Improvement on Speech Depression Recognition Based on Deep Networks

1st Jinming Li
College of Information Engineering
Capital Normal University
Beijing, China
ljmch7@163.com

2nd Xiaoyan Fu
Beijing Key Laboratory of Electronic
System Reliability Technology
Capital Normal University
Beijing, China
fuxiaosg@cnu.edu.cn

3rd Zhuhong Shao
Beijing Advanced Innovation Center
for Imaging Technology
Capital Normal University
Beijing, China
shaozh2015@163.com

4th Yuanyuan Shang
Beijing Engineering Research Center of High Reliable Embedded System
Capital Normal University
Beijing, China
syy@bao.ac.cn

***Abstract*—To reduce the burden of clinicians diagnosing a large number of depressive symptoms, the field of artificial intelligence researchers are increasingly interested in designing automatic recognition systems for depression. Depressed patient have different speech signal from normal people. Here, we present a deep model, Depression AudioNet, which encodes depression-related features in the vocal tract and provides a more comprehensive audio representation. Firstly, the Mel-frequency cepstral coefficients (MFCCs) were extracted from raw audio data. Secondly, the robust emotions features were acquired by Multiscale Audio Delta Normalization (MADN), which is a data processing algorithm we proposed. Finally, the MFCCs and the emotions features of two adjacent segments of local audio were fed into the Depression AudioNet in turn to train the network. This method solves the problem of less training data and low precision by increasing the length information of the sample without reducing the number of samples. Experiments are conducted on AVEC2014 dataset, and the results shows that the proposed method is more effective and accurate than the existing speech depression recognition algorithms.**

***Index Terms*—automated depression diagnosis, speech processing, deep learning, feature extraction**

I. INTRODUCTION

In recent years, more and more attention has been paid to the study of mental health. Major depression disorder (MDD), usually simply named depression, is a serious threat to human mental health [1]. People with depression may feel sad, helpless, empty, anxious, anorexic, irritable or upset, and severe depression may even lead to suicide [2]. Depression can seriously affect people's normal life, but it can be cured with medication, psychotherapy and physical therapy [3]. At present, the diagnosis of MDD basically requires a comprehensive assessment by professionals with clinical experience. This is largely limited by the subjective observation of doctors and the lack of long-term follow-up diagnosis. As the number of MDD patients increases, it puts a burden on doctors to accurately diagnose the degree of depression. Therefore, providing an objective assessment and rapid diagnosis through machine learning methods can help with MDD therapy.

More and more attention has been paid to the study of automatic mental health assessment in recent years. Machine learning methods can be used to learn emotions and expressive behaviors directly related to depression. Research shows that the speech production of patients with depression is different from that of normal people [4] [5], so some methods of diagnosing depression using audio cues are proposed [6] [7] [8] [9] [10]. Speech has been one of the prime modalities explored for depression detection. Accordingly, we merely focus on the audio cues for depression detection in our work. Automatic Speech Depression Detection (ASDD) is using computer to analyze the speakers voice signal and its change process, to find their inner emotions and psychological activities. Currently, the methods of ASDD can be divided into two categories: traditional machine learning methods and deep learning methods.

Feature selection is the key to the traditional ASDD machine learning method, which is directly related to the accuracy of recognition. At present, the most commonly used feature include energy correlation feature, zero crossing rate,pitch frequency feature, resonance peak feature, spectral feature, etc. After feature extracted, the machine learning method is used to study the relationship between feature and depression. These machine learning methods include Gaussian Mixture Models (GMM) [11], Partial Least Square (PLS) and Support Vector Regression (SVR) [12], Extreme Learning Machine (ELM) [13]. The main advantage of this approach is that the model can be trained without requiring large amounts of data. The disadvantage is that it is difficult to judge the quality of features, and some key features may be lost, thus reducing the accuracy of recognition.

