, the attention layer enables the proposed deep learning network to concentrate on emotional features with different weights, and construct a model to analyze the relationship between emotions and mental health.

Specifically, as shown in Figure 4(a), the vectors $E_1, E_2, \ldots, E_n$ represent different groups of emotional features. The weighted features fusion layer F is computed as concatenated weighted fusion of these group features where the weights $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ are computed in a method illustrated in Figure 4(b). The weights vector A is calculated as

$$\begin{cases} S_1 = \tanh(EW_1)^T \\ S_2 = S_1 W_2 \\ A = softmax(S_2) \end{cases} \tag{1}$$

where $E = [E_1^T; E_1^T; \ldots; E_1^T]$ with a shape of $d \times n$, W_1 is a vector of parameters with size $d \times 1$ and the size of W_2 is $n \times n$. S1 and S2 are the middle layers, which are composed of n neurons. The *softmax*() function ensures all the computed weights sum up to 1. After obtaining the fusion weights, the fusion layer F can be represented as

$$layerF = \{\alpha_1 E_1, \alpha_2 E_2, \ldots, \alpha_n E_n, \tag{2}$$

3) OVERALL FEATURE FUSION SYSTEM

Figure 5 shows the fusion model for audio features and behavior features. The input of the model can be divided into 2 fields. In the first field, 17 selected behavior features are inputted to DNN network to extract high-level behavior features. In the second field, five groups of emotional speech features are put into the LSTM-based network and the weights of the first two LSTM layers are initialized by using the method of Figure 3. F1 is the weighted fusion layer which fuses 5 groups of high-level speech features through the attention block 1. Besides, the combined LSTM features and high-level DNN features are concatenated on the layer F2 by attention block 2.

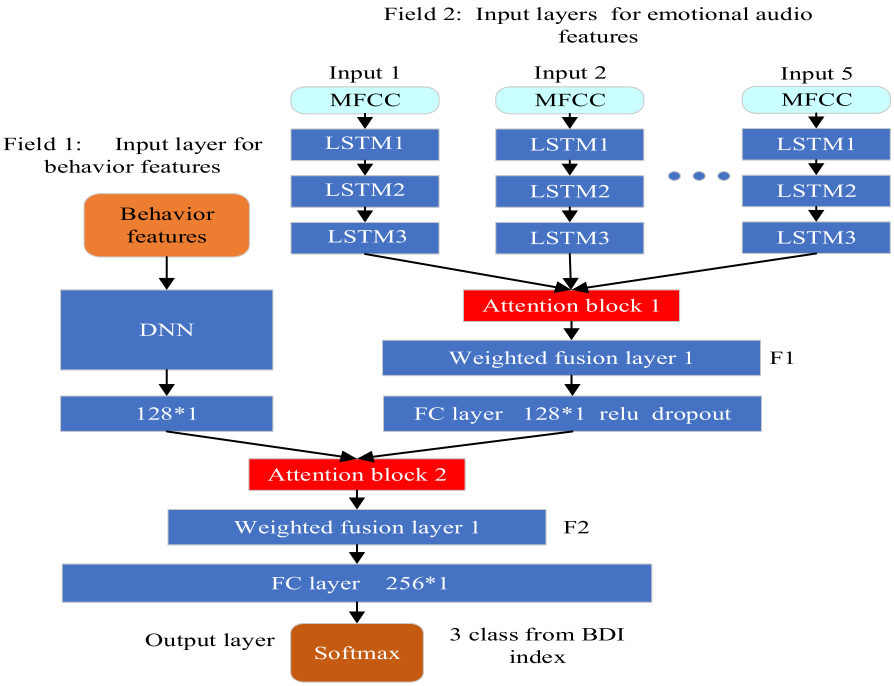


FIGURE 5. The overall architecture of the fusion model.

