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

Speech based Depression Severity Level Classification Using a Multi-Stage Dilated CNN-LSTM Model

Nadee Seneviratne, Carol Espy-Wilson

University of Maryland - College Park, USA

nadee@umd.edu, espy@umd.edu

Abstract

Speech based depression classification has gained immense popularity over the recent years. However, most of the classification studies have focused on binary classification to distinguish depressed subjects from non-depressed subjects. In this paper, we formulate the depression classification task as a severity level classification problem to provide more granularity to the classification outcomes. We use articulatory coordination features (ACFs) developed to capture the changes of neuromotor coordination that happens as a result of psychomotor slowing, a necessary feature of Major Depressive Disorder. The ACFs derived from the vocal tract variables (TVs) are used to train a dilated Convolutional Neural Network based depression classification model to obtain segment-level predictions. Then, we propose a Recurrent Neural Network based approach to obtain session-level predictions from segment-level predictions. We show that strengths of the segment-wise classifier are amplified when a session-wise classifier is trained on embeddings obtained from it. The model trained on ACFs derived from TVs show relative improvement of 27.47% in Unweighted Average Recall (UAR) at the session-level classification task, compared to the ACFs derived from Mel Frequency Cepstral Coefficients (MFCCs).

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

1. Introduction

With more than 264 million people suffering worldwide [1], Major Depressive Disorder (MDD) is one the most critical mental health disorders that affects the quality of life. MDD can even lead to suicidality and that urges the requirement of timely diagnosis and prompt treatments. Previous studies have shown that vocal biomarkers developed using prosodic, source, and spectral features [2, 3, 4] can be very useful in depression detection and severity prediction.

Articulatory Coordination Features (ACFs) have yielded successful results in distinguishing depressed speech from non-depressed speech by quantifying the changes in timing of speech gestures [5, 6, 7, 8]. These changes in articulatory coordination happens as a result of neurological condition called psychomotor slowing, a necessary feature of MDD that is used to evaluate the severity of MDD [9, 10, 11]. Previously, the correlation structure of the formants or MFCCs were used as a proxy for articulatory coordination to derive indirect ACFs which showed promise in the depression detection task [6]. Authors of this paper showed in their previous work, that by using Vocal Tract Variables (TVs) as a direct measure of articulation to quantify changes in the way speech is produced by depressed and non-depressed subjects can yield significantly better results in depression detection task [7, 8]. In recent work, the authors applied the time-delay embedded correlation matrix derived from TVs as ACFs to train a generalized deep learning based model for the first time with speech data sourced from two depression databases with different characteristics . It was shown that TV based ACFs show promise as a robust set of features for depression by generalizing well across the two databases [12].

While a lot of studies have looked into predicting the depression severity scores using regression [13, 14], most previous studies on depression classification focused on detecting whether a subject is depressed or not [15, 16] or detecting high or low depression [3, 17]. A very few studies have looked into performing classifications across more than just 2 classes [18, 19], but not using deep learning based models. Hence we extend our work to perform a depression severity level classification across 3 classes (normal, moderate and severe) using TV based ACFs. This helps to identify those who are at critical stages with severe depression, allowing to prioritize the allocation of limited resources. Then we propose a multi-stage model to perform session-wise classifications using segment-level classifications. We show that this technique can result in significant improvements in final classifications than training models using features extracted directly from full audio recordings by helping to avoid overfitting issues due to high dimensionality of the input features and low amount of training samples. We perform experiments using multiple feature sets (MFCCs, Formants, openSMILE features) for comparisons.

The paper is organized as follows: section 2 explains the methodology involving feature extraction and model architectures. Section 3 presents the experiments conducted and results obtained. Section 4 analyses the results in detail with potential future directions.

2. Method

2.1. Depression Databases

We use speech data from two depression databases [20, 21] (Table 2). Two clinician (CL)-rated depression assessment scales: the 17-item Hamilton Depression Rating Scale (HAMD) [22] and the 16-item Quick Inventory of Depressive Symptomatology (QIDS) [23] were provided which were used to define the severity levels of depression (Table 1). For the 3-class severity level classification task, data in levels 4-5 and 2-3 for was combined for classes 'severe' and 'moderate' respectively. Data in level 1 was used for 'normal' class.