Compared with the traditional machine learning method, the deep learning has the following advantages: it can extract the high-level features [14], [15], and it has been shown

978-1-7281-1312-8/18/$31.00 ©2018 IEEE

to exceed human performance in visual tasks [16], [17]. At present, many researchers have applied deep learning methods to ASDD. Huang et al. [10] introduced a deep learning method for audio-based classification of depression at the 2016AVEC, in which two layers CNN, one LSTM layer and two fully connected layers are designed to predict whether audio subjects are depressive. In [18], authors designed the median robust extended local binary patterns (MRELBP) as hand-crafted features. They extracted hand-crafted features and used Deep Convolutional Neural Networks (DCNN) to predicted the depression score. In [19], Chao et al. extracted the features of audio and video, and fused them as a sign of abnormal behavior, and then described the dynamic time information by using the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). They adopted multi-tasking learning to improve the accuracy of the results and assessed the predictive power of the model on the AVEC2014 dataset.

Though previous studies have achieved some results, the accuracy of recognition still needs to be improved. Research shows that depression patients and normal people have significant differences in their emotions, such as depression, sadness, anxiety and worry [20]. In order to make full use of the emotion features and address the problems of less training data, this paper designs the Multiscale Audio Delta Normalization (MADN) Algorithm, using the contextual emotional information. More training data can be obtained by increasing the length of the sample without reducing the number of samples. Then, the MFCCs and the emotions features of two adjacent segments of local audio were fed into the Depression AudioNet to encode the depression related characteristics.

The outline of the paper is as follows: In Section 2, the method and network proposed in this paper are introduced in details. Section 3 illustrates and analyzes the experimental results. Section 4 concludes the paper.

II. Proposed Algorithms

Since human emotions and voices in depression are theoretically different from that of normal people, we attempt to address the problem of depression scale prediction by combining the emotional characteristics of the context (MADN) with the description of local voice expression. This paper proposes a depression audio network to learn depression expression and emotional characteristics in vocal clues and automatically predicts the Beck Depression Inventory scale of depression. Fig. 1. shows the framework for automatic depression recognition.

A. MFCCs Feature extraction

MFCCs is the most commonly used feature, which has the advantages of accord with human hearing and low dimensional. MFCCs use a nonlinear frequency scale based on auditory perception, namely mel scale. A mel is the unit of pitch frequency. Equation (1) can convert the frequency scale of audio to mel scale.

$$f_{mel} = 2595lg\left(1+\frac{f_{HZ}}{700}\right) \tag{1}$$

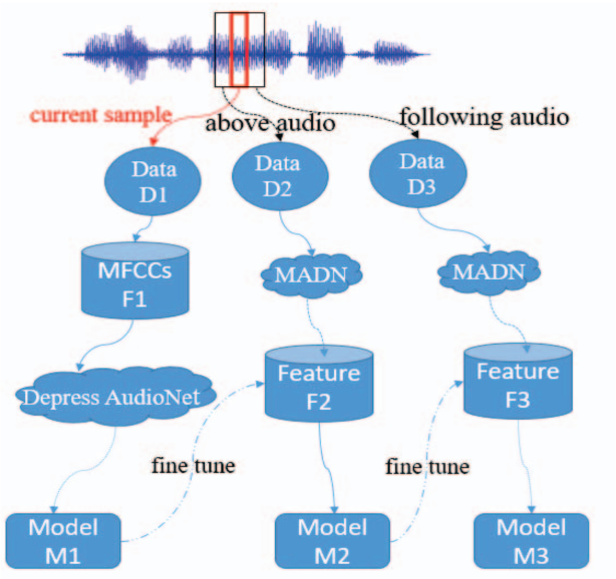

Fig. 1. The framework of automatic diagnosis of depression is presented in this paper.