The loss function of the model is categorical cross entropy function. The optimizer of training is Adam (Adaptive Moment Estimation) and the initial learning rate is 0.00001 while the batch size is 16. In order to avoid overfitting, we set early stopping and its patience is 15 epochs. This means that the training will automatically stop if the accuracy of the validation set is not improved in 15 epochs. The model for single audio features is trained in GTX 1070 and the fusion model is trained in GTX1080Ti. The total training time of 5-fold cross-validation is approximately 18 hours..

III. EVALUATION AND ANALYSIS

A. EXPERIMENTAL RESULTS

In this experiment, $F_1 - Score$ is used to measure the prediction results. It considers the Precision and Recall at same time and can be regarded as a weighted average of the Precision and Recall. Thus, $F_1 - Score$ helps objectively analyze the performance of the classifier.

Precision and Recall are defined as:

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

where tp is true positive, fp is false positive, fn is false negative. Precision is also referred to as positive predictive value (PPV), and it is the fraction of correctly predicted positive samples to the total predicted positive samples. Recall is also referred to as the true positive rate or sensitivity, it can be represented as the fraction of correctly predicted positive samples to the total positive samples in actual label.

The $F_1 - score$ is calculated by using Precision and Recall with same weight:

$$F_1 - score = 2\frac{Recall \times Precision}{Recall + Precision} \tag{5}$$

All test results are obtained using a 5-fold cross-validation strategy which balances the training accuracy of each round and the total training time. The overall performance is computed by averaging the results from all 5 iterations. 20% of the data is used as a test set for each iteration. Our cross-validation method is similar to the subject cross validation[40]. The method of dividing the training set and validation set is shown in Figure 1. The training and testing sets are split by subject. The dataset contains speech and behavior data for 60 subjects and the raw data are stored in different folders with subjects' numbers. In each iteration of cross-validation, the training set and validation set are divided by this number. For example, the data with numbers 01 to 48 is the training set, and the data with 49 to 60 is the test set. Thus, the training set and validation set are independent of each other and the data of one subject may only be in the training or test set. Overlapping windows are only used in feature extraction process. We performed feature extraction after dividing training set and validation. Thus, the training features and validation features are extracted separately. Therefore, there is no overlapping data between training set and validation set.

The results of the source task are shown in Figure 6. The pre-train task is an emotion classification for EMO-DB audio dataset and training on task can initialize the weights of the LSTM based network. CNN-based networks are usually used

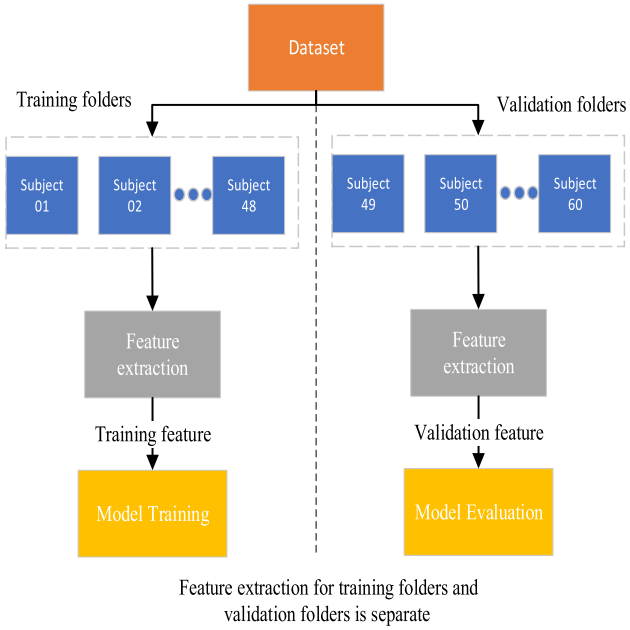


FIGURE 6. The method of dividing the training set and validation set.

for speech classification. The figure compares classification performance of the LSTM-based model and a powerful CNN-based network: VGG-net. The input of VGG-net is the same as the input of LSTM since they are both 40-length MFCCs. In addition, the time step is zero padded to 40. Thus, the input size of VGG-net is 40×40 (since the time step of each speech segment in EMO-DB is less than 40). It is illustrated that the LSTM-based model has relatively better performance and the average $F_1 - Score$ of LSTM-based model is 8.7% higher than VGG-based model. The main reason is that LSTM block extracts features of time dimension more efficiently for short-term sequences whereas CNN network with deeper layer is difficult to train for small datasets.

