ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747462

MULTIMODAL DEPRESSION CLASSIFICATION USING ARTICULATORY COORDINATION FEATURES AND HIERARCHICAL ATTENTION BASED TEXT EMBEDDINGS

Nadee Seneviratne and Carol Espy-Wilson

University of Maryland - College Park, USA

ABSTRACT

Multimodal depression classification has gained immense popularity over the recent years. We develop a multimodal depression classification system using articulatory coordination features extracted from vocal tract variables and text transcriptions obtained from an automatic speech recognition tool that yields improvements of area under the receiver operating characteristics curve compared to unimodal classifiers (7.5% and 13.7% for audio and text respectively). We show that in the case of limited training data, a segment-level classifier can first be trained to then obtain a session-wise prediction without hindering the performance, using a multi-stage convolutional recurrent neural network. A text model is trained using a Hierarchical Attention Network (HAN). The multimodal system is developed by combining embeddings from the session-level audio model and the HAN text model.

Index Terms— depression detection, multimodal, vocal tract variables, articulatory coordination features, hierarchical attention

1. INTRODUCTION

Major Depressive Disorder (MDD) is a mental health disorder with serious consequences. Previous studies have shown that vocal biomarkers developed using prosodic, source, and spectral features [3] can be very effective in automatic depression detection to enable timely diagnosis and prompt treatments.

MDD is known to cause changes in articulatory coordination of speech due to a neurological condition called psychomotor slowing which is a necessary feature of MDD [4, 5, 6]. Articulatory Coordination Features (ACFs) have yielded successful results in distinguishing depressed speech from non-depressed speech by quantifying these changes in the timing of speech gestures [7, 8, 9]. Previously, the correlation structure of the formants or Mel Frequency Cepstral Coefficients (MFCCs) was used as a proxy for articulatory coordination to derive indirect ACFs which showed promise in the depression detection task [7]. 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 depressed and non-depressed speech can yield relatively better results in the depression detection task [8, 9] and for depression severity level classification task [10]. TV-based ACFs also showed promise as a robust set of features for depression classification by generalizing well across the two databases [11].

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

Deep learning based depression detection is a highly researched area with promising results [12, 13]. Among these works, multimodal depression classification and severity prediction attempt to further improve the performance through inter-learning among different modalities [14, 15, 16, 17]. Several studies have used the approach of aggregating segment-level predictions to obtain a final subject/session-level prediction using techniques such as plurality voting (PV) [18, 19], max-pooling [20] or recurrent neural networks (RNN) [21, 10]. This is especially useful in a setting where there's a lack of training samples to train a deep learning model using full audio recordings which can lead to overfitting issues. It was shown empirically that the strengths of the segment-level classifier are amplified as a result of its repeated usage in the session-level classifier [19, 10]. In a multimodal setting, most of these aggregating approaches have used one-to-one correspondence among segments from different modalities [21, 19].

The key contributions of this paper are as follows:

(1) The development of a multimodal system using depression corpora that contain only speech data. Utilizing ASR to obtain text transcriptions, we show that for the first time, the performance of binary depression classification can be improved by using TV-based ACFs and textual features. The generalizability is improved by sourcing data from two different depression databases. The fusion strategy of the proposed architecture enables us to segment data from different modalities independently, in the most optimal way for each modality, when performing session-level classification from segment-level classification.

(2) The analysis of the constraints to be satisfied by the segment-level classifier to yield a stronger session-level classifier. Using a multi-stage convolutional recurrent neural network, these analytical findings are validated empirically using different sets of ACFs and openSMILE features.

2. FEATURE EXTRACTION

2.1. Audio Features

Vocal Tract Variables (TVs): Developed based on Articulatory Phonology [22], TVs 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 [23] 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 (APP) detector [24] to represent the glottal TV. At this time, we provide no information on the velum.

For comparison purposes, we also trained models using ACFs derived from formants and MFCCs. The first three formant frequen-

cies were obtained using the Karma formant tracking tool [25]. 12 MFCC time series were extracted (window size of 20ms, overlap of 10ms) discarding the (1st MFCC coefficient. We created two more sets of ACFs by appending the same glottal parameters to formants (FMT+GL) and MFCCs (MFCC+GL) to investigate the effect of adding voice source information.