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

In this study, we used recordings of free speech (FS) where

http://dx.doi.org/10.21437/Interspeech.2021-1967

Table 2: *Details of Depression Databases*

Database	**MD-1 [20]**	**MD-2 [21]**
Longitudinal	6 Weeks	4 Weeks
Study Type	Observational	Clinical trial
# Subjects	20 F, 15 M	104 F, 61 M
Demography	31 Caucasian 1 African American 1 Bi-racial 1 Greek, 1 Hispanic	125 Caucasian 26 African American 4 Asian 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)	

patients describe how they feel emotionally, physically and their ability to function.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 severity level were used.

2.2. Estimation of Vocal Tract Variables (TVs)

Acoustic-to-Articulatory Speech Inversion(SI): We use the TVs estimated by a speaker independent, deep neural network (DNN) based SI system to estimate the low-level feature vectors used to compute ACFs. Articulatory Phonology (AP) [24] views speech 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 vocal tract variables (TVs) are defined by the constriction degree and location of these five constrictors. The SI system computes the trajectory of the TVs that represent constriction location and degree of articulators located along the vocal tract [25]. 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: To achieve a complete representation of TVs described by the AP, TVs corresponding to the glottal state should also be included. The same DNN based SI system could not be trained due to unavailability of ground-truth articulatory data as a result of difficulty in placing sensors near the glottis. Therefore, we augment the previous 6 TVs with the periodicity and aperiodicity measures obtained from the Aperiodicity, Periodicity and Pitch detector [26] which are used as glottal TVs. In [8], these glottal parameters boosted the classification accuracies for depressed and non-depressed speech classification by about 8%.

2.3. Other Feature Extraction

In order to compare the performance of the model trained using TV based ‘direct’ ACFs, we also trained models with the same architecture using widely used MFCCs and Formants based ACFs which are considered as proxies for actual articulatory coordination. 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). The first three formant frequencies were extracted using Praat [27]. Settings were: tracking 5 formants, 5500 Hz maximum formant, window length of 25ms and time step of 10ms.

To train a baseline model for comparison purposes, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [28] was extracted using the openSMILE toolkit [29]. This 23-dimensional feature set consists of prosodic features like pitch, loudness, jitter, shimmer and other spectral parameters. The motivation behind using this feature set as the baseline case is because previous work on speech based depression recognition has shown that prosodic features like reduced speaking intensity, reduced pitch range and slower speech are characteristics of depressed speech [30]. These features were computed with a window size of 20ms with an overlap of 10ms.

2.4. Articulatory Coordination Features

We use the channel-delay correlation matrix proposed in [31] as the ACFs for this work. This matrix overcomes the limitations found in the conventional approach [6] such as repetitive sampling and matrix discontinuities at the borders of adjacent sub-matrices.

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:

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

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:

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

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

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

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. Before computing the ACFs, feature vectors were standardized individually.

2.5. Baseline Model

We trained a CNN using openSMILE features to be used as a baseline model for the session-wise classification. The input is passed through two sequential 1-D (across time axis) convolutional layers ($Conv1$, $Conv2$). Each convolutional layer is followed by batch normalization, leaky ReLU activation, dropouts (D_1, D_2) and a max-pooling layer ($MP1$, $MP2$). The output from the second max-pooling layer is flattened and passed through a fully connected layer ($Dense1$) with ReLU activation with l_2 regularization of $\lambda = 0.01$ to perform 3-class classification in the output layer.

The output of the $Dense1$ dense layer is extracted and used as the input to the LSTM model described in section 2.6 to obtain session-wise classifications.

2.6. Multi-Stage Dilated CNN-LSTM Model

A multi-stage approach to obtain session-wise classification predictions from segment-wise predictions is motivated by the fact that the original number of speech recordings were insufficient to train a deep neural network with high-dimensional input features (since each recording serves as a training sample). To overcome this drawback, we propose a two-stage neural network architecture.

