INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic Depression Level Detection via ℓ_p-norm Pooling

Mingyue Niu[1,2], *Jianhua Tao*[1,2,3], *Bin Liu*[1], *Cunhang Fan*[1,2]

[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

niumingyue2017@ia.ac.cn, {jhtao, liubin, cunhang.fan}@nlpr.ia.ac.cn

Abstract

Related physiological studies have shown that Mel-frequency cepstral coefficient (MFCC) is a discriminative acoustic feature for depression detection. This fact has led to some works using MFCCs to identify individual depression degree. However, they rarely adopt neural network to capture high-level feature associated with depression detection. And the suitable feature pooling parameter for depression detection has not been optimized. For these reasons, we propose a hybrid network and ℓ_p-norm pooling combined with least absolute shrinkage and selection operator (LASSO) to improve the accuracy of depression detection. Firstly, the MFCCs of the original speech are divided into many segments. Then, we extract the segment-level feature using the proposed hybrid network, which investigates the depression-related information in the spatial structure, temporal changes and discriminative representation of short-term MFCC segments. Thirdly, ℓ_p-norm pooling combined with LASSO is adopted to find the optimal pooling parameter for depression detection to generate the utterance-level feature. Finally, depression level prediction is accomplished using support vector regression (SVR). Experiments are conducted on AVEC2013 and AVEC2014. The results demonstrate that our proposed method achieves better performance than the previous algorithms.

Index Terms: depression detection, MFCC, a hybrid network, segment-level feature, ℓ_p-norm pooling, utterance-level feature

1. Introduction

Depression is a psychiatric disorder and deprives people of confidence and pleasure in life. More seriously, it even lead to suicidal behavior [1]. According to the World Health Organization [2], the number of people with depression in the world is about 350 million. In order to prevent the occurrence of misfortune, early diagnosis and intervention are particularly important for depression patients. However, the diagnosis process is rather subjective and mainly relies on patients' self-report and doctors' clinical experience [3]. In addition, doctors need to spend a lot of energy in the process, which will increase the probability of misdiagnosis in subsequent works. Thus, it is necessary to develop an automatic method to assist doctors.

Physiological studies [4, 5, 6] have shown that there are some differences in speech between depressed and normal individuals. Based on these facts, many researchers [7, 8, 9, 10] apply machine learning methods to explore the relationship between speech and Beck Depression Inventory-II (BDI-II) scores [11], which is a scale to measure the severity of depression and involves depression score ranging from 0 to 63 (0-13 no depression, 14-19 mild depression, 20-28 moderate depression and 29-63 severe depression). However, there are some limitations in previous methods. Firstly, most of them [8, 10] predict individual depression level through hand-crafted features, which lose some useful patterns related to depression [12]. Secondly, some methods [13, 14] only use spatial structure or temporal changes to describe the spectrograms and MFCC. In this way, the corresponding temporal or spatial information that is helpful to detect depression may be missed. Thirdly, the feature pooling or aggregation methods [7, 15, 16] used to characterize long-term speech are not necessarily optimal for depression detection task.

In order to alleviate the above issues, this paper proposes a novel approach to detect depression level using MFCC, which is considered to be a discriminative biomarker between depressed patients and normal individuals [6]. In particularly, we firstly obtain the MFCCs of the original speech and divide it into many segments. Then, the segment-level feature is extracted by our proposed hybrid network, which effectively integrates convolution neural network (CNN), long short term memory (LSTM) and deep neural network (DNN) to explore the information about depression in spatial structure, temporal changes and discriminative representation of short-term MFCC segments. Thirdly, this paper combines ℓ_p-norm pooling with LASSO (The least squares problem with ℓ_1 regularization term is called LASSO [17]) to find the optimal pooling parameter for depression detection and generates the utterance-level feature through pooling those segment-level features. Finally, SVR is employed to predict the depression level. We conduct experiments on Audio/Visual Emotion Challenge (AVEC) 2013 [18] and AVEC2014 [15]. The results indicate the superiority of our method.

The rest of this paper is organized as follows. In section 2, we review the works related to automatic depression detection. In section 3, we provide a detailed description of the method in this paper. Our experimental results and discussion are presented in section 4, and section 5 concludes the paper.

