EasyChair Preprint

№ 2371

Mental Health Detection from Speech Signal: A Convolution Neural Networks Approach

Haizhen An, Xiaoyong Lu, Renjun Li, Daimin Shi, Jingyi Yuan and Tao Pan

January 12, 2020

Mental Health Detection from Speech Signal: A Convolution Neural Networks Approach

1st Haizhen An
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, China
an1832616110@163.com

2nd Xiaoyong Lu
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, China
luxy@nwnu.edu.cn

3rd Renjun Li
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, China
603897587@qq.com

4th Daimin Shi
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, China
xst18419140068@163.com

5thJingyi Yuan
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, China
m765415210@163.com

6th Tao Pan
Lanzhou Resources and Environment
Voc-Tech College
Lanzhou, China
pant_revt@126.com

***Abstract*—Mental health disorder is a global topic, the current situation is particularly serious in China. The objective and automated detecting of mental health using speech signal has become popular. In the absence of depressed speech corpus, the authors regard depression as a negative emotion, and build the model by Convolution Neural Networks (CNNs), a machine learning method for detecting mental health disorder interchanging with emotional speech. In this experiment, the segmented speech was represented as a spectrogram in the frequency-time domain via a Short-Time Fourier Transform (STFT), and these images were as input of the CNNs model. It highlights some advantages that CNNs can offer mental health detection. Results indicate that it is a good attempt and this method can be directly utilized by interchanging with emotional speech.**

Keywords—mental health, negative emotion, detection, Convolutional Neural Networks

I. INTRODUCTION

Without national health, there will be no moderate prosperity for the whole country. Frankly speaking, mental health is the most import composing parts of national health [1]. Currently, with the process of industrialization and urbanization, social competition is intensifying in the process of economic construction and industrialization, resulting in a sharp increase in various psychological stress factors, mental disorders and mental health problems are becoming increasingly prominent. The negative consequences of mental health disorder tend to be explosive and lead to extreme behavior [2]. Mental health disorders in one's life typically included autism, depression, and Alzheimer's disease [3][4][5]. Among them, depression severely impact quality of life, is costly both for affected individuals and society, and increases risk for suicide [6]. According to data from the World Health Organization (WHO) in 2017 show that, more than 300 million people worldwide suffer from depression, and that number is increasing every year [7]. Among which, depression affects human health in the 21st century as the main factors, the impact of physical and mental health for young people can not be ignored.

Current researches show that potential advantages for detecting mental health related disorders by machine learning methods to capture and model key behavioral signals [8][9]. Behavioral signal processing research indicates the benefits to using signals – such as speech, facial and eye tracking to automatically and objectively detect mental health disorder, such as depression. From the machine learning perspective, mental health disorder detection can be considered as a regression or classification problem [10][11]. Depression is a classical mental health disorder, and is a strong and persistent negative emotion in psychology. So in this paper, the focus is primarily on depression, and we regard depression as a negative emotion. Our goal is to detect depression in negative emotion perspective, i.e. depression classification between non-depressed and depressed categories.

Deep learning has undoubtedly improved the state-of-the-art performance of machine learning models across a variety of machine learning applications. In this regard, Convolutional Neural Networks (CNNs) [12] have become increasing popular in deep learning research. Moreover, speech has long been recognized as a key component for any behavioral based mental health detection system [13]. Clinically, speech's abnormalities associated with an individual's mental health state are well documented [14]. For example, speech in patients with depression is often described diminished prosody, monotonous and "lifeless" [15].

The remainder of this paper is laid out as follows. Section II briefly discusses background and related work. Section III outlines more implementation details about the proposed framework of Convolutional Neural Networks. Section IV introduces our machine learning method for classification experiment. The obtained results are given in Section V, before presenting conclusions and future works are discussed in Section VI.

II. BACKGROUND AND RELATED WORK

Despite tremendous efforts by domestic and overseas scholars in the past decade, depression detection related issues have not been resolved. Different scholars hold different viewpoints upon depression detection. In psychology, depression is a negative emotion, which mainly contains pain with anger, sadness, and guilty and so on. It is a big challenge how to fast and effective and objective detect the presence of depression, even for clinicians. Current the state-of-the-art depression diagnostic methods such as the Hamilton Rating Scale for Depression (HAMD), the Beck Depression Index (BDI), Patient Health Questionnaire (PHQ) and the Quick Inventory of Depressive Symptomatology (QIDS) [16][17][18][19], which include on the basis of an interview style assessment between a clinician and a patient, patient self-reporting, and typical

rating scales . However, these tests are subjective and single in nature, low diagnostic accuracy and lack an objective predictor of depression. Therefore, to enhance current detection methods based on behavioral signals, is needed.