2.1.1. 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 time series signals such as TVs were used.

We use the channel-delay correlation matrix proposed in [26] 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 maintaining a low input dimensionality. Each $R_{i,j}$ will be processed as a separate input channel in the CNN model. Before computing the ACFs, feature vectors were standardized individually.

2.1.2. Baseline Acoustic Features

We trained a baseline model using the openSMILE features (window size of 20ms, overlap of 10ms) to benchmark the performance of models trained using ACFs. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [27] was extracted using the openSMILE toolkit [28]. This 23-dimensional feature set consists of spectral, cepstral, prosodic and voice quality parameters.

2.2. Textual Features

Linguistic features reveal important information about the mental health of a depressed subject. Therefore adding semantic contextual information should help to improve our models. We used Google speech-to-text API to obtain transcribed text of the Free Speech (FS) 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-dimensional) [29] to initialize the embedding layer of the text model.

3. SEGMENT TO SESSION-LEVEL CLASSIFICATION

Let $\mathbf{h}$ denote the segment-level classifier, $\mathbf{H}$ denote the session-level classifier, $\mathbf{C}$ be an oracle which provides ground-truth, $\{C_0, C_1\}$ be the classes and S be a session consisting of $S_1, S_2, ..., S_N$ segments. Let's consider an arbitrary class C_0 and prove that the recall of class C_0 of a PV session-level classifier is better than that of the segment-level classifier it is based on. We use a PV classifier here for simplicity, however RNN based approaches yield more generalizable classifiers.

We assume $p_j = \Pr(h(S_i) = C_j \mid C(S_i) = C_0)$ to be the same $\forall i$ and each segment classification is independent. Note that p_0 is the recall of class C_0 and $p_0 + p_1 = 1$. Consider a PV classifier which breaks ties by randomly selecting one class. Let P_0 denote the recall of the combined classifier.

$$P_0 = \Pr(H(S) = C_0 \mid C(S) = C_0) = \sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) p_0^k p_1^{N-k} \text{ where } d(k) = \begin{cases} 1/2, \text{if } k = N/2 \\ 1, \text{otherwise} \end{cases} \tag{4}$$

Using $\binom{N}{k} = \binom{N}{N-k}$ and $d(k) = d(N-k)$, the sum of coefficients of (4) can be written as,

$$\sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) = \sum_{k=\lceil N/2\rceil}^{N} \binom{N}{N-k} d(N-k) = \sum_{k=0}^{\lfloor N/2\rfloor} \binom{N}{k} d(k)$$

With that we have,

$$2 \sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) = \sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) + \sum_{k=0}^{\lfloor N/2\rfloor} \binom{N}{k} d(k) = \sum_{k=0}^{N} \binom{N}{k}$$

Therefore we have the sum of coefficients of (4),

$$\sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) = \frac{\sum_{k=0}^{N} \binom{N}{k}}{2} = (1+1)^N/2 = 2^{N-1}$$

Consider the difference of recalls. We use (4) to substitute for P_0 and artificially multiply p_0 by a term equal to 1 to ensure that the coefficient sums of both the terms are equal. Since $p_0 + p_1 = 1$,

$$P_0 - p_0 = \underbrace{\sum_{k=\lceil N/2\rceil}^{N} \binom{N}{k} d(k) p_0^k p_1^{N-k}}_{\text{Coefficient sum is } 2^{N-1}} - p_0 \underbrace{(p_0+p_1)^{N-1}}_{\text{Coefficient sum is } 2^{N-1}}$$
$$= \sum_{k=\lceil N/2\rceil}^{N} \left[d(k)\binom{N}{k} - \binom{N-1}{k-1} \right] p_0^k p_1^{N-k} - \sum_{k=1}^{\lceil N/2\rceil - 1} \binom{N-1}{k-1} p_0^k p_1^{N-k} \tag{5}$$