Our project dataset description and collection experiment has been covered in the Section II. The classification index of the dataset in our experiment is based on three questionnaires of BDI assessing depression, SAI assessing state anxiety, and TAI assessing trait anxiety, which avoid the contingency of individual indicator results. Besides, we have divided the data into three labels for every index, the 27% lowest scores are the low class of depression, the 27% highest scores are the high class and another 46% middle scores are the middle class. 27% is a common criterion for dividing the ratio of high and low in psychological experiments. This method is named as high-low-27-percent group method [41].

Tables 2 to 4 show the F1-score results of BDI, SAI and TAI class, respectively. The input is the audio features with single emotions and these tables compare the classification performances for audio input with different emotions. Figure 7 shows the average F1-scores for these three indices. The result reveals that the accuracy of depression and anxiety classification under the emotions of happiness, fear and anger are higher than that under the other two emotions. This rather interesting result shows that depression and anxiety are more easily detected through speech when the emotions of the participants are anger, fear and happy.

TABLE 2. Description for analysis of activity features.

Class	ID	Feature	Description
Time domain	1	Mean	Average value of samples in a window
	2	STD	Standard deviation of samples
	3	Minimum	Minimum of samples in a window
	4	Maximum	Maximum of samples in a window
	5	Energy	The energy of samples in a window
	6	Variance	Variance of samples in a window
	7	Range	Maximum minus minimum
	8	Entropy	Information entropy of signal in a window
Frequency domain	9	DC	Direct component of a FFT window
	10-13	Amplitude Features	

TABLE 3. F1-score results for BDI level under one emotion channel audio features.

Class	Emotions of the input audio features				
	Anger	Fear	Sad	Neutral	Happy
Low	0.570	0.562	0.573	0.633	0.660
Median	0.761	0.669	0.675	0.631	0.685
High	0.723	0.698	0.540	0.690	0.762
Average	0.685	0.643	0.596	0.651	0.702

TABLE 4. F1-score results for SAI level under one emotion channel audio features.

Class	Emotions of the input audio features				
	Anger	Fear	Sad	Neutral	Happy
Low	0.526	0.614	0.608	0.488	0.576
Median	0.781	0.712	0.735	0.705	0.774
High	0.669	0.674	0.533	0.571	0.629
Average	0.658	0.667	0.625	0.588	0.660

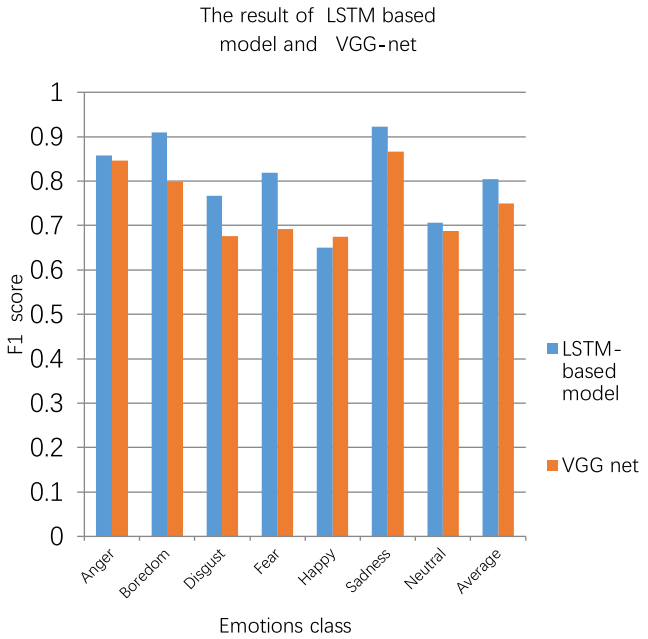


FIGURE 7. The classification results of pre-train dataset.