From the machine learning standpoint, depression detection can be considered as a regression or classification problem. In the AVEC depression sub-challenge, this paper [20] explored the suitability of CNNs for depression detection. In this experiment, the CNNs set up was used. And the authors utilized the down-sampling methodology during training for the 'not-depressed' class. In [21], the same authors trained 4systems, a combination of gender specific models for 'depressed' and 'not-depressed' classes. Their approach also reached good results. This paper [22] proposed a method that the authors use the raw and spectrogram CNNs to model the characteristic information of depression. The experimental results show the proposed method can improve the performance of depression recognition.

III. CONVOLUTION NEURAL NETWORKS

Convolutional Neural Networks is one of the representative algorithms of deep learning. Its structure is based on biological visual perception mechanism [23]. CNNs is mainly used in image recognition, video analysis and natural language processing. CNNs is mostly used for image processing. At present, it has been successfully applied in speech analysis and research.

Convolutional Neural Networks is generally divided into three layers: convolution layer, pooling layer and full connection layer. Its general structure framework is shown in Figure 1. Convolution layer is composed of convolution core and offset in each two-dimensional plane. Weight sharing is used to reduce the scale of network parameters [24]. Convolution layer convolutes different convolution kernels with each channel of input image by sliding window to extract different features of input image. The pooling layer is the feature mapping layer. The feature obtained by the convolution layer is sampled down, which reduces the size of the input data and obtains the local optimal value. Pooling operation can reduce the computational cost of feature mapping dimension and training process, so the pooling layer is mainly used for feature selection and selection of more effective features. Full connection layer mainly carries out model training, learning high-order features, and ultimately realizes effective classification, which can be said to be the output stage of CNNs.

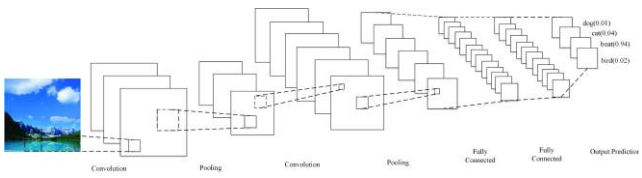

Fig. 1. A General overview of CNNs architecture.

IV. METHODOLOGY

A. Mental health detection using CNNs

In the experiment, CNNs was used to classify whether negative emotions were binary or not, that is to identify negative emotions and neutral emotions. Because CNNs is mostly used for classification and recognition of pictures, it is necessary to convert speech into pictures in order to recognize speech through CNNs. So in the experiment, the speech is transformed into a spectrogram by using short-time Fourier transform, and then the spectrogram is used as the input of CNNs. In the spectrogram, the audio power level of a given frequency and time is expressed mainly by the gray level of the picture. At this time, the prosodic features of speech are reflected in the spectrogram of each speech. Then we use the convolution layer of CNNs to learn the prosodic characteristics of negative emotion and neutral emotion to extract features.

Negative emotion and neutral emotion were trained respectively. Sliding filter is used to learn the characteristics of negative emotion and neutral emotion individuals. These acquired features can provide different prosodic features. These prosodic features represent potential differences between negative and neutral emotions, and ultimately result in classification. Figure 2 shows the structure of negative emotion and neutral emotion recognition using CNNs.

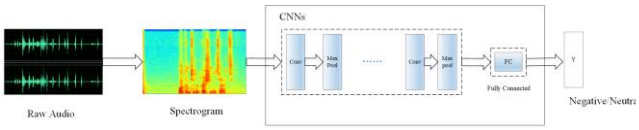

Fig. 2. Illustration of the proposed method using CNNs.

The specific process of recognizing negative and neutral emotions through CNNs is as follows: After transforming speech into spectrogram by short-time Fourier transform, it is divided into training set and testing set. Firstly, the training set is used for model training, and then the trained model is tested with the test set. Then dictionaries for negative and neutral speech are created by ID of each speech. We set the label of negative emotions to 1 and the label of neutral emotions to 0. Secondly, the spectrum is divided into four-second-wide speech by Hamming window, and then the segmented speech is sampled randomly. Then the spectrum of the sampled speech is used as the input of CNNs to train and test the model. Finally, the ROC curve is drawn to evaluate the model.