Where f_{mel} means that the frequency of mels and f_{HZ} is the normal frequency. In general, MFCCs calculations use a set of L filters, each of which is triangular and evenly spaced on the mel scale. Let $l(l)$, $c(l)$ and $h(l)$ respectively be the lower limit, center and upper limit frequency of the l triangle filter, then the lower limit, center and upper limit frequency of the adjacent triangle filter have the following relationship:

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

The output of each triangular filter is then calculated as follows:

$$m(l) = \sum_{k=l(l)}^{h(l)} W_l(k)\left|X_n(k)\right| \quad l = 1, 2, ..., L \tag{3}$$

$$W_l\left(k\right) = \begin{cases} \frac{k-l(l)}{c(l)-l(l)} & l\left(l\right) \leq k \leq c\left(l\right) \\ \frac{h(l)-k}{h(l)-c(l)} & c\left(l\right) \leq k \leq h\left(l\right) \end{cases} \tag{4}$$

where $X_n(k)$ is the discrete Fourier transform (DFT) of a speech input $x[n]$.

Then the log-energy mel spectrum and discrete cosine transform (DCT) are calculated as follows:

$$v_{mfcc}\left(i\right) = \sqrt{\frac{2}{N}}\sum_{l=1}^{L} lgm\left(l\right) cos\left(\left(l-\frac{1}{2}\right)\frac{i\pi}{L}\right) \tag{5}$$

Typically, the number of filters L is between 20 and 40, and the number of kept coefficients is 13. The research shows that in the case of 32-35 filters, the accuracy rate of speech recognition is the highest [21].

B. Multiscale Audio Delta Normalization (MADN)

It is common knowledge that the combination of emotional information of two adjacent segments of local audio is very helpful for us to judge the mood and mental state of the speaker at that time. In other words, the more audio information we get from the speaker, the more accurate the results will be. However, for machine learning, the large dimension of a single sample results in much more computational complexity. Moreover, the data set of depression is not large enough, so we proposed Multiscale Audio Delta Normalization (MADN) algorithm to acquire the low-level emotional characteristics of the two adjacent segments of local audio. First order difference of MFCCs coefficient, short-term energy, zero-crossing rate and formant frequencies features, such derivative based features contain certain emotional information and are not easily affected by the speakers personalized speech features. In order to eliminate the differences in the numerical values of different features, we normalize the different differences features separately. For clarity, the MADN algorithm process is as follows:

- Input: Original audio data.
- Step 1 Read and pre-processing audio data from file.
- Step 2 The MFCCs coefficient, short-term energy, zerocrossing rate and formant frequencies are obtained. These features are represented by feature vector $V(n,f)$, component $n = 1,...,N$, frame $f = 1,...,F$.
- The Delta matrix $D(n,f)$ is obtained by differentiating the features of the front and following frames.

$$D\left(n,f\right) = V\left(n,f+1\right) - V\left(n,f\right) \quad f = 1,2,...,F-1 \tag{6}$$

- Step 4 Different features were normalized at different scales as shown in Equation(6). Where, the difference between F_n and f_n represents different scales.

$$\begin{gathered} F\left(n,f\right) = \frac{D(n,f)-D_{min}(n,f_n{:}F_n)}{D_{max}(n,f_n{:}F_n)-D_{min}(n,f_n{:}F_n)} \\ n = 1,2,...,N-1 \end{gathered} \tag{7}$$

- Output: The features $F(n,f)$ after normalized at different scales.

C. Depression AudioNet

The deep convolutional neural network has a good performance in the field of pattern recognition. It can generate advanced features and enrich the hand-crafted features. In this paper, a new deep model, namely Depression AudioNet, is proposed to encode the depression related temporal clues in speech mode and to predict the degree of depression in the audio sample. The flowchart of Depression AudioNet as shown in Fig. 2. In the conventional DCNN for pattern recognition, the input image, convolution kernel shape and pooling shape are often square. For the speech signal, the data dimensions is one-dimensional, so cannot be modeled directly using the image-processing application. To handle this issue, we extracted the MFCCs, zero crossing rate, energy and formant frequencies in frames for each audio segment.

In the representation of audio features, the horizontal and vertical axes have different meanings, the horizontal axis represents time, and the vertical axis represents frequency information. The simple convolution-pooling operation in the CNN, while introducing translation invariability either in frequency (or in time or in both), would cause confusion among speech classes and reduce the discrimination ability [22]. Therefore, we attempt to use the one-dimensional convolution layer on the whole frequency axis instead of using square-sized filter to solve this problem. The convolution neural networks is expected to capture rich high-level audio structures effectively. The purpose of the pooling layer is to reduce the resolution of the feature map, that is, the size of the feature, and introduce invariance for small changes in the location. The input features becomes a one dimensional data after operating of convolution. Next, these features are feed into the LSTM layers to extract long-term dependencies. Finally, at the end of the network architecture are two fully connected layers designed to encode long-term variability on the timeline and predict depression scores.

