Interspeech 2022
18-22 September 2022, Incheon, Korea

Multimodal Depression Severity Score Prediction Using Articulatory Coordination Features and Hierarchical Attention Based Text Embeddings

Nadee Seneviratne[1], *Carol Espy-Wilson*[1]

[1]University of Maryland - College Park, USA

nadee@umd.edu, espy@umd.edu

Abstract

Multimodal approaches to predict depression severity is a highly researched problem. We present a multimodal depression severity score prediction system that uses articulatory coordination features (ACFs) derived from vocal tract variables (TVs) and text transcriptions obtained from an automatic speech recognition tool that yields improvements of the root mean squared errors compared to unimodal classifiers (14.8% and 11% for audio and text, respectively). A multi-stage convolutional recurrent neural network was trained using a staircase regression (ST-R) approach with the TV based ACFs. The ST-R approach helps to better capture the quasi-numerical nature of the depression severity scores. A text model is trained using the Hierarchical Attention Network (HAN) architecture. The multimodal system is developed by combining embeddings from the session-level audio model and the HAN text model with a session-level auxiliary feature vector containing timing measures of the speech signal. We also show that this model tracks the severity of depression for subjects reasonably well and we analyze the underlying reasons for the cases with significant deviations of the predictions from the ground-truth score.

Index Terms: depression, multimodal, vocal tract variables, articulatory coordination, staircase regression

1. Introduction

Major Depressive Disorder (MDD) is a mental health disorder that has taken a massive toll on society both socially and financially. Timely diagnosis of MDD is extremely crucial to minimize serious consequences such as suicide. Prosodic, source and spectral features [1] are found to be very effective in speech based depression detection and severity prediction.

Articulatory Coordination Features (ACFs) developed based on psychomotor slowing (a condition of slowed neuromotor output) which is a key feature of MDD [2, 3], quantifies the changes in timing of speech gestures that helps to distinguish depressed and not-depressed speech. Previously, the correlation structure of formants or mel-frequency cepstral coefficients (MFCCs) were used as a proxy for underlying articulatory coordination [4]. Authors of this paper showed that ACFs derived from a set of direct articulatory parameters called Vocal Tract Variables (TVs) are more effective in depression classification [5, 6, 7, 8].

While the changes in the coordination of articulatory gestures convey a lot of information about the mental state of a person, there are other modalities that provide complementary information such as facial expressions, physical gestures and language. Clinicians and psychologists make use of cues from all of these modalities when making a decision about their patient's mental health condition. Recent studies that developed speech based automatic systems to assess depression show the synergies of combining multiple modalities compared to unimodal systems [9, 10, 11, 12]. In [13], we show that the performance of binary depression classification can be improved by using TV-based ACFs and textual features obtained through Automatic Speech Recognition (ASR).

Depression assessment scales evaluate different items pertaining to depression symptoms whose itemized scores add up to the final severity score assigned to a subject under diagnosis. Given that a set of individual items contribute towards the overall severity score, these quasi-numerical scores have an inherent ordinal component. Thus, depression score prediction is even more challenging compared to depression classification. A lot of ongoing work in the speech based depression assessment domain attempts to improve the performance of the depression score prediction task [14, 10, 15, 4, 12, 16].

In this experiment, we gauged the usefulness of TV based ACF in predicting the depression severity score task for the first time. The key contributions of this paper are as follows:

(1) The development of a multimodal system using TV based ACFs for the first time along with textual features to improve the performance of the depression severity score prediction. Generalizability is improved by combining two speech depression databases with different characteristics.

(2) Application of the idea of staircase regression in a deep learning setting for the first time and incorporating the performance boosting of the segment to session approach from [13].

2. Feature Extraction

2.1. Articulatory Coordination Features (ACFs)

ACFs can be used to characterize the level of articulatory coordination and timing. To measure the coordination, assessments of the multi-scale structure of correlations among the TVs were used.

We use the channel-delay correlation matrix proposed in [17] as the ACFs in this work. For an M-channel feature vector $\mathbf{X}$ (such as TVs or formants), 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:

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

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

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

Information pertaining to multiple delay scales are incorporated into the model by using dilated Convolutional Neural Network (CNN) layers with corresponding dilation factors while

10.21437/Interspeech.2022-11099

maintaining a low input dimensionality. Each $R_{i,j}$ will be processed as a separate input channel in the CNN model.