In this experiment, a six-layer CNNs model is used, which consists of two convolutional layers with two max-pooling layers and two fully connected layers as shown in Figure 3. Max-pooling is an operation of window sliding to obtain maximum eigenvalues on each sub-window of feature mapping. The size of feature mapping can be reduced by max-pool operation, which is controlled by pool size and spanning over-parameter [25]. Each spectral input is an image with a size of 513 x 125, representing 4 seconds of audio and 0 to 8 kHz frequency.

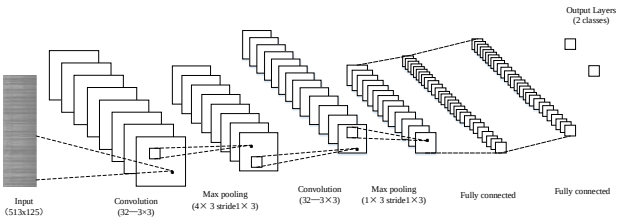

Fig. 3. The structure for experiments applying CNNs

B. Experimental Settings

- the segmented speech will be represented as a spectrogram in the frequency-time domain via a short-time Fourier transform.
- Build subjects' dictionaries data for the each class and values with the matrix representation of the entire segmented wav file's spectrogram.
- Perform normalization on the spectrogram and preps the images for Keras. Then trains and evaluates the network.

V. RESULTS

The model of this study was obtained by training 9 students with negative and neutral emotions, with average 33 sentences per student (resulting in 594 sentences in total). The training set consisted of 6 students with negative emotions and 6 students with neutral emotions, totaling 396 sentences. The test set consisted of 3 students with negative emotions and 3 students with neutral emotions, totaling 198 sentences as shown in Table 1. The results showed that the accuracy of the training model was 78.4%, the accuracy of the test set was 70.5%, and the experimental results of the final model were over 70.5%.

TABLE I. EXPERIMENTAL DATA FOR TRAINING AND TESTING

	Train	Test	Total
Negative	198	99	297
Neutral	198	99	297
Total	396	198	594

At the end of the experiment, we evaluated the model by F1 score as in (1) and accuracy. Table 2 shows the evaluation score. Figure 4 shows the AUC score of the CNNs model used in the study.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (1)$$

TABLE II. MODEL EVALUATION RESULTS

F1 score	precision	recall	accuracy
0.606	0.909	0.455	0.705

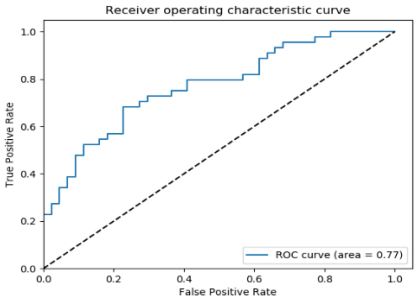

Fig. 4. ROC curve of CNNs model

In this study, the CNNs method was used to collect the speech of negative and neutral emotions of many students, and to establish a binary prediction model for whether the high-dimensional speech data are negative emotions or not. Through the research results, we can see the significance and value of speech as a fast emotional recognition. In view of the achievements and shortcomings of this study, future research can, on the one hand, increase the research data set of negative emotions, and conduct more accurate model training, so as to improve the accuracy rate. On the other hand, more feature extraction methods and classification methods can be used for repeated verification to improve the accuracy of the experiment.

VI. CONCLUSIONS AND FUTURE WORKS

This work explored the effects of mental health detection based on CNNs using speech signal. Our analysis indicates that this method can be directly utilized by interchanging with emotional speech in the absence of depressed speech. But we can also observed the method were not always stable and could produce results of widely varying quality. So to continue testing the performance of the CNNs method in practical application, there is still a need for large amounts of real data. However, the work of collecting such data is costly, time consuming and not quite easy.

Future work will continue to seek some appropriate methods to improve the accuracy of the model. In order to solve the problem about the lack of depressed speech data, we plan to utilize Transfer Learning for depression detection. Finally, we also plan to collect further multimodal dataset available on depression.

ACKNOWLEDGMENT

This research has received funding from the academic requirements for the National Science Foundation of China (NSFC) under grant No. 31860285 and No. 31660281. Additionally, part of this work is performed in the Scientific Research Project in Higher Education Institutions of Gansu Province (Grant No. 2017A-165). We also want to thank the reviewers for their thoughtful comments and efforts towards improving our paper.