In the first stage, a **dilated CNN** proposed in [31] was trained using the ACFs to predict the segment-wise classifica-

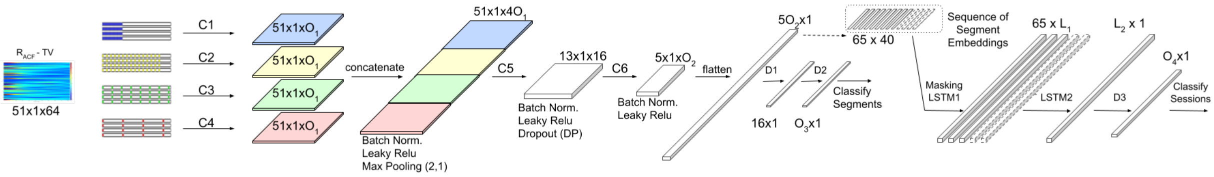


Figure 1: *Dilated CNN-LSTM Architecture for Session-Level Classification*

tions (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 two fully connected (dense) layers $(D1, D2)$ to perform segment-level 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.

The flattened output of $C6$ is passed as input to the second stage. Since the segments of the sessions can be represented in sequential order, these embeddings are passed through an **LSTM based RNN** model to perform the session-level classification. The input is first passed through two LSTM layers ($LSTM1$ with L_1 units, returning sequence outputs and $LSTM2$ with L_2 units), followed by a Dense layer ($D3$) with ReLU activation and O_4 output units. Finally, the output layer with Softmax activation performs the session-level classification. Recurrent dropout probabilities of DP_1 and DP_2 are applied to the two LSTM layers respectively.

3. Experiments & Results

3.1. Dataset Preparation

Originally there were 472 (35 speakers) and 753 (105 speakers) FS recordings from MD-1 and MD-2 respectively. The 140 speakers were divided into train / validation / test splits $(60 : 20 : 20)$ preserving a similar class distribution in each split and ensuring that there are no speaker overlaps. For the models trained on TV, MFCC and Formant based ACFs, to increase the number of samples in order to train a deep neural model and to make the model resilient to translation, we segmented the audio recordings in the train and validation splits 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 amount of speech data available after the segmentation. Test data was split into non-overlapping equal-length segments close to 20s. For the baseline model trained on openSMILE features, all these audio segments were truncated at 10s (minimum length of the available audio segments) to have fixed sized inputs to the CNN. Before extracting the low-level features, all audio segments were normalized to have a maximum absolute value of 1.

Table 3: *Available Data for the Dilated CNN model in hours/ # segments/ # sessions*

Database	Severe	Moderate	Normal
MD-1	4.13 / 763 / 43	6.57 / 1215 / 68	2.17 / 396 / 22
MD-2	8.8 / 1635 / 137	5.63 / 1044 / 95	0.87 / 164 / 17

3.2. Model Training

Baseline Model: Kernel size and number of output filters for $Conv1$ is 8 and 256 respectively and for $Conv2$ 8 and 128 respectively. The pool size of $MP1$ and $MP2$ is 8. D_1 and D_2 were tuned to 0.5 and 0.7 respectively. $Dense1$ has 64 units. These hyper-parameters were tuned using a grid search. A learning of $2e - 5$ was used. Each dimension of each input openSMILE feature vector was individually standardized.

Multi-Stage Dilated CNN-LSTM Model: Using a maximum delay D as 50 (empirically determined), the dilated CNN model was trained using a learning rate of $2e - 5$. 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 input ACFs were standardized using the mean and the standard deviation of the training data. The LSTM model was trained using an adaptive learning rate starting from $2e - 4$ and it was decayed by 50% every 10 epochs until it reached $2e - 5$. LSTM model hyper-parameters can be found in Table 4.