In [13], we showed that TV based ACFs outperformed the ACFs derived from MFCCs and formants and the baseline openSMILE features in the binary depression classification task. Hence, we use TV based ACFs in the depression severity score prediction task as well. TVs are developed based on Articulatory Phonology [18] and define the kinematic state of 5 distinct constrictors (lips, tongue tip, tongue body, velum, and glottis) located along the vocal tract in terms of their constriction degree and location. We use a speaker-independent deep neural network based speech inversion system [19] to estimate 6 TVs for 3 of the constricting organs - Lip Aperture, Lip Protrusion, Tongue Tip Constriction Location, Tongue Tip Constriction Degree, Tongue Body Constriction Location and Tongue Body Constriction Degree. In addition, we use the periodicity and aperiodicity measures obtained from an Aperiodicity, Periodicity and Pitch detector [20] to represent the glottal TV. Before computing the ACFs, TVs were standardized individually.

2.2. Auxiliary Audio Features

We computed additional timing measures to be used as prosodic information which were used as auxiliary speech features to the audio model. These features are speaking rate (number of syllables per second), pause percentage, speech to pause ratio, mean pause duration and standard deviation of pause durations. [21] states that differences in these features can be seen between depressed subjects and those who are in remission. To extract the timing measures we used the algorithm implemented by [22] in Praat that uses the intensity contour of the speech signal.

2.3. Textual Features

Language conveys a great amount of information about people's emotions, behavioral characteristics and social relationships. Therefore, adding language information should help to improve our models. We used the Google speech-to-text API to obtain transcribed text of the free speech recordings that were used to train the audio models. Since the Hierarchical Attention Network (HAN) can be expected to explicitly capture contextual information, we decided to use context-independent GloVe word embeddings (100-dimensions) [23] to initialize the embedding layer of the text model.

3. Model Architectures

3.1. Audio Model - Staircase Regression Approach ($\mathbf{M_{acf}}$)

We extended the segment-to-session-level architecture used in [13] to incorporate staircase regression to predict the severity score. Staircase regression which was previously used in [24, 25] defines an ensemble of models trained on multiple partitions of the same training data set. The outcomes of these individual models are fused via a regressor to obtain the total HAMD score prediction. Staircase regression is particularly interesting as its structure is essentially attempting to answer a collection of simpler questions and build the final prediction on top of those. This approach is able to better capture the quasi-numerical nature of the HAMD scores better.

Inspired by this approach, we trained four segment-level classifiers with 4 different partitions of the dataset as follows: class 0 (low) ranges were 0-7, 0-13, 0-18, 0-22 and class 1 (high) ranges were the complements of these, given the HAMD range from 0 to 52. These range boundaries were chosen according to the standard severity level boundaries for HAMD (Figure 1). The architecture of the segment-level classifier can be found in Figure 2.

Figure 1: *Data partitions used to train the segment-level classifiers in staircase regression approach*

The segment-to-session level model used in [13] was modified to be used as the audio model in this experiment. The output of the first dense layer ($D1$ in Figure 2) of each of these pretrained best segment-level classifiers denoted by $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$ and $\mathbf{X_4}$ were concatenated and passed as the input to the session-level classifier. The dimensions of $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$ and $\mathbf{X_4}$ are 16, 8, 8, and 16, respectively.

A sequence of these concatenated hidden embeddings from the segment-level classifier is passed through two LSTM layers with the second LSTM layer returning a fixed size summary vector. The five dimensional session-level prosodic feature vector is concatenated to this summary vector. It is passed through a dense layer with ReLU activation. The output regression layer with linear activation predicts the z-normalized HAMD score. The estimated score is de-normalized using the precomputed training statistics bringing the HAMD scores to the original range.

3.2. Text Model ($\mathbf{M_t}$) - Hierarchical Attention Network (HAN)

We trained a Bidirectional LSTM based HAN model to obtain a session-level classification for the text model. HAN applies the attention mechanism in word-level and sentence-level taking the hierarchical structure of the transcribed session text into consideration [26]. This allows the model to learn the important words and sentences taking the context into consideration. The embedding layer was fine-tuned for the task by allowing it to train on the errors back-propagated from the output layer.