Given that

$$d(k)\binom{N}{k} - \binom{N-1}{k-1} = \binom{N-1}{k-1}(d(k)N/k - 1) = \begin{cases} 0, \text{if } k \in \{N/2, N\} \\ > 0, \text{ otherwise} \end{cases}$$

we can group the terms in the expansion of (5) into pairs of the form

$$p_0^r p_1^l (p_0^t - p_1^t) \text{ where } r > 0, l \geq 0, t \geq 0 \text{ and } r + l + t = N \tag{6}$$

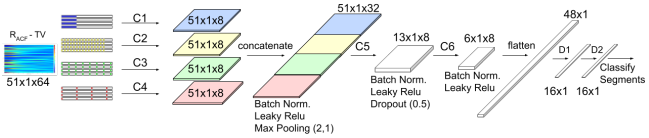
which are non-negative when $p_0 \geq p_1$. Therefore we establish that $P_0 - p_0 \geq 0$ or $P_0 \geq p_0$ when $p_0 \geq p_1$. That is, if all the classes have better than 50% recall in the segment-level classifier, PV based session-level classifier would result in a better recall for all the classes.

4. MODEL ARCHITECTURES

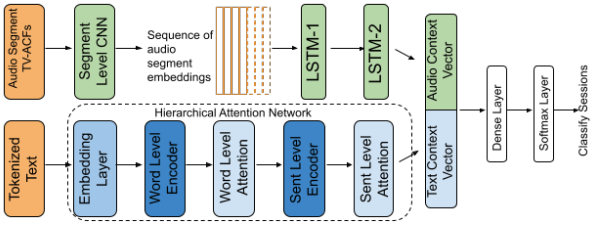
4.1. Audio Model

Baseline Segment-Level Classifier: We trained a CNN using the openSMILE features as a baseline model for this task. The input is passed through two sequential 1-D (across time axis) convolutional layers. Each convolutional layer is followed by batch normalization, leaky ReLU activation, dropouts and a max-pooling layer. The output from the second max-pooling layer is flattened and passed through two dense layers to perform classification at the output layer. The output of the second dense layer is extracted and used as the input to the session-level classifier.

Dilated CNN based Segment-Level Classifier for ACFs: A dilated CNN proposed in [26] was trained using the ACFs to classify



(a)



(b)

Fig. 1. (a) Dilated CNN architecture for segment-level classification. Hyper-parameters for the best performing audio model (TV-based ACF) are mentioned. Kernel size of C5 and C6 is (3,1). (b) Architecture of the multimodal classifier. 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 100 HU each and 0.3 and 0.1 DP, respectively. The Dimension of attention layers are 64. The final Dense Layer (before the Softmax Layer) has 64 HU.

the segments (Fig. 1a). The input $\widetilde{R}_{ACF}$ is fed into 4 parallel convolutional layers with different dilation rates $n = \{1, 3, 7, 15\}$ and a kernel size of $(15, 1)$ which resembles the multiple delay scales. The outputs of these 4 parallel layers are concatenated and then passed through two sequential convolutional layers. This output is flattened and passed through two dense layers to perform segment-level classification in the output layer. All convolutional layers used LeakyReLU activation, whereas the dense layers used ReLU activation with l_2 regularization ($\lambda = 0.01$). The flattened output of $C6$ is passed as input to the session-level classification.

RNN based Session-Level Classification: The segment embeddings are extracted from the segment-level classifiers as a sequence and are passed through a Long Short-Term Memory (LSTM) based RNN model to perform the session-level classification. The input is passed through two LSTM layers followed by a Dense layer with ReLU activation. Finally, the output layer with Softmax activation performs the session-level Classification. Recurrent dropout probabilities are applied to the two LSTM layers.

4.2. Text Model

We trained a Bidirectional LSTM based HAN model shown in Fig. 1 to obtain a session-level classification for the text model. HAN applies the attention mechanism in two levels: word-level and sentence-level taking the hierarchical structure of a document into consideration [30]. This allows the model to learn the important words and sentences taking the context into consideration. In this work, a document corresponds to the transcribed text of a session. The embedding layer was fine-tuned for the task by allowing it to back-propagate the error from the output layer.