III. Experiments

A. AVEC2014 depression database

The proposed approach is evaluated on the AVEC2014 depression dataset [23]. The database consisted of audio data from 84 German subjects, who were recorded between one and four times. Two consecutive records spaced two weeks apart. Subjects were between 18 and 63 years old, with an average age of 31.5 years and a standard deviation of 12.3 years.

The recordings in the AVEC2014 subset have only two tasks : Northwind and Freeform, which are provided as separate recordings, with a total of 300 audio. The NorthWind refers to participants reading aloud an allegory from The North Wind and The Sun. The Freeform asked subjects to answer a series of questions, such as what was your favorite dish? What's your best gift? Why is that? And discuss a sad childhood memory. The challenge data was divided into three parts, the training, development and testing set, each with 50 pairs of Northwind/Freeform pairs, and a total of 300 task recordings.

B. Experimental Settings

1) Data pre-processing: In order to obtain the optimal feature, the audio sample is preprocessed to remove the silence. Firstly, in each audio file, long pauses are manually removed and the rest of the voice containing the individual is linked together to create a single file. The effectiveness of depression audio can be enhanced by eliminating long pauses in speech. Then, each valid audio signal file is split into segments of the same length, without overlapping. Totally, there are about 11000 audio segments extracted from the AVEC2014. Each segment consists of 60 frames, the size of the Hanning window $w_h$is 1024, the hop size is 512, half of the analysis window, and the audio sample rate is 44100 Hz, so the time that an audio segment covers is ((60+1)*1024/2)/44100=0.708s. Finally, after standard normal variate normalization, the MFCCs, zero crossing rate, energy and formant frequencies are calculated for each segment of speech.

2) Performance Measure: During the test, the predicted depression score of the audio input was calculated by averaging the predicted score of all sampled segments. Recognition performance can be measured by two commonly used evaluation

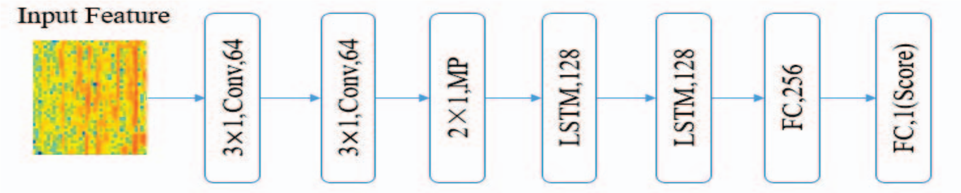

Fig. 2. Detailed network structure of the Depression AudioNet.

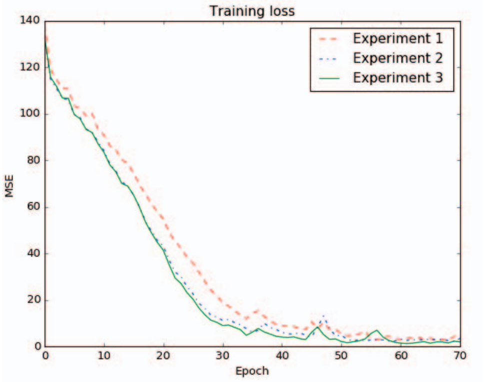

Fig. 3. The MSE loss versus number of iterations with different experiments on the AVEC2014 dataset.

indicators: mean absolute error (MAE) and root mean square error (RMSE).The MAE is computed by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \tilde{y_i}\right|. \tag{8}$$

The RMSE is computed by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y_i})^2}. \tag{9}$$

Where N denotes the total number of audio segments, y_i is the ground truth depression score of i_{th} audio segments, and $\tilde{y_i}$ is the predicted score of i_{th} audio segments.