In this paper, the attention block has been designed to produce the appropriate combination weights for the emotion features as well as behavior features and to enable the model focus on more relevant features. In addition, we compared the fusion method based on attention block with the direct connection method which uses same weights on features.

Figure 8 compares the average F1-scores of these two fusion methods. It is seen that attention block can improve the classification accuracy because better dynamic combination weights are obtained by training the attention model. The detailed results of the two fusion methods will be compared in Table 5.

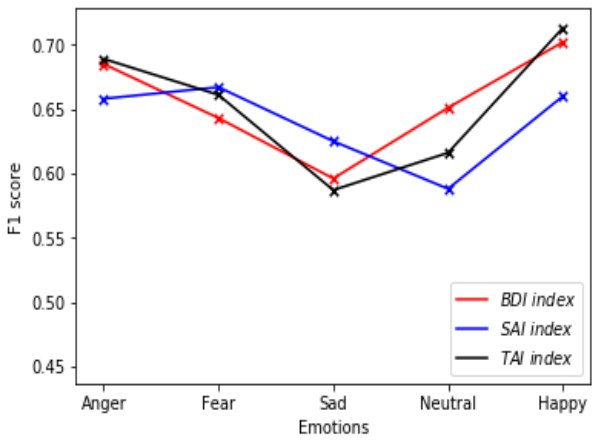


FIGURE 8. The average F1-scores under single emotion input.

Figure 9 shows that the fusion weights for speech features under different emotions. This provides a degree of explain ability of the deep learning model which helps to reveal the contributions of various emotional features to mental state

TABLE 5. F1-score results for TAI level under one emotion channel audio features.

Class	Emotions of the input audio features				
	Anger	Fear	Sad	Neutral	Happy
Low	0.624	0.600	0.558	0.672	0.641
Median	0.729	0.753	0.686	0.693	0.770
High	0.716	0.630	0.519	0.684	0.730
Average	0.689	0.661	0.587	0.616	0.713

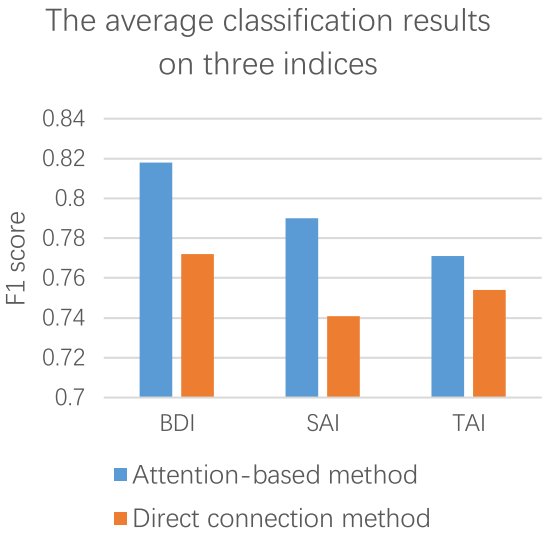


FIGURE 9. Comparison of two fusion methods.

recognition. For prediction of different indices, the generated fusion weights are different whereas several common phenomena can be found. This is clearly visible in Figure 9 that the attention model has highlighted the importance of speech features under emotions of anger and fear in each index prediction. On the separate hand, the speech features under neutral and sadness contribute less to the classification of mental states. Besides, the weight changes little for different mental states.

High-level behavior features and emotional audio features are fused in attention block 2 and their fusion weights in each index are shown in Figure 10. It is seen that the weight of the emotional audio features is much greater than the behavioral features. Therefore, emotional audio features contribute more to the classification of mental states.