3.3. Multi-modal Architecture

The multi-modal regressor in Figure 3 was developed with a late fusion approach to perform severity score prediction. The context vector from the second LSTM layer of $\mathbf{M_{acf}}$ and the session-level text vector of $\mathbf{M_t}$ were concatenated with the auxiliary session-level timing feature vector and passed through a Dense layer with ReLu activation to perform HAMD score prediction at the output layer. The late fusion helps to overcome the requirement to have one-to-one correspondence between the audio segments and text sentences and allows us to create segments of different modalities independently in the most optimal way for each modality.

4. Experimental Setup

4.1. Dataset Preparation

Similar to our previous work [7, 13], we used free speech data from two databases: MD-1 [27] and MD-2 [21]. Both databases were collected in a longitudinal study where subjects diagnosed with MDD participated over a period of 6 and 4 weeks, respectively. The clinician rated bi-weekly HAMD scores were used to determine groundtruth labels for the segment level classifiers and also were used as groundtruth scores for the final regression task. Originally there were 472 (35 speakers) and 753 (105 speakers) 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

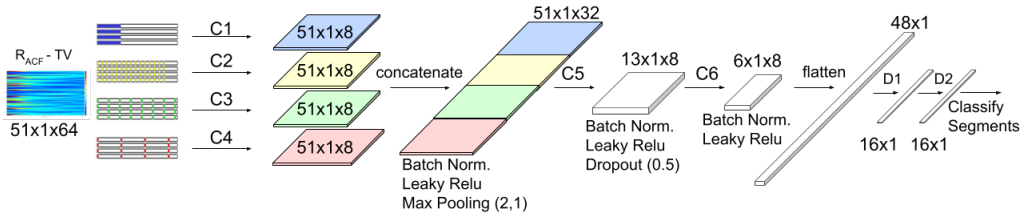


Figure 2: *Dilated CNN architecture for segment-level classification*

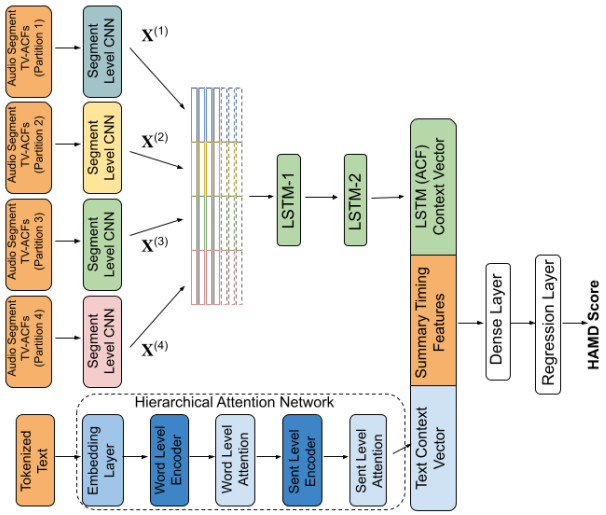


Figure 3: *Staircase regression based multi-modal architecture that uses TV based ACFs, auxiliary timing related summary features and GLoVe embeddings as inputs. LSTM-1 and LSTM-2 have 128 and 64 hidden units (HU) and 0.7 and 0.7 dropout probabilities (DP), respectively. The word-level encoder and sentence-level encoder have 128 and 100 HU and 0.3 and 0.1 DP, respectively. The Dimension of attention layers are 64. The final Dense Layer (before the Regression Layer) has 16 HU.*

in each split and ensuring that there are no speaker overlaps. For the segment-level models, 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. Before extracting the low-level features, segments were normalized to have a maximum absolute value of 1. Output variable (HAMD score) was z-normalized using the mean and variance statistics of the training set. This helped the models to achieve better model convergence. Table 1 summarizes the amount of speech data available after the segmentation for the case when HAMD $>$ 7 was considered as 'depressed' and HAMD $\leq$ 7 was considered as 'not-depressed'.

Table 1: *Available Data in hours/ # segments/ # sessions*

Database	Depressed	Not-depressed
MD-1	11.8 / 2131 / 111	2.5 / 444 / 22
MD-2	16.8 / 3056 / 232	1 / 183 / 17

Before extracting GLoVe embeddings for the text data, the transcribed text was preprocessed by removing punctuation, expanding contractions, lemmatizing and removing stop words (except negation words to preserve the contextual meaning).