REFERENCES

[1] World Health Organisation, Preventing Suicide: A Global Imperative . Geneva, 2014.

[2] K. Hawton, C. I. C. Carolina, C Haw, et al. Risk factors for suicide in individuals with depression: A systematic review. Journal of Affective Disorders, 2013, 147(1-3):17-28.

[3] M. Valstar, J. Gratch, B. Schuller, F. Ringeval,et al. Depression, mood, and emotion recognition workshop and challenge. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. AVEC'16, ACM, Amsterdam, Netherlands, 2016, pp. 3–10.

[4] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, et al. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. Proceedings INTERSPEECH 2016. 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, CA, 2016, pp. 1210–1214.

[5] O. B. Tysnes, A. Storstein. Epidemiology of parkinson's disease. Neural Transm. 124 (2017) 901–905.

[6] C.M. Bell, J.A. Ridley, J.C. Overholser, et al. The Role of Perceived Burden and Social Support in Suicide and Depression. Suicide and Life-Threatening Behavior, 2017.

[7] World Health Organization, Global Tuberculosis Report 2017. Geneva, 2017.

[8] G. Gosztolya, T. Grósz, R. Busa-Fekete, L. Tóth. Detecting the intensity of cognitive and physical load using adaboost and deep rectifier neural networks. Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, Singapore, 2014, pp. 452–456.

[9] H. Jing, T.-Y. Hu, H.-S. Lee, W.-C. Chen, et al. Ensemble of machine learning algorithms for cognitive and physical speaker load detection. Proceedings INTERSPEECH 2014, 15th Annual Conference of the

International Speech Communication Association, ISCA, Singapore, 2014, pp. 447–451.

[10] M. Valstar, B. Schuller, K. Smith, F. Eyben, et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 3–10.

[11] M. Valstar, B. Schuller, K. Smith, T. Almaev,et al. Avec 2014: 3d dimensional affect and depression recognition challenge. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 3–10.

[12] A. Krizhevsky , I. Sutskever, G. Hinton . ImageNet Classification with Deep Convolutional Neural Networks. NIPS. Curran Associates Inc. 2012.

[13] N. Cummins, J. Joshi, A. Dhall , et al. Diagnosis of depression by behavioural signals: a multimodal approach. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, 2013.

[14] N. Cummins, A review of depression and suicide risk assessment using speech analysis. Speech Communication 71(2015):10-49.

[15] C. Sobin, H A. Sackeim. Psychomotor symptoms of depression. American Journal of Psychiatry, 1997, 154(1):4-17.

[16] Takahashi. Rating scale for depression. Journal of Neurology Neurosurgery & Psychiatry, 1998, 23(1):: 56–62.

[17] A T. Beck, R A. Steer, R. Ball, et al. Comparison of the Beck Depression inventories IA and II in psychiatric outpatients. Journal of Personality Assessment, 1996, 132(3):381-385.

[18] K. Kroenke, R L. Spitzer, J B W. Williams . The PHQ-9 : Validity of a Brief Depression Severity Measure. Journal of General Internal Medicine, 2001, 16(9):606-613.

[19] J.A. Rush, M.H. Trivedi, H.M. Ibrahim, T.J. Carmody, et al. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biological Psychiatry, 2003, 54(5):573-583.

[20] L. Yang, D. Jiang, X. Xia, E. Pei, et al. Multimodal measurement of depression using deep learning models. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC'17, ACM, Mountain View, CA, 2017, pp. 53–59.

[21] L. Yang, H. Sahli, X. Xia, E. Pei, et al. Hybrid depression classification and estimation from audio video and text information. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, ACM, Mountain View, CA, 2017, pp. 45–51.

[22] N. Cummins, B. Alice , B W. Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. Methods, 2018:S1046202317303717-.

[23] H. Xing, G. Zhang, M. Shang. Deep Learning. International Journal of Semantic Computing, 2016, 10(03):417-439.

[24] A. Krizhevsky, I. Sutskever, G. Hinton . ImageNet Classification with Deep Convolutional Neural Networks. NIPS. Curran Associates Inc. 2012.

[25] W. Feng, W. Yu, T. Yu, Z. Ping, et al. Pattern recognition and prognostic analysis of longitudinal blood pressure records in hemodialysis treatment based on a convolutional neural network. Journal of Biomedical Informatics,2019,98.