4.3. Multimodal Depression Classifier

The multimodal system (Fig. 1b) is constructed with embeddings from the session-level audio classifier ($\mathbf{M_a}$) and HAN-based text classifier ($\mathbf{M_t}$). The context vector from the second LSTM layer of $\mathbf{M_a}$ and the attention-weighted sentence level context vector of $\mathbf{M_t}$ were concatenated and passed through a Dense layer with ReLu activation to perform final binary classification at the output layer. This late fusion structure helps to avoid overfitting issues that can occur as a result of the high dimensionality of input features when using early fusion. It also 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 (overlapping segments for audio and sentences for text).

5. EXPERIMENTS AND RESULTS

5.1. Dataset Preparation

Similar to our previous work [11], we used FS data from two databases: MD-1 [1] and MD-2 [2]. Both databases were collected in a longitudinal study where subjects diagnosed with MDD participated over a period of 6 and 4 weeks, respectively. For the binary classification problem, ground truth labels were determined by the bi-weekly scores provided for the clinician-rated 17-item Hamilton Depression Rating Scale (HAMD). Sessions with HAMD > 7 were considered as 'depressed' and sessions with HAMD ≤ 7 were considered as 'not-depressed'. Due to the availability of 2 clinician-rated depression scores in MD-2, the agreement between the two scores in terms of the severity level was considered (see Table 1 in [11]). 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 segment-level models trained on ACFs, 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 1 summarizes the amount of speech data available after the segmentation. For the baseline model trained on openSMILE features, all 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, segments were normalized to have a maximum absolute value of 1.

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).

5.2. Model Training

Hyper-parameters of the models were tuned using a grid search. The ranges used were as follows: the kernel size {(3,1), (4,1)} and the number of output filters {128, 64, 32, 16, 8} of the convolutional layers, the number of hidden units of dense layers {16, 8}, the number of hidden units of LSTM layers {128, 100, 64, 32}, the dimension of the attention vectors {128, 100, 64} and dropout probabilities {0.4, 0.5, 0.6, 0.7}. The models were optimized using an Adam Optimizer for the Binary Cross Entropy loss. The models were trained with an early stopping criteria based on validation loss (patience 20 epochs) for a maximum of 300 epochs. A batch size was 128. 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 classifiers were trained using an adaptive learning rate starting from $2e-4$ and it was decayed by 50% every 10 epochs until it reached $2e-5$. 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, the Area Under the Receiver Operating Characteristics Curve (AUC-ROC), Unweighted Average Recall (UAR) and F1 scores were used.

Table 2. Classification Results - Audio Model

Model	Features	AUC-ROC	UAR	F1 (D/ND)
Segment-Baseline	openSMILE	0.6300	0.5602	0.86/0.22
Segment-Level Classifier (ACFs)	**TV**	**0.7408**	**0.6961**	**0.87/0.37**
	MFCC	0.6031	0.5412	0.88/0.19
	FMT	0.5714	0.5688	0.77/0.22
	MFCC+GL	0.6042	0.5474	0.87/0.20
	FMT+GL	0.4806	0.5045	0.79/0.16
Session-Baseline	openSMILE	0.6673	0.6613	0.87/0.35
Session-Level Classifier (ACFs)	**TV**	**0.8246**	**0.8024**	**0.91/0.52**
	MFCC	0.7016	0.6452	0.85/0.32
	FMT	0.75	0.7238	0.88/0.42
	MFCC+GL	0.6794	0.6452	0.85/0.32
	FMT+GL	0.6552	0.6734	0.73/0.31

5.3. Segment-Level to Session-Level Classification

We trained 6 different audio-based segment-level to session-level classifiers using the ACFs derived from various feature vectors and openSMILE features (baseline). Results can be found in Table 2. In general, for all feature sets, there is a performance boost in the session-level classifier compared to the segment-level classifier. In both segment-level and session-level classifications, TV-based ACFs outperform the other features. TV-based ACFs yield a relative AUC-ROC improvement of 11.3% and a relative UAR improvement of 15.3% in session-level classification compared to segment-level classification. The AUC-ROC and UAR of session-level TV-based ACF classifier are 9.9% and 9.8% higher than the second best performing session-level classifier which was trained using formant-based ACFs. The chance-level F1 scores for depressed/not-depressed classes were 0.64/0.18 (segment-level) and 0.64/0.19 (session-level).