4.2. Model Training

Hyper-parameters of the models were tuned using a grid search. Parameter values for the best performing multimodal regressor are given in Figure 3. All models were optimized using an Adam Optimizer. Loss functions used for the segment-level and session-level models were, Binary Cross-Entropy loss and Mean Squared Error loss, respectively. The models were trained with an early stopping criteria based on validation loss (patience was 20 epochs for the segment-level classifiers and 15 epochs for the session-level regressors) for a maximum of 300 epochs. The batch size for the segment-level classifiers was 128. The session-level models were trained with a batch size of 32. To address the class imbalance issue in the segment-level classifiers, class weights were assigned to both training and validation splits during the training process for all the segment level models. All seed values were set to 1729 for training. A learning rate of $2e-5$ was used for the segment-level classifier. The session-level unimodal and multimodal models were trained using an adaptive learning rate starting from $3e-3$ and it was reduced by 50% every 10 epochs until it reached $3.75e-4$. AUC-ROC was used as the performance metric to evaluate the segment-level classifiers. Since predicting the HAMD severity score is a regression task, we used the RMSE and MAE to evaluate the performance of the session-level regressors. Additionally, we also computed the Spearman's rank correlation coefficient (ρ) defined as $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ where d_i is the difference between the two ranks of each observation and n is the sample size. This measure evaluates the correlation of the relative ranking of the groundtruth and predicted severity measures.

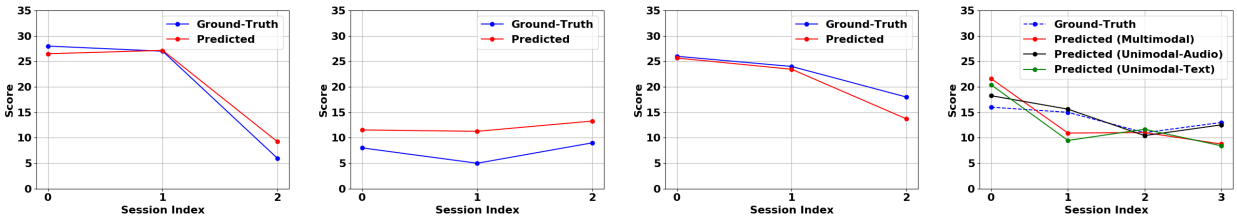
4.3. Experiments and Results

We trained both unimodal and multimodal systems that perform session-level HAMD score prediction. The results are given in Table 2.

Table 2: *HAMD score prediction results*

Feature Set	RMSE	MAE	ρ
Formant_ACF (MD-1 only) [4]	5.99	-	0.48
TV_ACF	6.28	5.23	0.51
TV_ACF + Prosodic	6.13	4.99	0.53
GLoVe	5.87	4.88	0.58
TV_ACF + Prosodic + GLoVe	**5.22**	**4.33**	**0.69**

Using the timing features in addition to the TV based ACFs improved the metrics in general. Therefore we decided to use both TV based ACFs and timing features as speech features in the multimodal regressor. It is interesting to see that the best performing text model outperforms the best performing audio model. The multimodal system was trained using TV based ACFs, timing features and GLoVe embeddings. RMSE of the multimodal system showed relative improvements of 14.82% and 11.03% compared to the best performing audio only and text only models, respectively. The respective relative improvements of ρ were 31.88% and 19.3%. While these improvements yield comparable metrics in comparison to the results from pre-



(a) (b) (c) (d)

Figure 4: *Predicted HAMD scores and groundtruth HAMD scores of subjects*

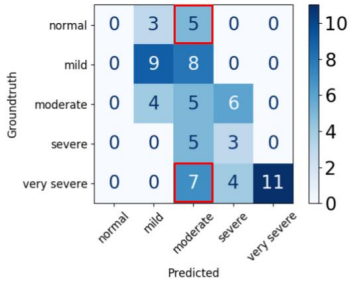


Figure 5: *Confusion matrix of groundtruth and predicted HAMD scores when associated with depression severity levels*

vious studies ([14, 10, 15]), due to the differences in databases used, these results cannot be directly compared. Note that in [4], subject adaptation was used.

4.4. Discussion and Error Analysis

The multimodal system produces better results as shown in table 2. This implies that different modalities provide complementary information to better estimate the depression severity score. When the unimodal regressors are considered, we see a clear improvement in the audio-only model when both TV based ACFs and timing related features were used. This is inline with the findings of previous studies that show that prosody related features help to detect depression.

