INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Generalized Dilated CNN Models for Depression Detection Using Inverted Vocal Tract Variables

Nadee Seneviratne, Carol Espy-Wilson

University of Maryland - College Park, USA
nadee@umd.edu, espy@umd.edu

Abstract

Depression detection using vocal biomarkers is a highly researched area. Articulatory coordination features (ACFs) are developed based on the changes in neuromotor coordination due to psychomotor slowing, a key feature of Major Depressive Disorder. However findings of existing studies are mostly validated on a single database which limits the generalizability of results. Variability across different depression databases adversely affects the results in cross corpus evaluations (CCEs). We propose to develop a generalized classifier for depression detection using a dilated Convolutional Neural Network which is trained on ACFs extracted from two depression databases. We show that ACFs derived from Vocal Tract Variables (TVs) show promise as a robust set of features for depression detection. Our model achieves relative accuracy improvements of $\sim 10\%$ compared to CCEs performed on models trained on a single database. We extend the study to show that fusing TVs and Mel-Frequency Cepstral Coefficients can further improve the performance of this classifier.

Index Terms: Depression, vocal tract variables, articulatory coordination, dilated CNN, generalizability

1. Introduction

Major Depressive Disorder (MDD) is a mental health disorder that is characterized by long-lasting depressed mood or loss of interest in activities that will cause significant impairment in daily life. There are more than 264 million people worldwide who suffer from depression [1]. The serious consequences of MDD such as suicidality necessitates the need of reliable automated solutions that could help clinicians and therapists diagnose and treat MDD patients early and effectively and help patients in monitoring themselves. Previous studies have shown that vocal biomarkers developed using prosodic, source and spectral features [2, 3, 4] can be useful in depression detection. There are multiple studies that have performed the detection of depression using various combinations of speech features such as Mel Frequency Cepstral Coefficients (MFCCs), formants and voice quality features [5, 6]. Prior to the rise of deep learning based approaches using network architectures such as Convolutional Neural Networks (CNNs) or Long Short Term Memory (LSTM) networks, earlier studies employed less data intensive machine learning models such as Support Vector Machines (SVM) or Gaussian Mixture Models (GMM).

Several recent studies found that successful results can be achieved by quantifying the changes in articulatory coordination to distinguish depressed speech from non-depressed speech [7, 8, 9, 10]. These differences in the timing of speech gestures are caused by a neurological phenomenon called psychomotor slowing, which is identified as a major characteristic of depression [11]. It is viewed as a necessary feature of MDD and a key component in evaluating severity of depression [12, 13]. Changes caused in speech due to psychomotor slowing such as more and longer pauses, slowed responses and monotonic phrases [14] lead to the usage of Articulatory Coordination Features (ACFs) to evaluate the severity of depression. ACFs are found to effectively capture information that can distinguish depressed speech from non-depressed speech using the multi-scale structure of correlations among the time series signals. This approach was predominantly validated using acoustic features such as formants and MFCCs as a proxy for underlying articulatory coordination [8, 15]. In our previous studies [9, 10] we showed that ACFs derived from direct articulatory speech features known as Vocal Tract Variables (TVs) are more effective in classifying depressed speech and non-depressed speech. These studies used the eigenspectra derived from the time-delay embedded correlation matrices as ACFs. An SVM classifier was used due to the limited availability of data. The magnitudes of the eigenvalues of the eigenspectra derived from the time-delay embedded correlation matrix was used as a measure of the complexity of articulatory coordination. The study in [16] explains that this channel-delay correlation matrix can be further optimized to eliminate repetitive sampling and matrix discontinuities. A more effective and scalable representation for ACFs was proposed in [16] utilizing dilated CNNs and incorporating more delays to the correlation matrix.

With the advent of Deep Neural Networks (DNNs), its applications in speech based depression detection and severity prediction increased rapidly, yielding promising results [17, 18, 19, 20]. However, the generalizability of these models is limited provided that these studies were performed on a single database. The characteristics of available depression databases differ depending on the acoustic variability in the speech recordings due to different speech types (free, read, sustained vowels etc.), speaking styles and rates, speaker demographics and different types of studies (observational studies/clinical trials), etc. Thus, findings from one study may not always be observed across different databases even within the same language. Therefore, the need to develop more generalized models prevails. Domain adaptation techniques have been explored to address this issue [21].