The evaluation results of attention-based fusion model and direct connection fusion model are tabulated in Table 5. The accuracy of the fusion models is significantly higher than that of the model under single emotion features. This illustrates the multiple emotional audio features are useful for analyzing mental wellbeing. Besides, the accuracy of attention-based model has obvious improvement compared with direct connection fusion model. Furthermore, the fusion of behavioral features slightly improved the classification performance.

BLSTM (Bi-directional Long Short-Term Memory) is another state-of-the-art learning algorithm for time series classification. Thus, two algorithms are conducted for comparison in both the single and the overall fusion model.

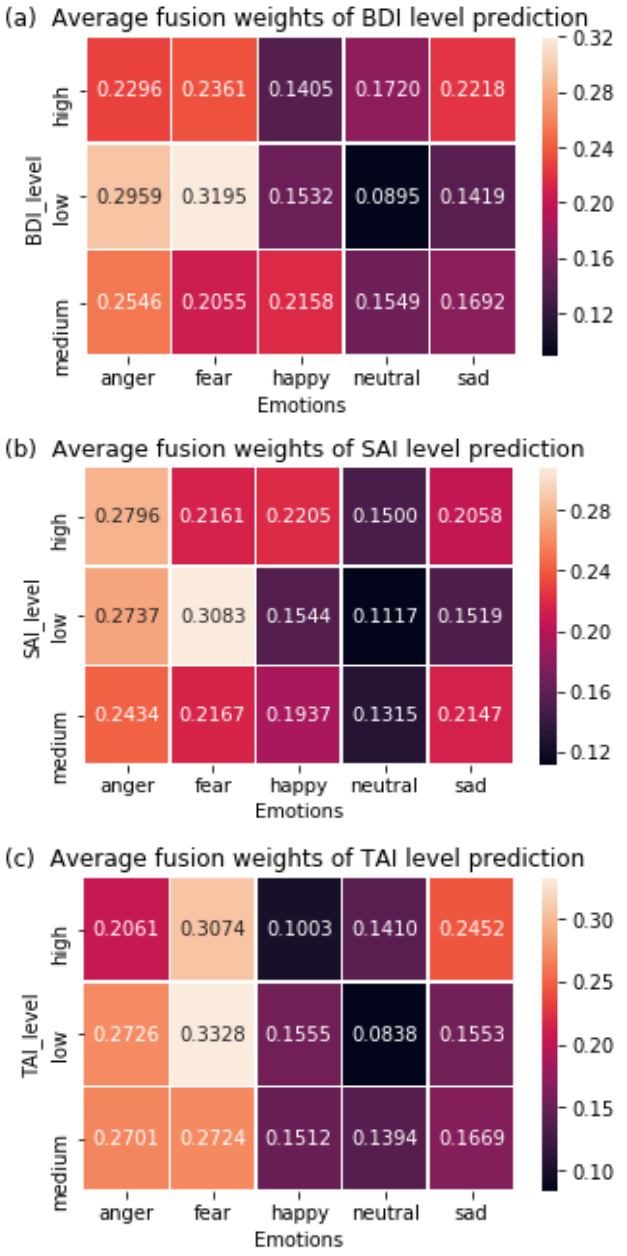


FIGURE 10. Average fusion weights produced by attention block 1 for (a) BDI level prediction, (b) SAI level prediction, (c) TAI level prediction.

Table 1 shows the classification results of LSTM-based model and BLSTM-based model. The input is the speech features with single emotion and the results are the average F1 scores. It is seen that BLSTM-based model slightly improves the classification performance compared with LSTM-based model. Besides, Table 2 compares the results of the overall fusion model with LSTM blocks and BLSTM blocks. It is seen that LSTM blocks and BLSTM blocks have overall similar results for the fusion model. One BLSTM layer is composed of two LSTM layers, the parameter and computational complexity of BLSTM is much greater. Therefore, LSTM is chosen in the overall fusion model.

FIGURE 11. Average fusion weights produced by attention block 2.