Even though we obtained sizable improvements for the performance of the score prediction task, there is still room for improvement. While the absolute score being off by a few points may not be super critical given the quasi-numerical nature of the severity scores, it is important that the predicted scores be in the same range as the severity level category of the corresponding groundtruth scores (Eg: HAMD 18 and 16 would both indicate moderate depression). When we categorize the groundtruth and predicted scores into the standard depression severity levels as shown in Figure 1, we obtained the confusion matrix shown in Figure 5. It seems like the model in general overestimates the HAMD score (because there aren't predicted scores that belong to the "normal" category and the summation of superdiagonal elements is higher than the summation of subdiagonal elements). The numbers highlighted by the red squares are off by two levels relative to the groundtruth score which indicates larger errors made by the model. When it comes to reducing the errors of the score prediction model, it is extremely important to reduce errors that significantly underestimate the severity score, as these cases could potentially have serious consequences for human safety. While not as critical, overestimating the score can lead to unnecessarily exhausting resources.

We also analyzed the ability of the multimodal regressor in tracking the depression severity longitudinally. For this, we chose those subjects in the test set who have data for at least 3 sessions for this analysis. For some subjects, the predicted severity scores are remarkably close to the groundtruth HAMD score as shown in Figure 4a. For some subjects, the pattern of changes in the predicted scores follow a similar pattern as seen in the groundtruth scores as shown in Figure 4b. Predicted scores of this subject were overestimated. We also analyzed a case where the model accurately predicted the scores of a majority of the sessions except the score of a single session which heavily deviated from the groundtruth as shown in Figure 4c. For this subject, the predicted HAMD score of the third session is lower than the actual score, while the other predictions are accurate. Inspecting the audio and the text for this case showed no signs of depression even though the session was assigned a high severity score.

While the multimodal system tracks the depression severity reasonably well in general, there are a few instances where the model has performed poorly as shown in Figure 4d. It can be seen that the trend from sessions 1-2 and 2-3 is not being followed by the predictions from the multimodal system and the predictions are very similar to the predictions from the unimodal text-only system. However, the unimodal audio-only system has closely followed the groundtruth scores. It seems like the multimodal system overlooked the information from the audio modality when predicting the severity score. This suggests that implementing an attention mechanism at the modality fusion stage may enable the model to prioritize modalities when predicting the severity score.

5. Conclusion

We presented a multimodal system to predict the depression severity score which utilizes speech data from two different depression databases and text data obtained by ASR. The approach of incorporating staircase regression in the segment-to-session level audio model proved to be effective in the score prediction task. We obtained noteworthy improvements of RMSE, MAE and Spearman's correlation coefficient when the multimodal system was developed combining TV based ACFs, timing features and GLoVe embeddings. We also showed that the model is capable of longitudinal tracking the severity of depression. It could be potentially improved by incorporating subject adaptation. In the future we plan to incorporate video modality which could potentially improve the results further.

6. Acknowledgements

This work was supported by the UMCP & UMB Artificial Intelligence + Medicine for High Impact Challenge Award and the National Science Foundation grant 2124270. We thank Dr. James Mundt for the depression databases MD-1&2 [27, 21] and Dr. Thomas Quatieri and Dr. James Williamson for granting access to the MD-2 database which was funded by Pfizer.

7. References

[1] 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

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

[3] 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

[4] 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

[5] 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

[6] 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

[7] N. Seneviratne and C. Espy-Wilson, "Generalized Dilated CNN Models for Depression Detection Using Inverted Vocal Tract Variables," in *Proc. Interspeech 2021*, 2021, pp. 4513–4517.

[8] ——, "Speech Based Depression Severity Level Classification Using a Multi-Stage Dilated CNN-LSTM Model," in *Proc. Interspeech 2021*, 2021, pp. 2526–2530.

[9] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4235–4239.

[10] 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

[11] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3946–3950.

[12] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated cnns," 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. 73–80. [Online]. Available: https://doi.org/10.1145/3347320.3357695

[13] N. Seneviratne and C. Espy-Wilson, "Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6252–6256.

[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] 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

[16] M. R. Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations 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. 55–63. [Online]. Available: https://doi.org/10.1145/3347320.3357694

[17] 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

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

[19] 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

[20] 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.

[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] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, May 2009. [Online]. Available: https://doi.org/10.3758/BRM.41.2.385

[23] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[24] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 41–48. [Online]. Available: https://doi.org/10.1145/2512530.2512531

[25] N. Cummins, V. Sethu, J. Epps, J. Williamson, T. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 10 2017.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489. [Online]. Available: https://aclanthology.org/N16-1174

[27] 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