In this paper, we present a novel approach of using these 'direct' articulatory parameters (TVs) in a deep learning setting for the first time to detect depression. We combine speech data from two depression databases with different characteristics to develop generalized CNN models to detect the presence of depression. Using the approach proposed in [16], we show that robust and generalized Dilated CNN models can be developed using the TV based ACFs to perform this task.

2. Database Descriptions

Speech data from two databases [22, 23] were combined for our experiments. We encountered two clinician (CL)-rated depression assessment scales used in these databases: Hamilton

http://dx.doi.org/10.21437/Interspeech.2021-1960

Depression Rating Scale (HAMD) and Quick Inventory of Depressive Symptomatology (QIDS). The severity level definition for each class can be found in Table 1. Data in levels 2-5 is combined for the 'depressed' category and data in level 1 is used for the 'non-depressed' category.

Table 1: *Severity level definitions of MDD assessment scales*

Severity Level	HAMD	QIDS
1. Normal	0 – 7	0 - 5
2. Mild	8 - 13	6 - 10
3. Moderate	14 - 18	11 - 15
4. Severe	19 - 22	16 - 20
5. Very Severe	23 - 52	21 - 27

Table 2: *Details of Depression Databases*

Database	MD-1 [22]	MD-2 [23]
Longitudinal	6 Weeks	4 Weeks
# Subjects	20 F, 15 M	104 F, 61 M
Demography	31 Caucasian	125 Caucasian
	1 African American	26 African American
	1 Bi-racial	4 Asian
	1 Greek, 1 Hispanic	10 Other
Assessment	HAMD-CL: Bi-weekly	HAMD-CL, QIDS-CL: Weeks 1,2,4
FS Lengths	Min: 2.5s, Max: 156.8s	Min: 2.6s, Max: 181.2s
Recording Type	Interactive Voice Response Technology (8kHz)	

Details of the depression databases are given in Table 2. In addition, MD-1 is an observational study where patients started on pharmacotherapy and/or psychotherapy treatment for depression close to the beginning of the study. MD-2 is a clinical trial where subjects weren't taking psychotropic medications at the beginning of the study and started on 50 mg/day of sertraline or placebo (double-blind and randomized) at baseline. In this study, we used recordings of free speech (FS) where patients describe how they feel emotionally, physically and their ability to function each week. Additionally these databases contain read speech recordings which were not used for this study. In MD-2, depression assessment scores were provided to only 105 subjects. Due to the availability of two CL-rated scores in MD-2, only the speech samples where both the scores belong to the same category were used.

3. Articulatory Coordination Features

3.1. Estimation of Vocal Tract Variables (TVs)

In Articulatory Phonology (AP) [24], speech is viewed as a constellation of overlapping gestures. These gestures are discrete action units whose activation results in constriction formation or release by five distinct constrictors along the vocal tract: lips, tongue tip, tongue body, velum, and glottis. The TVs are defined by the constriction degree and location of these five constrictors. A speaker independent, DNN based Speech Inversion (SI) System is used to compute the trajectory of the TVs that represent constriction location and degree of articulators located along the vocal tract [25]. The model was trained using the Wisconsin X-Ray Microbeam (XRMB) database [26]. The six TVs estimated by the SI system are – Lip Aperture, Lip Protrusion, Tongue Body Constriction Location, Tongue Body Constriction Degree, Tongue Tip Constriction Location and Tongue Tip Constriction Degree. For detailed information about the SI system, the reader is referred to [25].

Glottal TV Estimation: For a complete representation of TVs described in the AP, TVs related to the glottal state need to be included. Due to the difficulty in placing sensors near the glottis to acquire ground-truth information, the DNN based system could not be trained using articulatory data. Hence, we used the periodicity and aperiodicity measure obtained from the Aperiodicity, Periodicity and Pitch detector developed in [27]. This program estimates the proportion of periodic energy and aperiodic energy and pitch of a speech signal based on the distribution of the minima of the average magnitude difference function of the speech signal. In [10], these glottal parameters boosted the classification accuracies for depressed and non-depressed speech classification by about 8%.

3.2. MFCCs Estimation

We use ACFs derived from MFCCs as a proxy for actual articulatory features to enable fair comparisons with the experiments conducted using TV based ACFs. For this, 12 MFCC time series were extracted by using an analysis window of 20 ms with a 10 ms frame shift (1st MFCC coefficient was discarded).