C. First Experiment

In the experiment 1, we merge the training and development sets as training set. The overall performance is reported for audio samples from the test set. The result is shown in the first row in Table 1. It can be seen that, the MAE and RMSE obtained are 7.52 and 9.70, respectively.

TABLE I
RESULTS OF THE THREE EXPERIMENTS ON DEPRESSION SCALE. MEASURED BOTH IN RMSE AND MAE AVERAGE OVER ALL SEQUENCES IN THE TEST SET.

Methods	*RMSE*	*MAE*
Experiment 1	9.70	7.52
Experiment 2	9.46	7.30
Experiment 3	9.15	7.07

D. Second Experiment

In the experiment 2, model M1 is selected to fine-tune the data D2. Note that, the sample of the data D2 at this point is the previous audio segment of the sample of the data D1 in experiment 1. The data D2 is processed by the MADN algorithm, then get feature F2. F2 expresses the low-level emotional features of the previous audio segment. It can be seen from the second row of Table 1, the MAE and RMSE obtained are 7.30 and 9.46, respectively.

E. Third Experiment

In the experiment 3, model M2 is selected to fine-tune the data D3. Similarly, the sample of the data D3 at this point is the following audio segment of the sample of the data D1 in experiment 1. The data D3 is processed by the MADN algorithm, then get feature F3. F3 expresses the low-level emotional features of the latter audio segment. It can be seen from the third row of Table 1, the MAE and RMSE obtained are 7.07 and 9.15, respectively. From Fig. 3 we can see that the loss function converges faster when the experiment on the data D3. Moreover, we can also see from Table 1 that the recognition result of the third experiment is the best. The predicted values of the third experiment and actual depression scale values on the test set are shown in Fig. 4.

TABLE II
PERFORMANCE COMPARISON AGAINST OTHER APPROACHES ON THE TEST PARTITION,MEASURED IN RMSE AND MAE.

Methods	*RMSE*	*MAE*
Baseline [23]	12.57	10.04
Jan [24]	11.30	9.10
Mitra [12]	11.10	8.83
He [18]	9.99	8.19
Our method	9.15	7.07

Performance comparisons against other techniques including the baseline can be seen in Table 2. As you can see from the table, our proposed method performs better than other methods in the AVEC2014 dataset. This further proves the validity of our model recognition effect.

IV. CONCLUSION AND PROSPECT

The major clinical feature of depression is marked and persistent mood depression, which is the main type of mood disorder. Computer-assisted techniques have been used to help psychologists assess depression levels. In order to improve the accuracy of automatic depression recognition based on audio

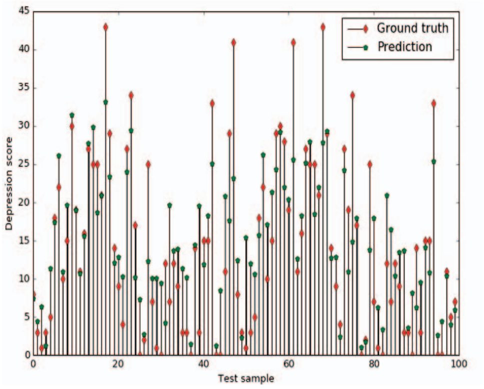

Fig. 4. Prediction bias of the proposed deep regression model on the test set.

signals, this paper proposes a method based on deep learning, which overcomes the problem that hand-crafted features will miss important information. In the proposed approach, MFCCs and emotion features (MADN) were used to solve the problem of less training data, and Depression AudioNet was used to extract the feature information of depression. we use the MFCCs and emotion features (MADN) to address the problem of less training data, and use Depression AudioNet to extract the characteristic information of depression. Moreover, we also proposed joint tuning layers using contextual emotional information, which can further improve the performance of the model in recognizing depression. Finally, we obtain about the results for MAE and RMSE achieved 7.07 and 9.15. In our future work, we will explore multimodal features, such as facial expressions and text messages, to more effectively recognition depression.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61876112,61601311, 61603022), Natural Science Foundation of Beijing (4162017)Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan (CIT&TCD20170322), Project of Beijing Excellent Talents (2016000020124G088), Beijing Municipal Education Research Plan Project (SQKM201810028018), Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds (025185305000/134/187/188/189) and Youth Innovative Research Team of Capital Normal University.