2. Related Work

Recently, depression detection using speech has attracted the attention of many researchers. Therefore, we briefly review some existing works in this section.

Valstar et al. [18] released a database for depression analysis in 2013 and supplied a baseline. They divided the speech into many fixed length segments, then 2268 features were extracted from these segments by open-source Emotion and Affect Recognition (openEAR) toolkit [19]. However, only using these hand-crafted low-level features might lost other information associated with depression [12]. In addition, they aggregated the features extracted from segments to generate the representation of the speech through average-pooling, which is a special case of ℓ_p-norm pooling [20] and not necessarily optimal for the depression detection. In [21], the authors used motion history histogram (MHH) to capture the dynamic changes in the speech and treated it as the corresponding feature. While speech sig-

http://dx.doi.org/10.21437/Interspeech.2019-1617

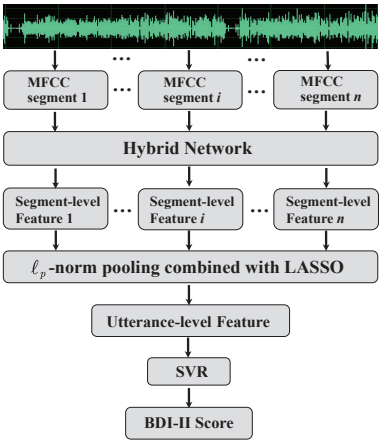

Figure 1: *The flow of our proposed method for automatic depression level detection.*

nals don't have as rich texture information as images, thus the extracted feature through MHH is limited in predicting depression levels. Jain et al. [16] adopted Fisher Vector to encode the original waveform to detect the depression level of individuals. However, the number of Gaussian components was not adapted to the depression detection task, which affected the accuracy of prediction [22]. He et al. [12] proposed a four-stream CNN to detect an individual depression level. Although, CNN is good at capturing spatial structure [23], it can not well explore the impact of temporal changes on depression detection. Moreover, the method augmented data to train the network through rotating and scaling transforms, which corrupted the meaning of data itself and degraded the performance of detection.

Different from the above works, this paper takes the advantage of representation ability of neural network to extract the high-level feature related to depression. Furthermore, we can find the suitable pooling parameter for depression detection task through optimizing the ℓ_p-norm pooling. The experimental results on AVEC2013 and AVEC2014 show the effectiveness of our method.

3. Proposed Method

Based on the fact that MFCC is a discriminative biomarker to detect depression disorder [6], this paper firstly divides the MFCCs of the speech into many segments. Then, the proposed hybrid network is used to extract the segment-level feature for each segment. Thirdly, the objective function of ℓ_p-norm pooling combined with LASSO is optimized to find the appropriate parameter for depression detection to generate the utterance-level feature. Finally, SVR is employed to predict individual depression level. The flow of our proposed method is shown in Fig. 1.

3.1. MFCC Segments

As mentioned above, the MFCCs extracted from the original speech are divided into a lot of segments through a fixed length window. There are two reasons for this processing: firstly, we can explore the detailed differences among individuals with different depression levels from the short-term MFCC segments. Secondly, more samples can be obtained to train the hybrid network. In addition, we regard the label of each segment as the BDI-II score of the corresponding speech.

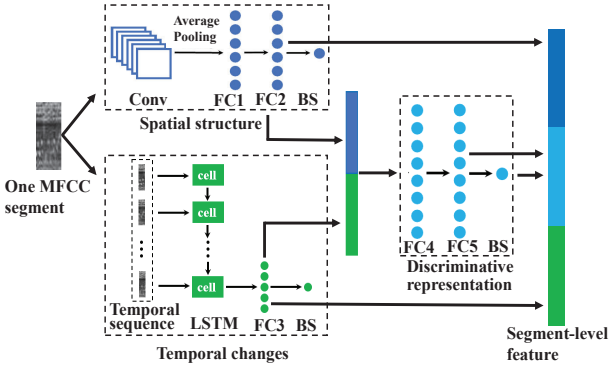

Figure 2: *The proposed hybrid network for extracting segment-level feature. Conv, FC and BS represent the convolutional layer, the fully connected layer and BDI-II score, respectively.*