3.3. Channel-delay Correlation Matrix

Conventionally for an M-channel feature vector (TVs or MFCCs), the channel-delay correlation matrix in [10] has dimensionality of $(kM \times kM)$ for $k(15)$ time delays per channel with a fixed delay scale (7 samples). In order to incorporate multiple delay scales p, these matrices computed at different delay scales will be stacked yielding a $p \times kM \times kM$ dimensional matrix [8]. In [16], the authors propose a novel method to construct the channel-delay correlation matrix that overcomes the limitations found in the conventional approach such as repetitive sampling and matrix discontinuities at the borders of adjacent sub-matrices.

In this work, we have adopted this new structure of channel-delay correlation matrix. For an M-channel feature vector $\mathbf{X}$, the delayed correlations ($r^d_{i,j}$) between i^{th} channel $\mathbf{x_i}$ and j^{th} channel $\mathbf{x_j}$ delayed by d frames, are computed as:

$$r^d_{i,j} = \frac{\sum_{t=0}^{N-d-1} x_i[t]x_j[t+d]}{N-|d|} \tag{1}$$

where N is the length of the channels. The correlation vector for each pair of channels with delays $d \in [0, D]$ frames will be constructed as follows:

$$R_{i,j} = \begin{bmatrix} r^0_{i,j}, & r^1_{i,j}, & \ldots & r^D_{i,j} \end{bmatrix}^T \in \mathbb{R}^{1\times(D+1)} \tag{2}$$

The delayed auto-correlations and cross-correlations are stacked to construct the channel-delay correlation matrix:

$$\widetilde{R}_{ACF} = \begin{bmatrix} R_{1,1} & \ldots & R_{i,j} & \ldots & R_{M,M} \end{bmatrix}^T \in \mathbb{R}^{M^2\times(D+1)} \tag{3}$$

It is important to note that the $\widetilde{R}_{ACF}$ matrix contains every correlation only once. With this representation, information pertaining to multiple delay scales can be incorporated into the model by using dilated CNN layers with corresponding dilation factors while maintaining a low input dimensionality. Each $R_{i,j}$ will be processed as a separate input channel in the CNN model, and thereby overcoming discontinuities.

4. Dilated CNN Classifier

The Dilated CNN architecture proposed in [16] was adopted in the experiments (Fig. 1). The input $\widetilde{R}_{ACF}$ is fed into four parallel convolutional layers $(C1, C2, C3, C4)$ with different dilation rates $n = \{1, 3, 7, 15\}$ and a kernel size of $(15, 1)$ which resembles the multiple delay scales in the conventional approach. The outputs of these four parallel layers are concatenated and then passed through two sequential convolutional layers $(C5, C6)$. This output is flattened and passed through

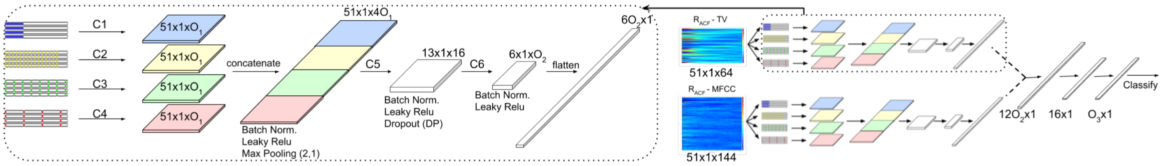


Figure 1: *Dilated CNN Architecture for TV and MFCC Fused Model*

two fully connected (dense) layers ($D1$, $D2$) to perform 2-class classification in the output layer. All convolutional layers used LeakyRelu activation, whereas dense layers used Relu activation with l_2 regularization of $\lambda = 0.01$. Batch Normalization, dropouts, and max-pooling layers were added as shown in the Fig. 1. The weight sharing nature of CNNs handles the high dimensional correlation matrices with a low number trainable parameters.