TABLE 6. F1-score results for feature fusion model.

Fusion Method	Class	Evaluation Index		
		BDI	SAI	TAI
Direct connection fusion of 5 emotional audio features and behavior features	Low	0.765	0.702	0.695
	Median	0.812	0.795	0.820
	High	0.711	0.688	0.701
	average	0.772	0.741	0.754
Attention-based fusion of 5 emotional audio features.	Low	0.774	0.721	0.704
	Median	0.803	0.833	0.849
	High	0.769	0.695	0.694
	average	0.786	0.765	0.768
Attention-based fusion of 5 emotional audio features and behavior features	Low	0.818	0.747	0.733
	Median	0.839	0.829	0.831
	High	0.783	0.765	0.708
	average	0.818	0.790	0.771

TABLE 7. F_1-score results of the LSTM-based model and the BLSTM-based model.

Single model	Emotion	Evaluation Index		
		BDI	SAI	TAI
LSTM-based model	Anger	0.685	0.658	0.689
	Fear	0.643	0.667	0.661
	Sad	0.596	0.625	0.587
	Neutral	0.651	0.588	0.616
	Happy	0.702	0.660	0.713
BLSTM-based model	Anger	0.721	0.664	0.672
	Fear	0.639	0.675	0.656
	Sad	0.604	0.627	0.581
	Neutral	0.655	0.594	0.611
	Happy	0.694	0.683	0.722

B. ANALYSIS AND DISCUSSION

Through the prediction results of these three indices, the relationship between mental health and multiple-sensor features can be analyzed objectively. Besides, it is shown that the fusion of multiple emotional features and behavioral features contributes to improving the classification accuracy.

TABLE 8. F_1-score results of the overall Fusion model with LSTM blocks and the overall fusion model with BLSTM blocks.

Fusion Method	Class	Evaluation Index		
		BDI	SAI	TAI
Attention-based fusion model with LSTM blocks	Low	0.818	0.747	0.733
	Median	0.839	0.829	0.831
	High	0.783	0.765	0.708
	average	0.818	0.790	0.771
Attention-based fusion model with BLSTM blocks	Low	0.796	0.755	0.679
	Median	0.845	0.838	0.829
	High	0.799	0.761	0.754
	average	0.819	0.795	0.768

We evaluated mental state through wearable devices and deep learning models. This is different from traditional method. The model we used is based on a supervised algorithm which requires automated features extractions and annotation of the data through the labels of the training set. Although the training set is implicitly derived from the questionnaires, these questionnaires are not used in a traditional sense to code up a system. In addition, the prediction process of the proposed system does not depend on the questionnaires.

The LSTM-based network of the fusion model is initialized by using the method. The fine-tuning method is derived from the parameters/model-based transfer learning. The source dataset is the EMO-DB and the task is emotion classification. The target dataset is from our experiments and the target task is the classification of mental states. The two datasets are similar, they are both human speech fragments and the input features are MFCCs. The difference between source and target is that the speech language of the two datasets are different and the classification task is different. According to the theory of transfer learning, the lower layers of the neural network can extract general features, while the specific features are extracted in the higher layer. Therefore, the lower layers are transferable even if there are several differences between source and target. In details, the first two LSTM layers are initialized by the model trained on the EBO-DB, other layers have random initialization and all layers are trainable. In this case, general speech features can be shared and the model converges faster.

IV. CONCLUSION

A wearable device with multiple sensors has been proposed and designed to collect social signals and continuously monitor the mental health status of the wearer. In addition, psychological experiments have been designed to analyze the degree of depression and anxiety. The speech as well as behavioral data have been collected by the wearable devices. By analyzing data and building models from more than 60 participants, the relationship between audio and behavioral features and degree of depression has been established. In particular, three indices of depression and anxiety have validated the proposed detection approach to ensure the objectivity of the results. Attention-based features fusion model has successfully demonstrated to achieve high level of performance accuracy in classifying depression and anxiety levels.

REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Kräger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, Nov. 2006.

[2] A. Pentland, "Social signal processing [exploratory DSP]," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[4] A. Pentland, "Socially aware computation and communication," *Computer*, vol. 38, no. 3, p. 199, 2005.

[5] S. Yang, B. Gao, L. Jiang, J. Jin, Z. Gao, X. Ma, and W. L. Woo, "IoT structured long-term wearable social sensing for mental wellbeing," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3652–3662, Apr. 2019.

[6] Y. Chen, B. Gao, L. Jiang, K. Yin, J. Gu, and W. L. Woo, "Transfer learning for wearable long-term social speech evaluations," *IEEE Access*, vol. 6, pp. 61305–61316, 2018.

[7] M. Xochicale, C. Baber, and M. Oussalah, "Analysis of the movement variability in dance activities using wearable sensors," in *Proc. Wearable Robot., Challenges Trends* 2017, pp. 149–154.

[8] R. Xie and J. Cao, "Accelerometer-based hand gesture recognition by neural network and similarity matching," *IEEE Sensors J.*, vol. 16, no. 11, pp. 4537–4545, Jun. 2016.

[9] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 326–330.

[10] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1547–1554.

[11] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[12] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 881–891, Aug. 2007.

[13] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van Hamme, "An exemplar-based NMF approach to audio event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[14] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 151–155.

[15] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, p. 26, Dec. 2015.

[16] M. Espi, M. Fujimoto, Y. Kubo, and T. Nakatani, "Spectrogram patch based acoustic event detection and classification in speech overlapping conditions," in *Proc. 4th Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, May 2014, pp. 117—121.

[17] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," 2017, *arXiv:1701.08071*. [Online]. Available: https://arxiv.org/abs/1701.08071

[18] K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane, "The CAPIO 2017 conversational speech recognition system," *arXiv:1801.00059*. Accessed: Dec. 1, 2017. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180100059H

[19] A. C. K. J. Han, J. Kim, and I. Lane, "DCASE 2016 acoustic scene classification using convolutional neural networks," presented at the Detection Classification Acoustic Scenes Events, Budapest, Hungary, Sep. 3, 2016.

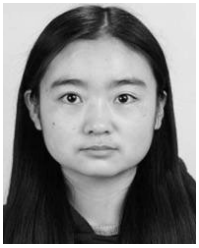
[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[21] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 46–54.
[22] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2006, pp. 120–128.
[23] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[24] J. Chen, B. Sun, H. Li, H. Lu, and X.-S. Hua, "Deep CTR prediction in display advertising," in *Proc. ACM Multimedia Conf.*, 2016, pp. 811–820.
[25] Y. Wang, B. Song, P. Zhang, N. Xin, and G. Cao, "A fast feature fusion algorithm in image classification for cyber physical systems," *IEEE Access*, vol. 5, pp. 9089–9098, 2017.
[26] T. Janani, S. L. JaffnaJaffna, and A. Ramanan, "Feature fusion for efficient object classification using deep and shallow learning," *Int. J. Mach. Learn. Comput.*, vol. 7, no. 5, pp. 123–127, 2017.
[27] M. H. Rahmani, F. Almasganj, and S. A. Seyyedsalehi, "Audio-visual feature fusion via deep neural networks for automatic speech recognition," *Digit. Signal Process.*, vol. 82, pp. 54–63, Nov. 2018.
[28] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.
[29] J. Xu, Y. Zhao, J. Jiang, Y. Dou, Z. Liu, and K. Chen, "Fusion model based on convolutional neural networks with two features for acoustic scene classification," presented at the Detection Classification Acoustic Scenes Events, Munich, Germany, Nov. 16, 2017.
[30] H. Poole, R. Bramwell, and P. Murphy, "Factor structure of the beck depression inventory-II in patients with chronic pain," *Clin. J. Pain*, vol. 22, no. 9, pp. 790–798, Nov. 2006.
[31] T. M. Marteau and H. Bekker, "The development of a six–item short–form of the state scale of the spielberger state—Trait anxiety inventory (STAI)," *Brit. J. Clin. Psychol.*, vol. 31, no. 3, pp. 301–306, Sep. 1992.
[32] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 73–76.
[33] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 1–18.
[34] B. Hammer, "On the approximation capability of recurrent neural networks," *Neurocomputing*, vol. 31, nos. 1–4, pp. 107–123, Mar. 2000.
[35] F. F. J. Informatik, Y. Bengio, P. Frasconi, and J. Schmidhuber, *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*. Hoboken, NJ, USA: Wiley, 2003, pp. 237–243.
[36] J. Schmidhuber, F. Gers, and D. Eck, "Learning nonregular languages: A comparison of simple recurrent networks and LSTM," *Neural Comput.*, vol. 14, no. 9, pp. 2039–2041, Sep. 2002.
[37] (1999). *Berlin Database of Emotional Speech*. [Online]. Available: http://emodb.bilderbar.info/docu/#download?tdsourcetag=s_pcqq_aiomsg
[38] N. S. Ashish Vaswani, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.
[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*. Accessed: Oct. 1, 2018. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D
[40] A. Dehghani, T. Glatard, and E. Shihab, "Subject cross validation in human activity recognition," *arXiv:1904.02666*. Accessed: Apr. 1, 2019. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190402666D
[41] T. Moses, "A review of developments and applications in item analysis," in *Advancing Human Assessment*. Cham, Switzerland: Springer, 2017.