3.2. Segment-Level Feature Extraction via the Hybrid Network

In order to fully explore the information related to depression in MFCC, this paper proposes the hybrid network as shown in Fig. 2 to extract the feature of each segment. On the one hand, considering that MFCCs have both spatial and temporal attributes [24], CNN and LSTM are adopted to explore the differences among different depression levels in the spatial structure and temporal changes of these short-term MFCC segments. On the other hand, for the sake of further enhance differentiation, we concatenate the two features extracted by CNN and LSTM as a new sample and input it into the DNN to obtain the discriminative representation. And there is a batch normalization layer before FC1. It is necessary to point out that these three subnetworks are trained separately.

In this paper, the output of each MFCC segment through the hybrid network is denoted as segment-level feature. In other words, this feature is composed of the output of FC2, FC3 and FC5, which reflect the information about depression in spatial structure, temporal changes and discriminative representation of each MFCC segment, respectively.

3.3. Utterance-Level Feature Generation by combing ℓ_p-norm Pooling with LASSO

Obtaining the speech representation is a key stage for depression detection. To this end, we combine ℓ_p-norm pooling with LASSO to find the suitable parameter for depression detection to aggregate the above segment-level features into an utterance-level feature.

For clarity, we present the process of ℓ_p-norm pooling for a speech sample in Fig. 3. In this figure, it is assumed that the MFCCs of the speech are divided into n segments. And $\mathbf{s}_i (i = 1, 2, \cdots, n)$ denotes the segment-level feature extracted from each segment using the hybrid network. Then, these features are arranged into a matrix as shown in Fig. 3. Finally, the ℓ_p-norm pooling result is obtained by calculating the ℓ_p-norm of each column of the matrix. The ℓ_p-norm definition of a vector is given by Eq. (1).

$$\|\mathbf{x}\|_p \triangleq \left(\frac{1}{n}\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}, \quad \forall \mathbf{x} = [x_1, \cdots, x_n]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}. \tag{1}$$

Following the above process, the ℓ_p-norm pooling result ($\mathbf{u}^{(k)}$) of the k-th speech sample can be written as the form of Eq. (2). It should be noted that the absolute value sign can

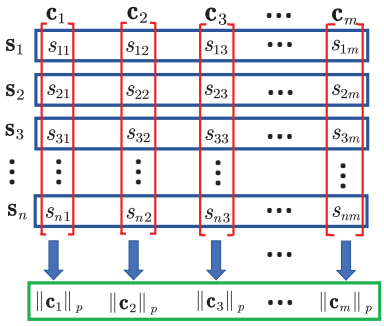

Figure 3: *The process of ℓ_p-norm pooling. $\mathbf{s}_i(i = 1, 2, \cdots, n)$ is the i-th segment-level feature. The column vectors surrounded by red boxes are denoted as $\mathbf{c}_j(j = 1, 2, \cdots, m)$. And the row vector surrounded by the green box is the result using ℓ_p-norm pooling.*

be removed because the segment-level features is min-max normalized to $[0, 1]$ in this paper. Moreover, it is easy to verify that the widely used average-pooling and max-pooling are special cases of ℓ_p-norm pooling when $p = 1$ and $p = \infty$.

$$\mathbf{u}^{(k)} = [(\frac{1}{n}\sum_{i=1}^{n}(s_{i1}^{(k)})^p)^{\frac{1}{p}}, \cdots, (\frac{1}{n}\sum_{i=1}^{n}(s_{im}^{(k)})^p)^{\frac{1}{p}}]. \quad (2)$$