5. Experiments and Results

5.1. Data Preparation

We conducted our experiments using free speech since previous work [10] has shown a larger difference of articulatory coordination between depressed and non-depressed speech for free speech than for read speech. This difference is probably due to the increased cognitive load associated with free speech. To increase the number of samples to train the CNN model and to make the model resilient to input translations, we segmented the audio recordings that are longer than 20s into segments of 20s with a shift of 5s. Recordings with duration less than 10s were discarded and other shorter recordings (between 10s-20s) were used as they were. Table 3 summarizes the total duration of speech data available after the segmentation. The speech samples were split into train / validation / test splits ($80 : 10 : 10$) preserving a similar class distribution in each split and ensuring that there are no speaker overlaps in the splits. The class imbalance issue is addressed during the training process by assigning class-weights to both training and validation sets. All audio segments were normalized to have a maximum absolute value of 1, prior to low-level feature extraction.

CNN models were trained using two sets of features: TVs corresponding to constriction degree and location and glottal parameters - 8 channels and MFCCs - 12 channels. Before computing the correlation matrices, feature vectors were standardized individually.

5.2. Model Training

All models were trained on $D = 50$ which was empirically determined and a learning rate of $1e-5$. The models were optimized using an Adam Optimizer for Binary Cross Entropy loss. For $C5$, kernel size was $(3, 1)$ with a stride of 2 and 16 output filters were used. For $C1-C5$, 'same' padding was used and for $C6$ 'valid' padding was used. All inputs were standardized using the mean and the standard deviation of the training data. All models were trained with an early stopping criteria based on validation loss (patience 15 epochs) for a maximum of 300 epochs. To evaluate the performance of models, overall accuracy, area under the receiver operating characteristics curve (AUC-ROC), and F1 scores were used.

Table 3: *Duration of Available Data in hours*

Database	Depressed	Non-depressed
MD-1	11.78	2.45
MD-2	15	1.2
MD-1&2	26.78	3.65

Table 4: *Grid Search Parameters - Best Model (for MD1&2)*

	C1-C4 Filter Outputs (O_1)	C6 Filter Output (O_2)	C6 Kernel (K_1)	D2 Output Size (O_3)	Dropout Prob. (DP)
Range	{16,32}	{8,16}	{(3,1), (4,1)}	{8,16}	{0.4,0.5}
TV	32	16	(3,1)	8	0.5
MFCC	16	8	(3,1)	16	0.5
Fused	32	8	(3,1)	8	0.5

Grid search was performed to tune the hyper-parameters using the ranges in Table 4. Models were trained with and evaluated on train-test combinations as shown in Table 5. Note that MD-1&2 represents the combination of MD-1 and MD-2 data.

5.3. Classification Results

The results are included in Table 5. We evaluate the performance of our models using AUC-ROC for fair comparison. Accuracy alone could not be used as an indication of model strength due to class imbalance. Compared to cross corpus evaluations (CCEs) of TV based models trained on a single database (i.e. train/test pairs of MD-2/MD-1 & MD-1/MD-2), the AUC-ROCs of the model trained on combined data (MD-1&2/MD-1 & MD-1&2/MD-2) indicate relative improvements of ∼19% and ∼14% respectively. For MFCC based models, the corresponding relative improvements are ∼12% and ∼15%. Therefore we believe the model trained on combined data performs better compared to single corpus models. Further, the MFCC based combined model seems to outperform MFCC based models that were trained and evaluated on the same database through learning on additional data. Due to generalization it is possible that with the combined data model performance for a particular database be sometimes slightly lower compared to the models trained solely on the same database as can be seen in TV based models.

5.4. Classification using Fused TVs and MFCCs

A fused model was trained combining TVs and MFCCs in order to investigate the inter-learning among ACFs derived from different feature vectors. Adopting a late fusion strategy, the model consists of two parallel structures similar to the previous model up to $C6$ using ACFs derived from MFCCs and TVs as inputs. Late fusion strategy was adopted to avoid the input fea-