References

[1] R. H. Belmaker, G. Agam, “Major depressive disorder,” New England Journal of Medicine, vol. 358, no. 1, pp. 55–68, 2008.

[2] S. Salmans, “Depression: questions you have-answers you need,” PeoplesMedical Society, 1995.

[3] H. Alan Pincus and A. R. Pettit, “The societal costs of chronic major depression,” The Journal of clinical psychiatry, pp. 5-9, 2000.

[4] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, “Acoustical properties of speech as indicators of depression and suicidalrisk,” IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 829-837, 2000.

[5] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, “Investigating voice quality as a speaker-independent indicator of depression and ptsd,” In Interspeech, pp. 847-851, 2013.

[6] Y. Yang, C. Fairbairn, J. F Cohn, “Detecting depression severity from vocal prosody” IEEE Transactions on Affective Computing, vol. 4, no.2, pp. 142-150, 2013.

[7] M. JH Balsters, E. J. Krahmer, M. GJ Swerts, and A. JJM Vingerhoets, “Verbal and nonverbal correlates for depression: A review,” Current Psychiatry Reviews, vol. 8, no. 3, pp. 227-234, 2012.

[8] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, “Multimodal assistive technologies for depression diagnosis and monitoring,” Journal on Multimodal User Interfaces, vol. 7, no. 3. pp. 217-228, 2013.

[9] Y. Niu, D.Zou, Y. Niu, “Improvement on Speech Emotion Recognition Based on Deep Convolutional Neural Networks,” International Conference, pp. 13-18, 2018

[10] M. Xingchen, Y. Hongyu, C. Qiang, H. Di, W. Yunhong, “DepAudioNet:An Efficient Deep Model for Audio based Depression Classification” The International Workshop, pp. 35-42, 2016.

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, “Vocal biomarkers of depression based on motor incoordination,” In International Workshop on Audio/Visual Emotion Challenge, pp. 41-48, 2013.

[12] V. Mitra, R. Ave, M. Park, E. Shriberg, R. Ave, M. Park, M. Mclaren, A. Kathol, R. Ave, M. Park, C. Richey, and M. Graciarena, “The SRI AVEC- 2014 Evaluation System,” International Workshop on Audio/visual Emotion Challenge, pp. 93-101, 2014.

[13] J. R. Williamson, T. F. Quatieri, B. S. Helfer, “Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing,” The International Workshop, pp. 65-72, 2014.

[14] Bengio, Yoshua, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, 2013.

[15] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,”Nature, vol. 521, no. 7553, pp. 436-444, 2015

[16] V. Mnih, K. Kavukcuoglu, D. Silver, “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[17] D. Silver, A. Huang, C. Maddison, “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484-489, 2016.

[18] L. He, C. Cao, “Automated depression analysis using convolutional neural networks from speech,” Journal of Biomedical Informatics, pp.103-111, 2018.

[19] L. Chao, J. Tao, M. Yang, “Multi task sequence learning for depression scale prediction from video,” International Conference on Affective Computing and Intelligent Interaction, pp. 526-531, 2015.

[20] World Health Organization, “Depression and other common mentaldisorders: global health estimates,” Tech. Rep, 2017.

[21] V. Tiwari, “MFCC and its applications in speaker recognition, International Journal on Emerging Technologies,” vol. 1, pp. 19-22, 2010.

[22] L. Deng, O. Abdel-Hamid, D. Yu, “A deep convolutional neural networkusing heterogeneous pooling for trading acoustic invariance with phonetic confusion,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6669-6673, 2013.

[23] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R.Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” International Workshop on Audio/Visual Emotion Challenge, pp. 3-10, 2014.

[24] A. Jan, H. Meng, Y. Gaus, “Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression,” International Workshop on Audio/visual Emotion Challenge, pp. 73-80, 2014.