Table 4: *LSTM Based RNN Model Parameters*

	LSTM1 Units (L_1)	LSTM2 Units (L_2)	LSTM1 Dropouts (DP_1)	LSTM2 Dropouts (DP_2)	D3 Output Size O_4
TV	64	64	0.4	0.3	32
MFCC	128	64	0.6	0.4	64
Formants	128	64	0.7	0.7	16

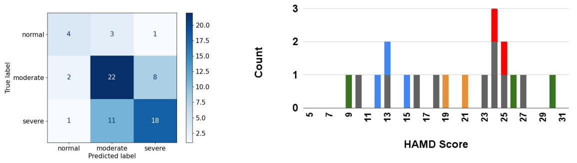
The models were optimized using an Adam Optimizer for the Categorical Cross Entropy loss. The models were trained with an early stopping criteria based on validation loss (patience 15 epochs) for a maximum of 300 epochs. Batch size of 128 was used. To address the class imbalance issue, class weights were assigned to both training and validation splits during the training process to both the models. To evaluate the performance of the model, overall accuracy, Unweighted Average Recall (UAR), and F1 scores were used. Grid search was performed to tune the hyper-parameters of the dilated CNN model using the ranges in Table 5.

Table 5: *Grid Search Parameters for Best Models*

	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	16	8	(4,1)	16	0.5
MFCC	32	16	(3,1)	8	0.5
Formants	32	8	(4,1)	16	0.4

3.3. Segment-Level Classification Results

Table 6 includes the accuracy and UAR results for the models trained using 4 sets of features. The model trained on TV based ACFs perform significantly better yielding a relative UAR improvement of 13.63% compared to the next best model that was trained on MFCC based ACFs. Performance of Formant based ACFs and openSMILE features seem very close in terms of UAR. The chance level F1 scores for normal (N), moderate (M) and severe (S) classes were 0.16, 0.38 and 0.39 respectively.



(a) *Confusion Matrix* (b) *HAMD of Misclassified Sessions*

Figure 2: *Analysis of Mis-classifications by TV based ACF Dilated CNN-LSTM Model*

Table 6: *Overall Results of Segment-Level Classification*

Model	Features	Accuracy	UAR	F1(N)/F1(M)/F1(S)
Dilated CNN	TV ACF	**50.84%**	**0.5010**	**0.33 / 0.50 / 0.58**
Dilated CNN	MFCC ACF	42.76%	0.4409	0.27 / 0.53 / 0.37
Dilated CNN	Formant ACF	38.22%	0.4089	0.26 / 0.42 / 0.40
CNN	Opensmile	40.20%	0.3980	0.21 / 0.42 / 0.48

3.4. Session-Level Classification Results

Using a low dimensional intermediate layer output of the segment-level classifier as en embedding, session-level RNNs were trained for all four feature sets. The overall session-level classification results can be found in Table 7. To benchmark the performance of the LSTM based session-level classifier, we used a conventional plurality voting approach where we used the mode of top 50% of the segment level predictions based on the confidence as the session-wise prediction (ties were broken by randomly selecting a class out of the tied classes). This was done based on the hypothesis that highly confident segment-level predictions can produce a more reliable final prediction.

The LSTM based RNN model trained using TV based ACFs achieved a relative improvement of 27.47% in UAR compared to the one trained using MFCC based ACFs. In general, the ACFs have yielded better UAR relative to the openSMILE features in session-level classification. Chance-level F1 scores for normal, moderate and severe classes for session-level classifications were 0.17, 0.39 and 0.38 respectively.

Table 7: *Overall Results of Session-Level Classification*

Model	Features	Accuracy	UAR	F1(N)/F1(M)/F1(S)
LSTM based RNN	TV ACF	**62.86%**	**0.5958**	**0.53 / 0.65 / 0.63**
	MFCC ACF	45.71%	0.4674	0.44 / 0.48 / 0.43
	Formant ACF	37.14%	0.4028	0.32 / 0.4 / 0.36
	Opensmile	37.14%	0.3729	0.21 / 0.39 / 0.43
Plurality Voting Classifier	TV ACF	52.86%	0.5861	0.4 / 0.43 / 0.29
	MFCC ACF	45.71%	0.5222	0.48 / 0.61 / 0.37
	Formant ACF	30.00%	0.3784	0.63 / 0.12 / 0.22
	Opensmile	42.86%	0.3243	0.0 / 0.42 / 0.51

4. Discussion

Based on the results reported in section 3.4, the ACFs derived from the TVs perform significantly better in this depression severity level classification task. This shows that using TVs as a direct measure for articulatory coordination can provide additional information in distinguishing different levels of depression severity. Our comparison among different feature sets helps to benchmark the performance of TV based ACFs with widely used other ACFs and openSMILE features.