JIKUN JIN received the B.Sc. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2017, where he is currently pursuing the M.Sc. degree in machine learning and intelligent system. His research interests include speech scene recognition, machine translation, and wearable data processing.

BIN GAO (Senior Member, IEEE) received the B.Sc. degree in communications and signal processing from Southwest Jiaotong University, China, in 2005, and the M.Sc. degree (Hons.) in communications and signal processing and the Ph.D. degree from Newcastle University, U.K., in 2006 and 2011, respectively. He worked as a Research Associate with Newcastle University on wearable acoustic sensor technology, from 2011 to 2013. He is currently a Professor with the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include sensor signal processing, machine learning, social signal processing, nondestructive testing, and evaluation, where he actively publishes in these areas. He is also a very active reviewer of many international journals and long standing conferences. He has coordinated several research projects from the National Natural Science Foundation of China.

SIHAO YANG received the B.Sc. degree from the School of Automation Engineering, Northeastern University, Qinhuangdao, China, in 2017. He is currently pursuing the M.Sc. degree in control engineering and science with the University of Electronic Science and Technology of China, Chengdu, China. His research interest includes wearable sensor.



BINGMEI ZHAO received the B.E. degree from the School of Bioinformatics, Chongqing University of Posts and Telecommunications, China, in 2016. She is currently pursuing the M.E. degree in biomedical engineering with the University of Electronic Science and Technology of China, Chengdu, China. Her research interest is autobiographical memory.

LIZHU LUO received the B.Sc. degree in psychology and the B.A. degree in English and the M.Sc. degree in psychology from Southwest University, China, in 2006, 2010, and 2012, respectively, and the Ph.D. degree in biomedical engineering from the University of Electronic Science and Technology of China. Her research interests are psychological behavioral and fMRI studies on social affective disorder.

WAI LOK WOO was born in Malaysia. He received the B.Eng. degree (Hons.) in electrical and electronics engineering and the Ph.D. degree from Newcastle University, U.K. He is currently a Senior Lecturer and the Director of Operations with the School of Electrical and Electronic Engineering. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of machine learning for signal processing, blind source separation, multidimensional signal processing, signal/image deconvolution, and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. He has published over 250 articles on these topics on various journals and international conference proceedings. He was awarded the IEE Prize and the British Scholarship to continue his research work. He is an associate editor of several international journals and has served as the lead editor for journals' special issues.