Table 3. Results of Classification Using Different Modalities

Model	Features	AUC-ROC	UAR	F1 (D) / F1 (ND)
Audio	TV_ACF	0.8246	0.8024	0.91/0.52
Text	GloVe	0.7802	0.7540	0.85/0.41
multimodal	**TV_ACF + GloVe**	**0.8871**	**0.8105**	**0.92/0.55**

5.4. Results of multimodal Classification

Using the best performing audio model which was trained using TV-based ACFs and HAN based text model, we trained a multimodal system that yields synergies by combining different modalities. According to Table 3, the multimodal system has a relative AUC-ROC improvement of 7.58% and 13.7% compared to the audio model and the text model, respectively.

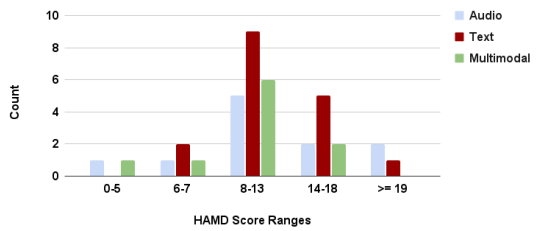


Fig. 2. Distribution of HAMD scores for the mis-classified samples categorized by severity levels: not-depressed (0-5), borderline not-depressed (6-7), mild (8-13), moderate (14-18), severe ($\geq$19)

6. DISCUSSION

Results reported in Table 2 show that TV-based ACFs are more effective compared to other feature sets in the binary depression classification task. These results support the hypothesis that TVs as a direct measure convey more distinguishing information regarding the articulatory coordination of depressed speech. Further, it is evident from the results that the performance of the session-level classifier heavily relies on the performance of the segment classifier. We observed that the segment-level TV-based classifier satisfied the constraints we derived in section 3 (the recall of depressed and not-depressed classes were 58.1% and 81.1%, respectively). Consequently, the TV-based session-level classifier yielded better results with recall increased to 75% and 85.5% for the two classes respectively.

According to Table 3, the audio-only model performs better than the text-only model. One possible reason for this could be errors introduced to the text transcripts by the ASR tool. We further investigated the sentence-level attention weights of the mis-classified sessions to understand the nature of the errors. Most of the mis-classifications are cases where the subjects use complex sentence structures (often with mixed sentiments) in their answers. However the model seems to capture the most frequently occurring raw sentiment and not take the sentence structure into account. A common pattern is when patients use contrast in their sentences. Consider the following example. Text: *"Like I said in the beginning, I am feeling much better. I'm not feeling sad. I'm not feeling guilty and feeling like I cannot do anything like before"*. Ground Truth: not depressed (misclassified as depressed). Note that the first segment of the text *"I am feeling much better. I'm not feeling sad. I'm not feeling guilty"* in isolation conveys a positive sentiment. The last segment of the text *"feeling like I cannot do anything like before"* conveys a negative sentiment to which the model has paid higher attention. However if the underlined *'not'* is also applicable to this segment, it results in double negation and consequently the segment conveys a positive sentiment. Another common pattern is when the patients excessively use negation. Consider the following example. Text: *"I do not feel like I do not want to do anything"*. Ground Truth: not depressed (misclassified as depressed). Note the usage of double negation which results in a complex sentence structure. In the future we plan to explore context aware embeddings which are capable of correctly processing such complex sentence structures.