Table 5: *Results for 2-Class Classifications*

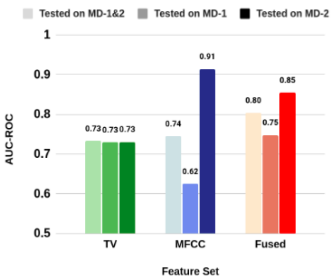
Feats	Train	Test	Accuracy	AUC-ROC	F1(D)/F1(ND)
TV based ACF	MD-1&2	MD-1&2	82.67%	0.7329	0.9/0.42
	MD-1&2	MD-1	78.46%	0.7298	0.87/0.44
	MD-1&2	MD-2	86.39%	0.7309	0.92/0.39
	MD-1	MD-1	82.31%	0.7757	0.9/0.41
	MD-1	MD-2	78.57%	0.6399	0.87/0.26
	MD-2	MD-2	85.71%	0.8014	0.92/0.43
	MD-2	MD-1	71.92%	0.6127	0.82/0.37
MFCC based ACF	MD-1&2	MD-1&2	81.77%	0.7447	0.89/0.36
	MD-1&2	MD-1	74.62%	0.6249	0.85/0.3
	MD-1&2	MD-2	88.10%	0.9141	0.93/0.44
	MD-1	MD-1	76.54%	0.6190	0.86/0.33
	MD-1	MD-2	81.29%	0.7923	0.89/0.25
	MD-2	MD-2	80.61%	0.7353	0.89/0.28
	MD-2	MD-1	78.46%	0.5583	0.88/0.18

ture dimensions being unnecessarily large which may lead to overfitting issues.

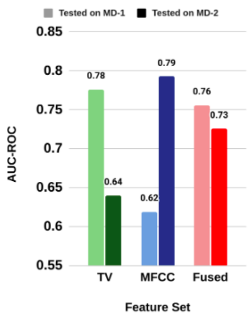
The outputs of the $C6$ layers of both structures were flattened and concatenated and then passed through two dense layers to perform the classification. Similar to section 5.3, the general trend of the model trained on combined data performing better than the CCEs of single corpus models is evident through the relative AUC-ROC improvements of ∼14% and ∼18% for MD-1 data and MD-2 data respectively (Table 6). Additionally, the fused model generally shows better AUC-ROC values when trained on the combined data compared to TV based and MFCC based models.

Table 6: *Results for Fused Models Classifications*

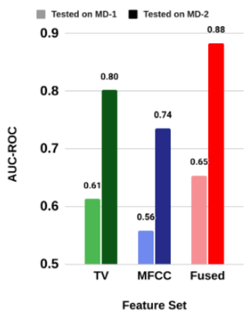
Train	Test	Accuracy	AUC-ROC	F1(D)/F1(ND)
MD-1&2	MD-1&2	87.73%	0.8036	0.93/0.39
MD-1&2	MD-1	83.07%	0.7479	0.9/0.39
MD-1&2	MD-2	91.84%	0.8546	0.96/0.4
MD-1	MD-1	83.84%	0.7555	0.9/0.48
MD-1	MD-2	82.65%	0.7254	0.9/0.11
MD-2	MD-2	77.21%	0.8824	0.86/0.36
MD-2	MD-1	71.54%	0.6535	0.82/0.31



(a)



(b)



(c)

Figure 2: *AUC-ROC for Models Trained on (a) MD-1&2, (b) MD-1 only and (c) MD-2 only when evaluated on MD-1&2 (for (a) only), MD-1 and MD-2 for 3 Feature Sets.*

6. Discussion

In this paper we showed that Dilated CNN based models trained for depression detection can be generalized by combining data from multiple databases using a concise representation of channel-delay correlation matrices of feature vectors. Relatively high AUC-ROC variance of MFCC based models across different experiments indicate that the performance of MFCC based models are database-dependent compared to TV based models. The TV based models seem to consistently perform with lower AUC-ROC variance (Figure 2a). Therefore, ACFs derived from TVs are desirable in general compared to those from MFCCs, indicating that TV based ACFs could act as more robust features in the generalization process. This observation is further supported as TVs provide a direct articulatory representation. Further fused model experiment suggests that combining both TVs and MFCCs helps to boost the classification performance of the generalized model.

We observe that TV based models perform better on MD-1 data while MFCC based models perform better on MD-2 data. Fused models seem to capture the strengths of both TVs and MFCCs and thereby perform better on both MD-1 and MD-2 data. This result suggests that ACFs derived from different features complement each other.

The findings of the cross corpus studies are encouraging. The TV based models perform significantly better than the MFCC based models when trained and evaluated on the same database (Figures 2b and 2c). On the other hand, MFCC based models show higher accuracy gains when evaluated across databases. However, we note that the minority class (non-depressed) F1 scores are lower, suggesting that the model is biased towards the majority class given the class imbalance in the test data. In comparison TV based models have higher F1 scores for the minority class suggesting that they generalize better on unseen data. This also leads to the hypothesis that the normalized TVs can provide a more speaker-independent representation compared to the MFCCs given that the TV based articulatory coordination features provide a more stable result across the two databases compared to MFCCs.