As described, if the number of speech samples is K, then the sample matrix is expressed as $\mathbf{U} = [\mathbf{u}^{(1)^{\mathrm{T}}}, \cdots, \mathbf{u}^{(K)^{\mathrm{T}}}]^{\mathrm{T}} \in \mathbb{R}^{K\times m}$, where, $\mathbf{u}^{(k)} \in \mathbb{R}^{1\times m}$ $(k = 1, 2, \cdots, K)$ is calculated by Eq. (2). Let $\mathbf{b} = [b^{(1)}, \cdots, b^{(K)}]^{\mathrm{T}}$ be the BDI-II score vector, which measures the depression level of each sample. In this way, we combine ℓ_p-norm pooling with LASSO and propose the objective function as illustrated in Eq. (3), where $\mathbf{w} \in \mathbb{R}^{m\times 1}$ is a weight vector, p is the pooling parameter to be optimized and λ is the trade-off. Noted that Eq. (3) does not explicitly contain p, but p is involved in the process of calculating $\mathbf{U}$. It is obvious that the appropriate pooling parameter (p) for depression detection is found in the process of learning the mapping relationship between utterance-level features and BDI-II scores. Hence, the optimized pooling parameter (p) is more suitable for depression detection task than average-pooling ($p = 1$) and max-pooling ($p = \infty$).

$$\min_{\mathbf{w},p} f(\mathbf{w}, p) = \frac{1}{2K}\|\mathbf{U}\mathbf{w} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{w}\|_1. \quad (3)$$

In order to solve the optimization problem in Eq. (3), we adopt the alternation gradient descent method [25]. For a given $\mathbf{w}$, the gradient of f with respect to p is presented by Eq. (4), where $d\mathbf{u}^{(k)}/dp$ is a row vector of the same size as $\mathbf{u}^{(k)}$ and each element of it is calculated using Eq. (5). Similarly, Eq. (6) provides the gradient of f with respect to $\mathbf{w}$ when p is fixed.

$$\frac{\partial f}{\partial p} = \frac{p}{K}\sum_{k=1}^{K}(\mathbf{w}^{\mathrm{T}}\cdot\mathbf{u}^{(k)} - b^{(k)})\cdot(\mathbf{w}^{\mathrm{T}}\cdot\frac{d\mathbf{u}^{(k)}}{dp}), \quad (4)$$

$$\frac{du_j^{(k)}}{dp} = \frac{p\sum_{i=1}^{n}[t^p\cdot\ln(t)]}{\sum_{i=1}^{n}t^p\cdot\ln(\frac{1}{m}\sum_{i=1}^{n}t^p)}, j = 1, 2, \cdots, m. \quad (5)$$

where $t = s_{ij}^{(k)}$ is the j-th component of the i-th segment-level feature of the k-th speech sample.

$$\frac{\partial f}{\partial \mathbf{w}} = \frac{1}{K}\cdot\mathbf{U}^{\mathrm{T}}(\mathbf{U}\mathbf{w} - \mathbf{s}) + \frac{\lambda}{m}\cdot\Theta, \quad (6)$$

where, Θ is a vector of the same size as $\mathbf{w}$ and each elements of it is calculated by Eq. (7).

$$\Theta_j = \begin{cases} 1, w_j > 0 \\ 0, w_j = 0 \\ -1, w_j < 0 \end{cases}, \quad j = 1, 2, \cdots, m. \quad (7)$$

For convenience, we assume that $\mathbf{w}^*$ and p^* are the solutions of Eq. (3), then the utterance-level feature is generated by two steps. Firstly, the result (denoted as $\mathbf{u}^{(k)^*}$) of ℓ_p-norm pooling is obtained by substituting $p = p^*$ into Eq. (2). Secondly, we sort the elements of $\mathbf{w}^*$ i.e., $|w_{i_1}^*| \geq \cdots \geq |w_{i_m}^*|$. If the number of selected feature is M, then the (i_1-th, ..., i_M-th) columns of $\mathbf{U}$ are the utterance-level features of all samples.

4. Experiments

In this section, we firstly introduce the databases used in experiments briefly, and then the experimental setups are given. Finally, the results and discussion are presented.

4.1. Databases and Evaluation Measures

Experiments are conducted on two public datasets i.e., AVEC2013 and AVEC2014. In AVEC2013 corpus, there are 150 videos from 82 subjects and these recordings are divided into three parts: training, development and testing, each has 50 samples. In AVEC2014 corpus, only two different tasks are involved i.e., "Northwind" and "FreeForm" [15]. In each task, there are 150 videos, which are divided equally into training, development and test sets. For the fairness of comparison, we also combine the training, development and test sets of these two tasks as the new data in the experiments. Namely, there are 100 samples in the training, development and test sets, respectively.