For the multimodal classifier, the recall of depressed and not-depressed classes were 87.1% and 75%, respectively. From a clinical perspective, the model is effective in recognizing depressed subjects, reducing type-II errors (classifying a depressed subject as not-depressed). Nine out of the total of 10 mis-classified sessions by the multimodal system are a subset of the 25 mis-classified sessions by unimodal classifiers. This shows that the inter-learning among different modalities can compensate for the errors made by individual modalities. It is worth noting that when a session is incorrectly classified by both the unimodal classifiers, it is always incorrectly classified by the multimodal classifier. According to Fig. 2, it can be seen that the multimodal classifier is able to correctly classify the depressed sessions in the severe category which were incorrectly classified by the unimodal classifiers.

7. CONCLUSION

We presented a multimodal system which utilizes audio data from two different depression databases and text data obtained by ASR. The proposed system performs better compared to unimodal classifiers. The robustness of the TV-based ACFs is evident from the performance of the audio-only models trained using a comprehensive set of features. We established the constraints that need to be satisfied by a segment-level classifier in order to yield a stronger session-level classifier. In the future we plan to incorporate other linguistic features to improve the text model and investigate the performance of TV-based ACFs in the depression severity score prediction task.

8. REFERENCES

[1] James C. Mundt, Peter J. Snyder, Michael S. Cannizzaro, Kara Chappie, and Dayna 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.

[2] James C. Mundt, Adam P. Vogel, Douglas E. Feltner, and William 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.

[3] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10 – 49, 2015.

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

[5] American Psychiatric Association, *Copyright*, Washington, DC, 2000.

[6] Daniel 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.

[7] James R. Williamson, Diana Young, Andrew A. Nierenberg, James Niemi, Brian S. Helfer, and Thomas F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," *Computer Speech & Language*, vol. 55, pp. 40 – 56, 2019.

[8] Carol Espy-Wilson, Adam C. Lammert, Nadee Seneviratne, and Thomas F. Quatieri, "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," in *Proc. Interspeech 2019*, 2019, pp. 1448–1452.

[9] Nadee Seneviratne, James R. Williamson, Adam C. Lammert, Thomas F. Quatieri, and Carol 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.

[10] Nadee Seneviratne and Carol Espy-Wilson, "Speech Based Depression Severity Level Classification Using a Multi-Stage Dilated CNN-LSTM Model," in *Proc. Interspeech 2021*, 2021, pp. 2526–2530.

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

[12] 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.

[13] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 2016, AVEC '16, p. 35–42, Association for Computing Machinery.

[14] Meng Niu, Kai Chen, Qingcai Chen, and Lufeng 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.

[15] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu 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*, New York, NY, USA, 2019, AVEC '19, p. 81–88, Association for Computing Machinery.

[16] 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.

[17] Weiquan Fan, Zhiwei He, Xiaofen Xing, Bolun Cai, and Weirui Lu, "Multi-modality depression detection via multi-scale temporal dilated cnns," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, New York, NY, USA, 2019, AVEC '19, p. 73–80, Association for Computing Machinery.

[18] Adrián Vázquez-Romero and Ascensión Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, 2020.

[19] Nujud Aloshban, Anna Esposito, and Alessandro Vinciarelli, "Detecting depression in less than 10 seconds: Impact of speaking time on depression detection sensitivity," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 79–87, Association for Computing Machinery.

[20] Amir Harati, Elizabeth Shriberg, Tomasz Rutowski, Piotr Chlebek, Yang Lu, and Ricardo Oliveira, "Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7273–7277.

[21] Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang, "A multimodal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, New York, NY, USA, 2019, AVEC '19, p. 65–71, Association for Computing Machinery.

[22] Catherine P Browman and Louis Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[23] Ganesh Sivaraman, Vikramjit Mitra, Hosung Nam, Mark Tiede, and Carol 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.

[24] 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.

[25] Daryush D. Mehta, Daniel Rudoy, and Patrick J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, Sep 2012.

[26] Zhaocheng Huang, Julien Epps, and Dale 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*. 2020, pp. 6549–6553, IEEE.

[27] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet 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.

[28] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, p. 1459–1462, Association for Computing Machinery.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543, Association for Computational Linguistics.

[30] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard 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, June 2016, pp. 1480–1489, Association for Computational Linguistics.