7. Conclusions and Future Work

In this paper we use TV based ACFs in a deep learning setting for the first time to establish that the robustness of TV based ACFs holds across databases and the results are generalizable. Further, we proposed a fused model which harnesses the information captured in TV based and MFCC based ACFs. Previous work on binary classification of depression using these two databases is not directly comparable given that the score thresholds for the two classes used in those work are different [28, 29] and experiments were conducted using only MD-1. To the best of our knowledge, this is the first time the MD-2 database has been used in the work presented at Interspeech. Hence, we plan to expand the study to compare the performance of different features such as formants and openSMILE features [30] etc. In the future, this study can be extended to more databases. In the case of different MDD assessment scales, consideration needs to be given to achieve uniformity in establishing ground-truth.

8. Acknowledgements

We would like to thank Dr. James Mundt for the depression databases MD-1&2 [22, 23]. We also thank Dr. Thomas Quatieri and Dr. James Williamson for granting access to the MD-2 database [23] which was funded by Pfizer. This work was supported by the UMCP & UMB Artificial Intelligence + Medicine for High Impact Challenge Award and a Faculty Student Research Award from UMCP.

9. References

[1] World Health Organization (WHO), “Depression,” 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10 – 49, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639315000369

[3] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 4 2013, pp. 1–8.

[4] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 857–861, 01 2013.

[5] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization." *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2997–3000, 01 2011.

[6] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 1–9, 09 2018.

[7] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 65–72. [Online]. Available: https://doi.org/10.1145/2661806.2661809

[8] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," *Computer Speech & Language*, vol. 55, pp. 40 – 56, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230817303510

[9] C. Espy-Wilson, A. C. Lammert, N. Seneviratne, and T. F. Quatieri, "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," in *Proc. Interspeech 2019*, 2019, pp. 1448–1452. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1815

[10] N. Seneviratne, J. R. Williamson, A. C. Lammert, and T. F. Q. C. Espy-Wilson, "Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression," in *Proc. Interspeech 2020*, 2020.

[11] J. R. Whitwell, *Historical notes on psychiatry*. Oxford, England, 1937.

[12] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC, 2000. [Online]. Available: https://dsm.psychiatryonline.org/doi/abs/10.5555/appi.books.9780890425596.x00pre

[13] D. J. Widlöcher, "Psychomotor retardation: Clinical, theoretical, and psychometric aspects," *Psychiatric Clinics of North America*, vol. 6, no. 1, pp. 27 – 40, 1983, recent Advances in the Diagnosis and Treatment of Affective Disorders. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0193953X18308384

[14] C. Sobin and H. Sackeim, "Psychomotor symptoms of depression," *The American journal of psychiatry*, vol. 154, pp. 4–17, 02 1997.

[15] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 11–18. [Online]. Available: https://doi.org/10.1145/2988257.2988263

[16] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNS for depression detection in naturalistic environments," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 6549–6553. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9054323

[17] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 45–51. [Online]. Available: https://doi.org/10.1145/3133944.3133950

[18] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 65–71. [Online]. Available: https://doi.org/10.1145/3347320.3357696

[19] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 81–88. [Online]. Available: https://doi.org/10.1145/3347320.3357697

[20] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7159–7163.

[21] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, "Domain Adaptation for Enhancing Speech-Based Depression Detection in Natural Environmental Conditions Using Dilated CNNs," in *Proc. Interspeech 2020*, 2020, pp. 4561–4565. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3135

[22] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0911604406000303

[23] 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, novel Pharmacotherapies for Depression. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006322312002636

[24] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[25] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019. [Online]. Available: https://doi.org/10.1121/1.5116130

[26] J. R. Westbury, "Speech Production Database User ' S Handbook," *IEEE Personal Communications - IEEE Pers. Commun.*, vol. 0, no. June, 1994.

[27] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 776–786, 9 2005.

[28] B. Helfer, T. Quatieri, J. Williamson, D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2172–2176, 01 2013.

[29] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5 2013, pp. 7542–7546.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246