The confusion matrix (Fig. 2a) obtained for the best performing model for the session-level classification shows that most of the confusions have occurred between adjacent classes (either normal-moderate or moderate-severe). We further analysed the HAMD scores of those sessions that were misclassified between the moderate and severe classes (Fig. 2b). There are instances where all or most of the sessions belonging to the same speaker (brightly-colored bars in the figure) are misclassified. Out of the 19 misclassified sessions, 10 (3-green, 3-blue, 2-red, 2-orange) belonged to 4 speakers. The remaining 9 gray colored bars are from 9 different speakers. This suggests that a speaker dependent system may perform better than a speaker independent system by learning unique characteristics inherent to a particular speaker.

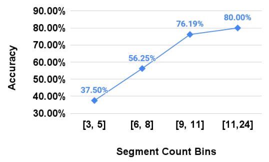


Figure 3: *Accuracy of session-wise classifications against the number of segments available for each session (TV based ACFs)*

The LSTM based RNN model for session-wise classifications is proposed to overcome the drawbacks in standard rule based approaches. A major drawback of these approaches is that there is no principled approach to choose rule thresholds. For instance, if we are using a plurality voting scheme which only considers K most confident predictions and K is chosen to maximize the metrics (K_{max}), the resultant scheme is biased towards the test set. Once K is chosen, there is no guarantee that it will generalize to unseen test data. An alternative would be to pick the segment level predictions where the confidence is greater than a certain threshold p_{th}, however it's possible that there are no segment level predictions with better confidence than the threshold. With the proposed approach, the performance of the model is comparable with the result achieved using K_{max} and the generalizability of the model is maintained.

Looking at the performance of the segment-level and session-level classifiers it is evident that the strengths of the segment-level classifier are amplified as a result of its repeated usage in the session-level classifier. This also suggests the notion that the session level classifier is stronger on sessions with more segments. This is empirically verified by calculating the accuracy for ranges of segment counts in a session (Fig. 3).

5. Conclusion and Future Work

In this paper we proposed a new multi-stage architecture trained on TV based ACFs for depression severity classification which clearly outperforms the baseline models. We also established that the robustness of ACFs based on TVs holds beyond mere detection of depression and even in severity level classification. In future, we will expand our work to perform severity level classification beyond 3 classes and to predict the depression severity scores. This work can be extended to develop a multi-modal system that can take advantage of textual information obtained through Automatic Speech Recognition tools. Linguistic features can reveal important information regarding the verbal content of a depressed patient relating to their mental health condition.

6. Acknowledgements

We would like to thank Dr. James Mundt for the depression databases MD-1&2 [20, 21]. We also thank Dr. Thomas Quatieri and Dr. James Williamson for granting access to the MD-2 database [21] 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.

7. 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] 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

[6] 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

[7] 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

[8] N. Seneviratne, J. R. Williamson, A. C. Lammert, T. F. Quatieri, and C. Espy-Wilson, "Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression," in *Proc. Interspeech 2020*, 2020, pp. 4551–4555. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2758

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

[10] 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

[11] "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.

[12] N. Seneviratne and C. Espy-Wilson, "Generalized dilated cnn models for depression detection using inverted vocal tract variables," in *Interspeech 2021*, arXiv:2011.06739.

[13] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction." New York, NY, USA: Association for Computing Machinery, 2019.

[14] 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.

[15] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," in *Proc. Interspeech 2017*, 2017, pp. 3266–3270.

[16] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," 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. 35–42. [Online]. Available: https://doi.org/10.1145/2988257.2988267

[17] 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.

[18] A. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–18, 2011.

[19] 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.

[20] 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

[21] 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

[22] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 23, no. 1, pp. 56–62, 1960. [Online]. Available: https://jnnp.bmj.com/content/23/1/56

[23] "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*, vol. 54, no. 5, pp. 573 – 583, 2003.

[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] 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.

[27] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 01 2001.

[28] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[29] 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

[30] H. Hollien, "Vocal indicators of psychological stress," *Annals of the New York Academy of Sciences*, vol. 347, no. 1, pp. 47–72, 1980. [Online]. Available: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1980.tb21255.x

[31] 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.