Currently, root mean square error (RMSE) and mean absolute error (MAE) are widely used indicators for evaluating depression detection algorithms. The calculation of RMSE and MAE is shown in Eq. (8) and Eq. (9), where N denotes the number of samples. y_i and $\hat{y}_i$ are true and predicted BDI-II score of the i-th sample, respectively.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \quad (8)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \quad (9)$$

4.2. Experimental Setups

As mentioned in Fig. 2, the proposed hybrid network contains CNN, LSTM and DNN. In the CNN, there are 32 convolution kernels with a size of 3×3 and the stride size of 1. Then, we adopt average-pooling with a size of 2×2 and a stride of 1 for the pooling layer. Finally, two fully connected layers with 1024 neurons are followed. The same number of neurons is also set in the FC3 and FC5 layer. Note that the three subnetworks are trained separately and the objective function are all RMSE.

For AVEC2013 and AVEC2014, we set the window size for dividing the MFCCs to be 500 and 50 frames with 50% overlap. Due to space limitation, λ in Eq.(3) is set to 0.5 and the number of feature selected M is set to 1024 without displaying related experimental results. It is necessary to point out that the dimension of speech representation without feature selection is $3 \times 1024 = 3072$. In addition, we use LIBSVM [26] with the radial basis kernel for depression level prediction.

4.3. Results and Discussion

In this section, we firstly show the performance with different pooling strategies and network structures. Then, the comparison between our proposed method and the previous works is presented. Finally, we give the discussion for these results.

4.3.1. Experimental Performance of Different Pooling Strategies and Network Structures

Based on the above settings, we examine the performance of depression detection using max-pooling, average-pooling and the proposed ℓ_p-norm pooling combined with LASSO on the development sets of AVEC2013 and AVEC2014. The results are illustrated in Fig 4. Obviously, average-pooling ($p = 1$) is better than max-pooling ($p = \infty$), because the feature extracted by the hybrid network is not sparse, so max-pooling lose some useful information. Furthermore, the proposed ℓ_p-norm pooling combined with LASSO ($p = 4.06$ in AVEC2013 and $p = 2.13$ in AVEC2014) achieves the best detection accuracy. This can be explained that the pooling parameter p found in the process of solving the mapping relationship between speech features and BDI-II scores is suitable for the depression detection task.

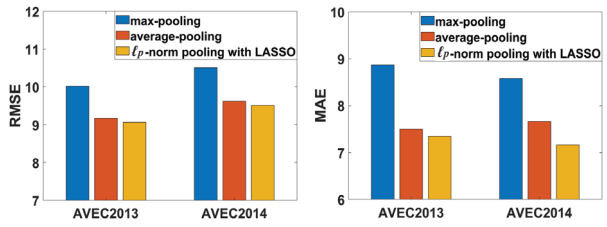

Figure 4: *Depression detection accuracy of three pooling strategies on the development sets of AVEC2013 and AVEC2014. (a) and (b) show two different indicators, respectively.*

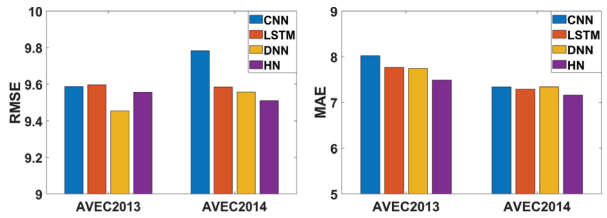

Figure 5: *Depression detection accuracy of four network structures on the development sets of AVEC2013 and AVEC2014. (a) and (b) show two different indicators, respectively. HN means the hybrid network.*

Besides, we compare the effects of CNN, LSTM, DNN and the hybrid network on depression detection. Fig. 5 shows the performance of them on AVEC2013 and AVEC2014 development sets. From them, it can be seen that LSTM gains better detection accuracy than CNN. Since MFCCs are essentially a sequence of time series vectors [27], the temporal changes can better reflect depression-related information than the spatial structure. While DNN extracts the discriminative representation through processing spatial-temporal feature of the MFCC, thus it has further improvement. On the whole, our proposed hybrid network acquires the best detection performance. The reasons are that we effectively integrate CNN, LSTM and DNN to extract the information in favor of depression detection from spatial structure, temporal changes and discriminative representation.

4.3.2. Comparison with Previous Works

We compare the proposed method with the previous works and the results are exhibited in Table 1 and 2. As shown, our method with feature selection is better than without feature selection on AVEC2013 and AVEC2014. The reasons lie in that the most helpful features for depression detection are selected by sorting the elements of $\mathbf{w}^*$ solved in Eq. (3).

Table 1: *Comparison with previous works on the AVEC2013 test set. Note that all methods only use speech data. FS means Feature Selection.*

Methods	RMSE	MAE
Valstar et al. [18]	14.12	10.35
Meng et al. [21]	11.19	9.14
He et al. [12]	10.00	8.20
Our method without FS	10.20	8.05
Our method with FS	**9.79**	**7.48**

Table 2: *Comparison with previous works on the AVEC2014 test set. Note that all methods only use speech data. FS means Feature Selection.*

Methods	RMSE	MAE
Valstar et al. [15]	12.56	10.03
Jain et al. [16]	10.25	8.40
He et al. [12]	9.99	8.19
Our method without FS	9.88	8.46
Our method with FS	**9.66**	**8.02**

From the presentation of these two tables, our method with feature selection provides the best prediction accuracy. On the one hand, we use the hybrid network to automatically extract the high-level feature about depression from MFCCs, which have better representation ability than those hand-crafted features used in the works of [18, 21, 15, 16]. Although CNN is applied in [12], it is limited in temporal information extraction. On the other hand, we optimize Eq. (3) to find the adaptive pooling parameter p^*, which is more suitable for depression detection than average-pooling used in [18, 15]. Jain et al. [16] gain the speech representation via Fisher Vector encoding, but they don't optimize the number of Gaussian functions to adapt to depression detection task.

5. Conclusions

Physiological studies have shown that MFCC is a discriminative biomarker to detect depression. Based on the facts, we use hybrid network to extract the high-level feature reflecting depression in the spatial structure, temporal changes and discriminative representation of short-term MFCC segments. Then, ℓ_p-norm pooling combined with LASSO is proposed to optimize the pooling parameter for depression detection to generate the utterance-level feature. Experimental results on AVEC2013 and AVEC2014 indicate that our method is superior than the previous approaches. In the future, we will consider other modal to further enhance the performance of depression prediction.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61771472), and the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100).

7. References

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

[2] W. H. Organization *et al.*, "Depression and other common mental disorders: global health estimates," 2017.

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

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

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

[6] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of affective disorders*, vol. 225, pp. 214–220, 2018.

[7] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena, "The sri avec-2014 evaluation system," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 93–101.

[8] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke, "Glottal source features for automatic speech-based depression assessment." in *INTERSPEECH*, 2017, pp. 2700–2704.

[9] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5795–5799.

[10] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification." in *INTERSPEECH*, 2016, pp. 485–489.

[11] C. A.MCPHERSON, "A narrative review of the beck depression inventory (bdi) and implications for its use in an alcohol-dependent population," *Journal of Psychiatric and Mental Health Nursing*, vol. 17, no. 1, pp. 19–30, 2010.

[12] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, 2018.

[13] M. N. Stolar, M. Lech, and N. B. Allen, "Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 987–991.

[14] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2018.

[15] 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," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[16] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 87–91.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[18] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Openearintroducing the munich open-source emotion and affect recognition toolkit," in *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*. IEEE, 2009, pp. 1–6.

[20] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[21] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 21–30.

[22] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, 2018.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Advances in Neural Information Processing Systems(NIPS)*, pp. 2017–2025, 2015.

[24] HassanEZZAIDI and JeanROUAT, "Comparison of mfcc and pitch synchronous am, fm parameters for speaker identification," in *INTERSPEECH*, 2000, pp. 318–321.

[25] M. Ranzato, C. Poultney, S. Chopra, and Y. Lecun, "Efficient learning of sparse representations with an energy-based model," *Advances in Neural Information Processing Systems(NIPS)*, pp. 1137–1144, 2007.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[27] D. Garreau, R. Lajugie, S. Arlot, and F. Bach, "Metric learning for temporal sequence alignment," *Advances in Neural Information Processing Systems(NIPS)*, pp. 1817–1825, 2014.